Translating and the Computer 44

24-25 November 2022

European Convention Center, Luxembourg

ISBN 978-2-9701733-0-4

September 2023. Editions Tradulex, Geneva

This document is downloadable from www.tradulex.com and www.asling.org

Acknowledgements

AsLing wishes to thank and acknowledge the support of the sponsors of TC44:

Gold Sponsors

Silver Sponsors

Preface

After two years of virtual events, the 44th Translating and the Computer conference (TC44), organised by the International Association for the Advancement in Language Technology (AsLing), once again took place in person, allowing fruitful exchanges among academics, developers, users, and vendors of computer aids for translators, of other translation technology tools, and increasingly, for interpreters and others performing new roles in our industry.

AsLing was able to host its first post-Covid in-person conference at the European Convention Center, in Luxembourg, a move that was very well accepted and brought us almost 230 participants, exceeding by far those who attended most of the earlier in-person events in London for TC1-TC41. After 2 virtual editions, as a transitional measure, TC44 was also fully web-streamed live, enlarging the audience to hundreds of people who could not travel to join us in-person. By meeting in Luxembourg AsLing also reduced the travel distance of most conference participants. We thank one of our sponsors, the Publications Office of the European Union, for making it possible to hold the conference in Luxembourg. Their support and that of our other sponsors made it possible to organize this "traditionally new" conference: thank you Terminotix, Star AG, Wordfast and Juremy, as well as the Publications Office.

TC44 featured two keynote addresses, 27 presentations, 13 workshops and three panel discussions. PowerPoint slides of presentations, panels and workshops are available on the AsLing website, together with short bios of the presenters, panellists and moderators of the workshops and panels. We strongly recommend browsing www.asling.org/tc44. The only thing that cannot be found there are the lively in-person discussions covering a broad range of subjects and tools, that brought together freelance and in-house translators, interpreters, researchers and businesspeople from translation companies, international organisations, universities and research centres, and offered them opportunities to exchange ideas and to learn about and discuss the latest developments in translation technologies.

Our two keynote speakers brought clarity to two areas of translation requiring increased attention. Valter Mavrič, Director-General for Translation at the European Parliament, provided an in-depth view of what has already been achieved and what is in development at the Parliament to improve accessibility for the deaf and hard of hearing through language technology, with the help of real-time speech-to-text and various machine translation tools. Luisa Bentivogli, from the Machine Translation unit at Fondazione Bruno Kessler, Italy, talked about the biases of MT systems, with a special focus on gender bias in MT, shedding light both on the various aspects of bias as well as on approaches for mitigation.

We thank all who submitted proposals to the conference and those authors who produced full versions of their papers for these Proceedings, as well as all whose slides are available on the AsLing website. A special thank-you goes to the Editors of these proceedings: without their hard work, the publication would have been impossible. We are grateful to the members of the Programme Committee who carefully reviewed the submissions as well as all additional reviewers who helped assess some of the final papers and to our fellow members of the Organising Committee, who played key roles in ensuring that this year's conference, again in-person, but at a new location, took place and linked people from all continents. Finally, we thank all those who lent their support, helping to make both the conference and these Proceedings possible.

Conference Chairs

João Esteves-Ferreira, Ruslan Mitkov, Maria Recort Ruiz, Olaf-Michael Stefanov

The Executive Committee of AsLing establishes several bodies each year, to organise and carry out the annual conference. Membership in these bodies overlap. The tables below show membership in these bodies for TC44.

Conference Organising Committee:
Denis Dechandon, European Union (Session Chair)
João Esteves-Ferreira, Tradulex (Conference Chair)
Ruslan Mitkov, Lancaster University (Conference Chair)
Joss Moorkens, Dublin City University/ADAPT Centre (Session Chair)
Maria Recort Ruiz, International Labour Office (Conference Chair)
Vilelmini Sosoni, Ionian University (Session Chair)
Olaf-Michael Stefanov, United Nations (ret.), (Conference Chair)
Coordinators: João Esteves-Ferreira and María Recort Ruiz

Editors of the Proceedings:
Joss Moorkens
Vilelmini Sosoni

Programme Committee:
Juan José Arevalillo, Hermes Traducciones
Frédéric Blain. Tilburg University
Lynne Bowker, University of Ottawa
Vicent Briva-Iglesias, Dublin City University
Burcu Can, University of Stirling
Sheila Castilho, Dublin City University/ADAPT Centre
David Chambers, AsLing Honorary Member
Dragoș Ciobanu, University of Vienna
Eleanor Cornelius, University of Johannesburg
Félix do Carmo, University of Surrey
Gökhan Doğru, Universitat Autònoma de Barcelona & Dublin City University
Joanna Drugan, Heriot-Watt University
Emmanuelle Esperança-Rodier, Université Grenoble Alpes
María Fernandez-Parra, Swansea University
David Filip, Huawei Ireland Research Center
Amal Haddad Haddad, University of Granada
Camelia Ignat, Joint Research Centre of the European Commission
Valentini Kalfadopoulou, Ionian University
Elpida Loupaki, Aristotle University of Thessaloniki.
Raisa McNab, UK Association of Translation Companies
Elizabeth Marshman, University of Ottawa
Ruslan Mitkov, Lancaster University
Joss Moorkens, Dublin City University/ADAPT Centre
Dora Murgu, Interprefy
Rozane Rebechi, University of Rio Grande do Sul
Vilelmini Sosoni, Ionian University
Cristina Toledo, University of Málaga
Paola Valli, Project Manager, Tamedia
Nelson Verástegui, International Telecommunications Union (ret.)
David Verhofstadt, European Investment Bank
Michal Ziemski, ETH Zürich

Contents

Section A: Accessibility and speech technology

Accessibility through language technology at the European Parliament 8
Valter Mavrič

Past, present and future of speech technologies in translation — life beyond the keyboard 16
Julian Zapata, Alina Secară and Dragoș Ciobanu

Section B: Machine translation and users

Impact of Domain-Adapted Multilingual Neural Machine Translation in the Medical Domain 27
Miguel Rios, Raluca-Maria Chereji, Alina Secară and Dragoș Ciobanu

A Study Towards a Standardized Typology of Machine Translation Post-Editing Guidelines: A Suggested Template for Language Professionals 39
Lucía Guerrero and Viveta Gene

Do translators use machine translation and if so, how? Results of a survey among professional translators 49
Michael Farrell

Peculiarities of Polish academic legal writing in English translation: field experts vs. algorithms 61
Anna Setkowicz-Ryszka

Why are generic MT engines of limited assistance to legal academics wishing to communicate in English as a Lingua Franca? A reviser's and post-editor's perspective 70
Anna Setkowicz-Ryszka

Evaluation of adaptive machine translation from a gender-neutral language perspective 82
Aida Kostikova, Todor Lazarov and Joke Daems

Introducing Fairslator: a machine translation bias removal tool 90
Michal Měchura

SECTION C: Multi-word expressions, terminology and corpora

Expert data: a French MWE Manually Annotated Corpus 97
Emmanuelle Esperança-Rodier, Fiorella Albasini and Yacine Haddad

gApp: a text preprocessing system to improve the neural machine translation of discontinuous multiword expressions 111
Carlos Manuel Hidalgo Ternero and Xiaoqing Zhou Lian

The use of CAT tools and corpus analysis in comparative literary translation research: an English-Arabic case study 119
Amal Haddad Haddad

HypoLexicon: A Terminological Resource for Describing Hyponymic Information 130
Juan Carlos Gil Berrozpe

Using bitext mining to identify translated material: practical assessment and new applications 147
Zhilu Tu, Minghao Wang, Mark Shuttleworth and Zhiwen Hua

Section A: Accessibility and speech technology

Accessibility through language technology at the European Parliament

Valter Mavrič

Based on the keynote speech to AsLing's 44th Translation and the Computer conference

1. Introducing Parliament's translation service

The European Parliament's translation service (DG TRAD) has a crucial role in safeguarding one of the essential principles of European democracy, which is firmly rooted in the European Treaties: multilingualism. In Parliament, this is very visible, as Members of the European Parliament have the right to use any of the 24 official EU languages in parliamentary meetings and have access to documents in their own language. The European Parliament's translation service translates all documents relating to Parliament's legislative, budgetary and scrutiny processes into all official EU languages. In this way, it contributes to the legitimacy and transparency of Parliament's functioning, since all citizens of the European Union have full access to documents of interest to them and can communicate with the European Parliament in their own language.

Parliament's translation service works in Parliament's three places of work, Luxembourg, Brussels and Strasbourg. It has evolved considerably over time; whereas in 1958, it worked in only four official EU languages, this number has increased with each successive round of EU enlargement to the current number of 24 official EU languages. This has led not only to an increase in the number of translated documents but has also added multiple layers of complexity. Currently, DG TRAD provides language services in no less than 552 language combinations and has around 1,300 language professionals and support staff to enable it to do so.

With the aim of reaching an even larger number of citizens, DG TRAD´s mission is to increase the accessibility of Parliament's content, not only through translating documents into all official EU languages, but also by promoting clear language, both through drafting support and multilingual translation and transcreation, in formats which allow audiences to easily find and use this information. In the European Parliament, this approach is called Citizens' Language: promoting clear language for all writers in 24 languages and in text, audio and video formats.

2. Citizens' language

2.1. What is Citizens' Language?

The work of the European Parliament has an impact on the lives of European citizens through the adoption of laws and resolutions, but it is often complex to follow Parliament's procedures and read its documents. The European Parliament is the European Union's only directly elected institution and, as such, it must be transparent and accountable to citizens. Therefore, creating clear, accessible and multilingual content is a vital part of this.

According to the International Plain Language Federation, a communication is in plain or clear language if its wording, structure, and design are so clear that the intended readers can:

1. easily find what they need;
2. understand what they find;

3. and use that information.

Clear language policies have gained widespread attention in academia over recent years. They have found their way into the public sectors of many democracies around the world. This is why, in 2019, Parliament's translation service laid the foundations for Parliament's very own Citizens' Language policy. This policy, formally launched in 2020, promotes the use of clear language for all writers in all 24 official EU languages, in text, audio and video formats. The Citizens' Language policy gives Parliament's Members and staff the knowledge and tools to make the language of their multilingual communication with each other and with citizens clearer. This is done with the help of Parliament's DG TRAD language professionals through drafting support and translation. They primarily translate laws and communication products, but they are also there to provide a wide range of language services, including localisation, audiovisual translation and adaptation, transcreation, post-editing of machine translation, proof-reading, innovative technology solutions, and finally, guidelines and training.

More specifically, the Citizens' Language policy translates into a number of specific projects and actions related to text, audio and video formats. Let's take a more detailed look into the current situation:

a. Text: DG TRAD's language professionals now translate texts intended for a wide audience in a way that is tailored to their specific needs or culture. This includes press releases and webpages and will soon extend to Parliament's political resolutions.

b. Audio: This easy-to-use format allows listeners to access content anywhere and anytime. It also opens up new opportunities for people who are blind or partially sighted to familiarise themselves with Parliament's work. DG TRAD produces audio podcasts in 24 languages based on Parliament's existing written content, creates voice-overs and even broadcasts content via a web-streamed radio station in six languages, called Europarl Radio.

c. Video: DG TRAD has become an expert in subtitling in 24 languages for hearing audiences but also for people who are deaf and hard-of-hearing. It subtitles a wide range of products, such as social media videos, speeches and even the films shortlisted for Parliament's annual LUX Audience Award.

2.2. Parliament's translation service as content provider

DG TRAD has consolidated its commitment to clear language in its strategic objective to support multilingual communication with citizens in clear language. In this framework, DG TRAD adapts Parliament's existing content, created by its research or communication services, into multilingual products in various formats. For example, every day, a selection of Parliament's press releases is condensed into a multilingual podcast called *News in Brief*, published on Parliament's channels in all 24 official EU languages, as well as in Ukrainian since the spring of 2022.

However, in some instances, DG TRAD has expanded its traditional role as a provider of language services to actively contribute to the creation of multilingual content. For example, with the *My House of European History*[1] platform. A spin-off of Parliament's *House of European History* museum in Leopold Park in Brussels, this online platform aims to collect personal testimonials from citizens linked to historic events and periods on the European continent.

People from all over the world are invited to submit their personal stories on the *My House of European History* website in any of the 24 official EU languages, in text, audio and video format. DG TRAD then not only translates a selection of these stories into other languages, but also adapts them into podcasts

[1] https://my-european-history.ep.eu/myhouse/allStories

or videos by interviewing the people concerned. In this way, DG TRAD is creating a virtual library of personal stories interwoven with European history. Like the story of Elza, a Slovenian Holocaust survivor who tells us about her imprisonment in Auschwitz, or Magdalena, who lived through the peaceful end of communism in Poland and the country's transition to democracy. Their testimonies offer a window into Europe's rich and diverse past and show how we are all connected through a shared European history.

2.3. Stand with Ukraine

More recently, DG TRAD has focused on the need for language services in non-EU languages, in particular Ukrainian. Russia's invasion of Ukraine has led to the expansion of language services in Ukrainian in order to provide access to information to Ukrainians about Parliament's efforts to support them. Soon after the invasion, Parliament set up the website Stand with Ukraine,[2] to which DG TRAD contributes with texts and audio podcasts in Ukrainian on an ongoing basis. In addition, DG TRAD is currently providing linguistic assistance to the Ukrainian Verkhovna Rada to help prepare for Ukraine's accession to the European Union.

The language services provided in Ukrainian include:

- translating relevant texts such as Parliament's resolutions, briefings and speeches given by President Metsola and other key political figures;
- facilitating the translation of and providing terminology support for the *acquis communautaire*;
- subtitling speeches and videos;
- producing the daily *News in Brief* podcast and other podcasts on relevant subjects.

3. Development of the translation profession

In parallel with DG TRAD's new focus on Citizens' Language, it became clear that the traditional role of the translator had to be developed and aligned with the new services DG TRAD was offering. Therefore, five new profiles were created, constituting the new family of language professionals in DG TRAD:

- Intercultural and language professionals;
- Legal language professionals;
- Proofreaders-language professionals;
- Clear language professionals;
- Innovation language professionals.

New staff are recruited into one of these language professional profiles and some of DG TRAD's existing job functions have been reassigned to the new profiles. Of course, extensive training was provided to all language professionals to help them acquire the skills needed for the tasks in their new job profiles, which are described in further detail below.

Intercultural and language professionals translate, adapt, transcreate and revise all types of content in their own language. They enable communication in their mother tongue, working from at least two other official EU languages, by means of intercultural and linguistic mediation, for example audiovisual content, subtitling and adapting. They provide drafting assistance in non-legislative matters, help with training measures, terminology work and the development of communication and IT tools. They maintain regular contacts with requesting departments, contribute to quality assurance and control processes, and optimise the quality of content delivered through following best practice.

[2] https://ukraine.europarl.europa.eu/en/home

Legal language professionals translate and revise legal texts in their own language and provide legal analysis and advice on legal terminology, while ensuring the coherence of texts throughout the translation process. Furthermore, they translate, adapt, transcreate and revise all types of content in their mother tongue, facilitate communication with citizens in plain language and provide legal advice on plain language in procedural texts. They optimise the quality of translations by following best practice. They maintain regular contacts with requesting departments and represent the department in project teams, internal and interinstitutional working groups, on professional bodies and/or at professional meetings. They also help with training measures and terminology work, as well as the development of communication and IT tools.

Proofreader - Language editors proofread texts to check spelling, grammar, syntax, punctuation, typography, formatting, technical compliance and compliance with the rules on external references, among other aspects of texts. They check texts for linguistic and terminological consistency and compliance with stylistic conventions and rules. They pre- and post-process documents using translation tools, word-processing software and other office applications, and search for existing texts, as well as incorporating them into translation memories. Proofreader - Language editors prepare draft translations of short non-legislative texts or parts thereof and assist translators and terminologists by carrying out terminology research, updating terminology databases, preparing terminology for translation in CAT tools and assist in carrying out technical and linguistic quality checks of texts. They also carry out other language-related tasks, such as transcription, checking transcribed files, checking texts in various file formats, checking and content moderation of texts on websites and social media sites. Finally, they also help with training and onboarding new colleagues, with providing training inside the department and with developing IT tools.

Clear language professionals revise, adapt, transcreate and summarise all types of content in their mother tongue and carry out linguistic editing of non-legislative texts, resolutions, questions and other documents. They provide clear language services to Parliament's writers, drafting assistance in non-legislative matters and facilitate communication with citizens in plain language in their mother tongue. They contribute to the quality assurance and control processes and optimise the quality of the content delivered by following best practice. They also help with training measures, terminology work and the development of communication and IT tools. Furthermore, they maintain regular contact with requesting departments and provide them with training and advice.

Innovation language professionals provide technological and content support aimed at ensuring multilingualism in the European Parliament. They actively keep track of technological advancements in fields relevant to their duties and propose innovative solutions benefiting all categories of language professionals. They coordinate and manage projects, including innovative and multilingual projects and procurement procedures, and participate in the development, testing and improvement of language tools and features, ensuring their effective and efficient use, and help to design and organise relevant training courses. They also devise, draw up, formalise, propose, implement and follow up on objectives and action plans. They write studies, notes, summaries and statistics, analyse, devise and prepare draft rules, and liaise with the departments involved and with their counterparts at interinstitutional level. Finally, they optimise the use of the department's resources to provide a quality service in their areas of activity.

With these new language profiles, DG TRAD has created the framework for language experts to become versatile language professionals, who are able to contribute to the production of multilingual content in text, audio and video formats. This is also a precondition for making an increased range of language services in the European Parliament available to a broader audience, including people who are deaf and hard of hearing, or blind and partially sighted.

4. Making Parliament accessible for all

4.1. Accessibility in the European Parliament

According to the World Health Organization, globally over 1.5 billion people live with hearing loss,[3] and at least 2.2 billion people have a near or distance vision impairment.[4] In the WHO's European Region, approximately 190 million people have some hearing loss or deafness[5] and over 30 million people experience visual impairment.[6] Parliament aims to make its work accessible and understandable to all EU citizens. To do so, it must go beyond ensuring multilingualism in the content Parliament produces, to provide equal access to information for people who are deaf or hard of hearing, and blind or partially sighted, both in its online communication and in the information provided on-site to visitors to Parliament's premises.

Following Parliament's principle of full multilingualism, DG TRAD approaches accessibility from a linguistic standpoint. Firstly, it makes sure that information and content is available in all 24 official EU languages. Allowing citizens to access Parliament's content in their own language is one means to increase trust, closeness and effectiveness. It implies respect for cultural and linguistic diversity, one of the cornerstones of European democracy. Beyond guaranteeing multilingualism, DG TRAD has also expanded its range of language services related to Parliament's online and on-site content in text, audio and video formats.

4.1.1. Accessibility online

Equal access to information is crucial to Parliament's online presence. In order to make Parliament's websites and online media content accessible and understandable to all citizens, DG TRAD follows the four principles of web accessibility described by the World Wide Web Consortium in its Web Content Accessibility Guidelines (WCAG),[7] the **'POUR'** principles.

Anyone who wants to use the web must have access to content that is:

1. **Perceivable:** content should be available to at least one of the user's senses. For example, images are described with an alternative text for visually impaired users.
2. **Operable**: content should be controllable with a variety of tools. For example, by using a keyboard only, for people who are not able to use a mouse.
3. **Understandable**: using clear and simple language, and predictable and consistent interfaces helps people with cognitive or reading disabilities.
4. **Robust**: the website or app must work well across different platforms, browsers and devices, including assistive technology.

Regarding accessibility online, DG TRAD's focus on formats other than text plays a crucial role. Its wide range of audio products and video subtitling services aims to provide a response to the needs of people who are deaf or hard of hearing, or blind or partially sighted. Listening to an audio summary can be an excellent alternative to reading a long text. Subtitles for hearing audiences or for the d/Deaf and hard of hearing allow people to watch Parliament's social media videos without sound and multilingual

[3] https://www.who.int/health-topics/hearing-loss#tab=tab_1
[4] https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment
[5] https://www.who.int/europe/news-room/questions-and-answers/item/ear-and-hearing-care
[6] https://www.who.int/europe/activities/supporting-member-states-to-reduce-avoidable-vision-impairment-as-a-public-health-problem
[7] https://www.w3.org/WAI/WCAG21/Understanding/intro#understanding-the-four-principles-of-accessibility

voice-overs make them perceivable to the blind or partially sighted. Learning about Parliament's activities starts online for most people, so that is where we must ensure access for all in the first instance.

4.1.2. Accessibility on-site

However, thousands of people each year also visit Parliament's premises in Brussels, Luxembourg and Strasbourg. Visitors to Parliament's various visitors' centres and museums, such as the Parlamentarium or the House of European History, obtain a unique insight into the work, functioning and history of the European institutions. Perceivable visual and auditive communication is of crucial importance for visitors with impairments. Again, this links to DG TRAD's audio services and subtitling of content, this time providing physical content available in the visitors' areas.

Currently, it is possible to make pre-prepared content perceivable in multiple formats, for example by subtitling or providing voice-overs for introductory videos about Parliament. However, DG TRAD is also working on live speech-to-text and machine translation technology that will allow speech to be transcribed and automatically translated into all 24 official EU languages (see below). This technology, which is currently being developed with the aim of making parliamentary debates available to a wider audience, is expected to have many practical uses that could be implemented elsewhere, for example, in making visitor programmes more accessible.

4.2. Content for the deaf and hard of hearing

4.2.1. Subtitling for the deaf and hard of hearing

DG TRAD subtitles a wide range of videos for the hearing audience in 24 official EU languages and in Ukrainian, demand for which is growing. These range from political speeches (e.g. speeches delivered by the President of the European Parliament, the Vice-Presidents and Members of the European Parliament), to informative content produced by the European Parliamentary Research Service, the Directorate-General for Communication or the My House of European History project, which gathers personal stories related to European History. DG TRAD also subtitles videos for social media, documentaries and the films nominated for the LUX Audience Award. The Award is presented by the European Parliament and the European Film Academy every year, in partnership with the European Commission and Europa Cinemas. The award celebrates European cinema and aims to raise awareness about social, political and cultural issues in Europe. Every year, DG TRAD subtitles the five nominated feature films in the 24 official EU languages.

Recently, DG TRAD has been making advances in terms of accessibility by producing several subtitled videos also for the deaf and hard of hearing audience. Since 2021, the subtitles of the winning films of the LUX Audience Award have been adapted in all 24 official EU languages for the deaf and hard of hearing audience, creating an equivalent viewing experience.

To tackle this task, extensive internal and external training was organised for members of the Core Team of the Subtitling and Voice-over Unit, as well as for language professionals who were tasked with subtitling for the deaf and hard of hearing audience. Furthermore, 24 detailed internal guidelines were drawn up for subtitling for the deaf and hard of hearing audience for all 24 official EU languages.

4.2.2. Real-time Speech-to-text and machine translation for 24 languages

While hearing audiences can listen to and follow parliamentary debates of the European Parliament in all 24 official EU languages, people who are deaf and hard of hearing cannot currently do this in real

time without difficulties, regardless of whether they are Members of Parliament or citizens. In order to increase the accessibility of these debates, the European Parliament entered an innovation partnership with an external partner, with the aim of developing and acquiring a prototype tool which automatically transcribes and translates multilingual parliamentary debates in all 24 official EU languages in real time. This tool will also enable Members of the European Parliament to follow debates on screen, for example in silent mode.

The real-time Speech-to-text and machine translation tool consists of several components. Firstly, the tool must identify automatically which of the 24 official EU languages the speaker is using. Secondly, the automatic speech recognition component transcribes the spoken word in real time. Lastly, the original transcript is machine-translated into any of the other 23 official EU languages.

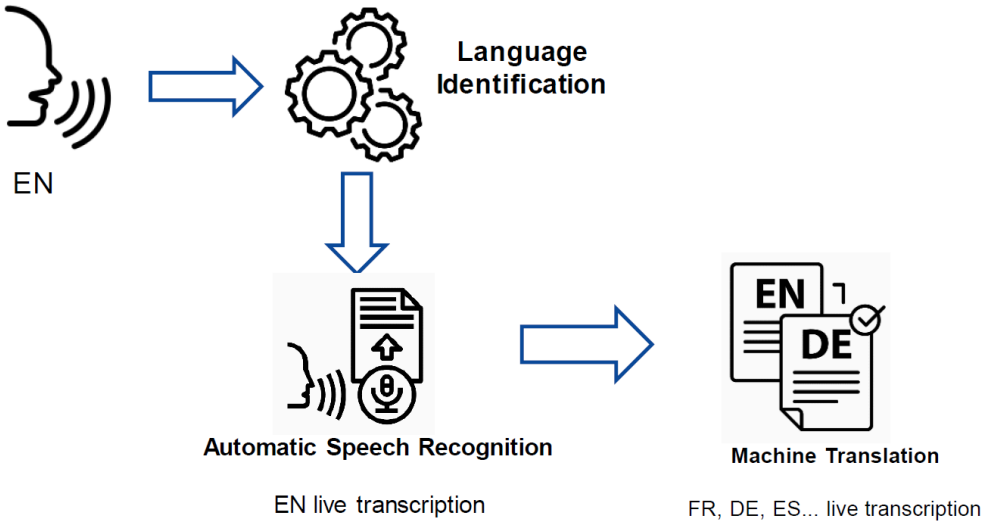


Figure 1: Components of the real-time Speech-to-text and machine translation tool
Source: European Parliament, DG TRAD, Speech-to-text Unit

The speech-to-text and machine translation project is divided into three stages. During stage 1, which started in September 2020, DG TRAD assessed the quality of the linguistic output in 10 languages (German, English, Greek, Spanish, French, Italian, Dutch, Polish, Portuguese and Romanian). In stage 2, commencing in November 2021, nine new language models were developed and successfully added to the tool (Bulgarian, Czech, Finnish, Croatian, Hungarian, Lithuanian, Slovak, Slovene and Swedish). Furthermore, the previous 10 language models were continuously improved. Stage 3 of the innovation partnership began in December 2022 and will see the addition of the remaining five language modules (Danish, Estonian, Irish, Latvian and Maltese). This stage is set to conclude by the end of 2023.

The prototype of the speech-to-text and machine translation tool was also deployed and tested in several real-life meetings and events to gather user feedback and gain further experience in actual use case scenarios.

4.3. Content for the blind and partially sighted

After consolidating its initiatives on better accessibility for people who are deaf and hard of hearing, DG TRAD is now also venturing into services tailored to the needs of people who are blind or partially sighted. This includes, in particular, services such as voice-over and audio subtitling, which enable blind and partially sighted people to access video content.

These services for people who are blind and partially sighted call for a different set of modal transformations. Images must be transformed into sounds. Both services need to be prepared differently and have their own technical and linguistic constraints. Whereas voice-overs require voice-recorded translations, audio subtitles require voice-recorded subtitles. Venturing out into different domains of audiovisual translation also creates challenges in terms of style, tone, register, synchronicity of information and taking a multi-cultural approach. Moreover, DG TRAD is developing workflows in all 24 official EU languages, which amplifies the scale and complexity of operations immensely.

5. Conclusion: taking the next steps for accessibility

Parliament's translation service continues to reconfirm its status of a world leader in translating legislation for Europe, alongside other translation services of the European institutions. Simultaneously, it is also on its way to becoming a world leader in accessibility through a broad range of multilingual audiovisual language services. It is gradually building on its decade-long experience in complex translation processes and on an unparalleled wealth of in-house skills and knowledge.

DG TRAD is no stranger to adapting to change. In recent years, the service has undergone a transformation into a modern language service provider, offering a wide range of localisation and audiovisual translation services. Translators and translation assistants have diversified their skills and activities through extensive on-the-job training. They have morphed into versatile language professionals who are present in all production steps of Parliament's multilingual content through translation, localisation, transcreation and adaptation, using modern and sophisticated technology, which is often homegrown.

This process has filled us with confidence for the future. Developing new services to improve the accessibility of Parliament's content is likely to give rise to challenges, but our solid foundations and our experience in change management will help us overcome any obstacles that may arise.

Past, present and future of speech technologies in translation — life beyond the keyboard

Julián Zapata
Toronto Metropolitan University
350 Victoria St, M5B 2K3
Toronto, Canada
julian.zapata@torontomu.ca

Alina Secară
University of Vienna
Universitätsring 1, 1010 Wien
Vienna, Austria
alina.secara@univie.ac.at

Dragoș Ciobanu
University of Vienna
Universitätsring 1, 1010 Wien
Vienna, Austria
dragos.ioan.ciobanu@univie.ac.at

Abstract

The idea of speaking to and with computers is as old as the idea of computers themselves. Today, after more than eight long decades of research and development in computing and natural language processing, such an idea is far from science fiction: systems that allow humans and computers to interact directly through speech are increasingly becoming part of our daily lives and are transforming the nature of human-computer interaction. Speech technologies have reached a point of maturity to be useful and effective in several domains. They can notably be found in customer and tech support services; virtual assistants, in-vehicle navigation systems; as well as the operating systems of personal computers, smartphones and tablets. In translation, certain researchers, trainers and practitioners are increasingly showing great interest in the use of speech technologies for translation purposes. Recent work has shown that both speech recognition and synthesis positively influence the output quality, language professionals' productivity, and workspace ergonomics associated with translation, revision, machine translation post-editing, audiovisual translation, as well as interpreting processes. This paper presents a brief history of speech technologies in translation and reports on a panel held within the framework of the 44th Translating and the Computer (TC44) conference. The panel members represented both academic and international organisations, and discussed some of the myths, successes, challenges, and opportunities of working with speech recognition and synthesis in translation research, teaching and practice.

KEYWORDS: speech technologies, speech recognition, speech synthesis, translation dictation, translation technologies, human-computer interaction

Introduction

Humans have used tools since time immemorial: to light fires, hunt, eat, build, travel, write and communicate. In translation, scholars rarely examine the use of tools throughout the centuries, although it is argued, sometimes satirically, that translation is one of the oldest professions in the world. Indeed, the history of translation goes back to the development of the human capacity to communicate; consequently, written translation can safely be said to have emerged in parallel to the invention of alphabets, writing systems and tools. Throughout history, translators have adopted different tools with the constant advances in science and technology: from stone-engraving tools, papyri and ink pens to typewriters and personal computers.

However, the physical act of writing is not very satisfying in the eyes of many translators, or even seasoned writers; they want their hands to move at the speed of their thoughts. In fact, either by preference or for health reasons, some writers and translators have opted for dictation, that is, to speak their texts into a recording device (also known as a Dictaphone) for subsequent transcription, or directly to a professional transcriptionist (Gingold, 1978; Hétu, 2012; Jiménez Ivars and Hurtado Albir, 2003; Jurafsky and Martin, 2009, 285; Laroque-Divirgilio, 1981). However, translation dictation, which was very common in the 1960s and 1970s, is far from being the norm in the translation industry today.

Nowadays, however, translators use a wide range of tools that have arrived with the enormous progress in computing and natural language processing (NLP). These tools, known

as computer-aided translation (CAT) tools, assist the translator in their work because they are, on the one hand, able to automate certain linguistic and technical tasks, and on the other hand, because they facilitate access to the information the translator needs to produce their translations.

Outside the translation sphere, a growing number of computer developments aim to explore alternatives to traditional input methods such as the keyboard and mouse; the latter have characterised the interaction of translators, and writers in general, with personal computers since the 1980s. Emerging interaction modes include speech technologies, which fundamentally change the way humans interact with machines to access, produce and use information. The quality of speech technologies is improving so fast for certain languages that the latest report published by the Language in the Human-Machine Era (LITHME) COST (2021, 6) action highlights "two imminent changes to human communication […]: speaking through technology and speaking to technology." In situations where two-way communication is not necessary, however, speech technologies have already been implemented to optimise monolingual and multilingual content production workflows: for over a decade, human-to-machine dictation with speech recognition (SR) (speech-to-text) has been the preferred mode of creating content of language professionals whose technological set-up allowed this kind of enhancement or who work in live contexts, such as broadcasting, where immediate access to the text produced is crucial.

In addition, speech synthesis (SS) (text-to-speech) has also been gaining ground in recent years. Research has shown that both SR and SS can positively influence the output quality, language professionals' productivity, and workspace ergonomics associated with translation, revision, machine translation post-editing (MTPE), audiovisual translation, as well as interpreting processes. Despite these demonstrated benefits, technology providers have for unknown reasons trailed behind in implementing speech technologies within current CAT/MTPE environments, although progress in enhancing subtitling, computer-aided interpreting, and even complete speech-to-speech translation tools has been more substantial. Moreover, there is some evidence that speech technologies are finding a place in translation training and research, particularly in respeaking (also called live subtitling/captioning or speech-to-text interpreting to produce live subtitles/captions via SR tools) in accessibility-related scenarios.

In this paper, we will discuss the past, present and future of speech technologies in translation. After a literature review of the field, we will report on a panel held within the framework of the Translating and the Computer (TC44) conference that took place on November 25, 2022, at the European Convention Centre in Luxembourg. The panel members represented both academic and international organisations and discussed, using a variety of examples from their academic or professional experiences, some of the myths, successes, challenges and opportunities of working with speech technologies in translation research, teaching and practice.

Speech technologies and translation: long-time allies

As stated above, speech technologies include SR and SS. In short, SR enables a computer system to recognise and process human speech, while SS uses acoustic models to emulate it.

Speech technologies are among the major developments in the NLP field, together with word processing, parsing (or syntactic analysis of texts), machine translation (MT), indexing or automatic term extraction. However, after decades of research in this area, speech technologies constitute the most salient element of these applications. In the early days of NLP, the dominant assumption within the research community was that prior knowledge of language would have enabled computers to proceed to the next stages of learning; a learning process

comparable to that of humans from early childhood. According to Alan Turing (1950, 460), one of the pioneers of computer science, computers should be able to process human language first before trying to achieve artificial intelligence.

As a result of the ideas put forward by Turing, several experiments took place to process natural languages and automate certain tasks: text processing and storage, MT and the design of conversational agents, i.e., systems with which the user can establish a conversation. Some conversational agents also integrated SR and SS: they could automatically "recognise" what a person was saying using SR and respond to them by emulating a human voice using SS (Llisterri, 2009, 11–12). However, advances in SS were said to exceed those in SR: while recognition systems could only process isolated words spoken by a single speaker, synthesis systems were already at the refinement stage, seeking to emulate human intonation. The great challenges of SR were then the analysis of continuous speech, due to the great variability that speech can present depending on the speaker, in addition to the accents and the multiple possible voice timbres.

From the early 1990s, new developments focused on conversational agents which also had the ability to translate, that is to say, equipped with an MT system; in other words, artificial interpreters. Furthermore, during the same decade, many labs invested in the development of SR systems while adding the possibility of issuing voice commands to the computer. However, despite these significant advances, the systems developed were functional only in specific professional fields, with limited vocabularies, and in noise-free environments.

Around the mid-1990s, research efforts for the adaptation of SR to human translation were first made. Several studies have gone beyond artificial interpreters. In designing an SR tool that can help a human translator, the focus was on reducing recognition error rates by coupling MT and SR. In other words, the translator dictates their translation to a hybrid MT+SR system (Brousseau et al., 1995; Brown et al. 1994; Dymetman et al., 1994). Such a system has access to the source text and uses probabilistic MT models to improve recognition.

Although advances in speech technologies in general were satisfying for some researchers at the time, especially in the field of telecommunications, others still could not see what Turing imagined in 1950: that humans and computers could process speech the same way. In translation, efforts to integrate SR into the translators' toolbox did not awaken the same interest among researchers, trainers and practitioners as other NLP applications. Indeed, research stalled for SR, but continued for the design of other tools, capable of supporting other peripheral tasks, allowing translators to achieve increased efficiency ─ CAT tools. In short, at the end of the 1990s, SR was not yet powerful enough to automate language tasks, including the transcription of dictations.

That said, at the beginning of this century, commercial SR systems gradually began to be part of the translator's toolkit, although translation was not their main field of application; in other words, SR software was not designed specifically for translation tasks (Gouadec, 2002, 133). In any case, some translators integrated off-the-shelf SR programs into their toolbox to dictate translations and to issue commands to their computers (Benis, 2002; Seaman, 2002; Stroman, 2002). Benis (2002), for example, addresses the issue of using SR to dictate translations; the main difficulties that he exposes are linked not only to recognition error rates and limited computational power of machines at that time, but also to the lack of dictation skills on the part of the users. However, his testimony, as well as that of his contemporaries, is sprinkled with positive comments towards this technology.

Research in SR quickly picked up momentum. During the 2000s, significant progress was made in the optimisation of SR systems: the improvement of accuracy rates; the creation of user-specific profiles; the adaptation of the technology to certain professional fields such as medical, legal and law enforcement; the addition of more voice commands, etc. These improvements convinced some translation researchers of the relevance of exploring the

advantages of translating using these systems. In the second half of the decade, several studies were conducted on the subject. Désilets et al. (2008), for example, conducted an experiment to assess productivity gains among translators in Canada who used a hybrid MT+SR system. These researchers were quite optimistic about SR and called for more research in the field. Further experiments took place in other research centres and their results point to the advantage of dictating translations using SR to gain productivity (Dragsted, Hansen and Sørensen 2009; Reddy and Rose 2010). In addition, a survey was conducted in 2009 among participants of the International Annual Meeting on Computer-Assisted Translation and Terminology (JIAMCATT) bringing together representatives of major translation services within international organisations. The survey sought to determine interest in adopting digital recording devices and SR software within their organisations. The survey suggested that the number of translation services using SR software was not negligible and that an innovative approach should be considered: teaming translators dictating translations using digital dictaphones with copyists transcribing recordings using SR software (Verástegui, 2009).

The history of speech technologies spans decades of research and development. That said, it is only recently that interest in translation research focused on these technologies has *truly* begun to awaken (Ciobanu, 2014, 2016; Ciobanu et al., 2019; Ciobanu and Secară, 2019; Garcia-Martinez et al., 2014; Herbig et al., 2020; Liyanapathirana and Bouillon, 2021; Liyanapathirana et al., 2022; Mees et al., 2013; Mesa-Lao, 2014; Teixeira et al., 2019; Wiesinger et al., 2022; Zapata, 2012; Zapata and Kirkedal, 2015; Zapata et al., 2017), in light of both the promising results of studies conducted over the past twenty years, examples of the successful use of these systems in various other fields, the increasing performance of this technology coupled with the multiplied processing capacity of computers, as well as the indisputable need to design ergonomic translation tools, that is, taking into account the human factor (O'Brien, 2012).

Speech technologies panel at TC44

The panel "Past, present and future of speech technologies in translation" took place on November 25, 2022, at the 44th Translating and the Computer (TC44) conference. It lasted for 1 1/2 hours and was moderated by Dragoș Ciobanu (DC) who introduced the topics, offered a demo of a range of speech technology applications, directed questions to the panellists and moderated the Q&A sessions, which concluded the panel. The panellists Marcin Feder (MF), Alina Secară (AS), Carlos Teixeira (CT) and Julián Zapata (JZ) offered academic, institutional and commercial perspectives regarding the use and implementation of speech technologies in translation practice, training and research.

Panelists' and moderator's profile

Marcin Feder has a PhD in linguistics (English/Computer Assisted Translation). Since 2019, he is the Head of the Speech-to-text Unit in DG TRAD, which is developing a live speech-to-text and MT tool that is able to automatically transcribe and translate multilingual parliamentary debates in real time.

Alina Secară has a PhD in audiovisual translation, is an accredited Stagetext theatre captioner and Senior Scientist in the University of Vienna Centre for Translation Studies, where she investigates accessibility practices and technologies, and teaches modules related to accessibility and audiovisual translation, as well as multimedia localisation processes and technologies.

Carlos Teixeira has a PhD in translation and intercultural studies and is an expert in translation technologies and processes. He works as a localisation engineer for IOTA

Localisation Services, Dublin, helping to optimise the translation workflows for their high-profile customers mainly in the software industry. He also teaches at the Masters in Professional Translation at Universitat Rovira i Virgili, Tarragona, Spain.

Julián Zapata has a PhD in translation studies and is Assistant Professor of Translation at the Toronto Metropolitan University. He is also a certified translator and an entrepreneur.

Dragoș Ciobanu is Professor of Computational Terminology and Machine Translation in the University of Vienna Centre for Translation Studies, where he leads the HAITrans research group. He collaborates with language service providers, international organisations, and freelance linguists to investigate ways to improve localisation workflows by integrating translation and speech technologies.

Structure of the Panel

The Panel started with a demo showcasing some SR and SS applications in the translation workflow. DC presented applications such as Voice Typing in Google Docs, which supports an extensive number of languages for dictation, the Dragon NaturallySpeaking SR software package, the Trados TTS Plugin deploying the Microsoft Azure Text-to-Speech solution to enable those translating in Trados to listen to the audio of a segment (source or target), the Hey memoQ app, which relies on Apple's SR technology to enable dictation in memoQ in over 30 languages, as well as the latest SR and SS functionalities integrated in Matecat. After this brief demo, the Panel focused on the following questions:

- Why are we only having a panel on speech technologies at a translation technology conference now, in 2022?
- In your area, where are we with the use of speech technologies and why are we there?
- What should the priorities now be in terms of integrating speech technologies in research and training?

We compiled the various answers to these and subsequent questions to provide a summary of them in the following sections.

Relevance of speech technologies for the language professions

All panellists agreed that SR and SS are highly relevant to the tasks translators (also known as language professionals or linguists) are facing, and that several factors have led to these technologies gaining more visibility.

First, JZ commented that the nature of human-computer interaction is changing. Linguists no longer interact with computers in the same way they did ten, twenty or thirty years ago. In the "ubiquitous computing era" technology is everywhere and uses such as voice notes or speech-supported interactions with devices are frequent. The mechanical keyboard is becoming an optional hardware for computing devices sold off-the-shelf. CT followed by reminding participants that SR and SS technology has only recently become mature enough to be seriously explored and therefore the interest we are seeing today is understandable. He mentioned his first experience with SR technology in the early 2000s, when IBM ViaVoice required extensive training and did not produce acceptable results, while Dragon NaturallySpeaking was only starting to show potential and was available only in a few languages.

In addition to the technology becoming mature enough, legal requirements and volume are additional factors influencing the use and visibility of SS and SR today. AS highlighted that recent legal changes such as The European Accessibility Act (Directive 2019/882), in addition to various national legislation implementations, requiring services to be accessible for persons with disabilities led to an increase in the need to provide audiovisual translation services such

as monolingual subtitling and audio description. This growth could be met by deploying SR and SS to support linguists delivering those services. MF added that, to the pure legal requirement, the institutions recognised an opportunity to use these technologies to enable the regular EU citizen to better access content related to the activities and decisions of the European institutions.

Current speech technologies implementation

CT offered insight into research in the field of multimodal interaction which includes speech input in combination with other emerging interactive modes such as touch and stylus. He presented a usability study he led in 2018 to test the assumption that we could eliminate the mouse and keyboard as main interaction modes. In his experiment, he asked translators to perform different tasks on a touch-and-speech activated interface using a research-level MTPE interface developed at the Adapt research centre in Dublin.

JZ mentioned that he has focused in the past few years on developing an introductory training course for students and professionals on interactive translation dictation. The course focuses on developing sight translation and dictation skills, exploring available free or paid SR systems, and learning about multimodal interaction. He also mentioned that over 100 students in Canada and France have participated in the tests and improvements of the training modules over a period of 2 years, and that the course continues to evolve with more exercises and language combinations coming up.

AS discussed the successful and wide implementation of SR in live television respeaking, where a 98% accuracy threshold is usually required and can already be achieved. Speech technologies are integrated via speech detection mechanisms for timing functionalities in subtitling environments, too. As far as SS is concerned, AS offered evidence from research carried out at her institution that SS can lead to positive results when used in revision and MTPE contexts. This technology can also be employed in the delivery of voiceover and audio description, and the reception by the public of these is fairly good.

MF mentioned that his team is working on the implementation of a system that combines SR and MT. The tool—which had been shown earlier during the TC conference during a keynote speech—is able to automatically transcribe and translate multilingual parliamentary debates in real time at the European Parliament in most EU languages, the aim being to cover all the official EU languages by the end of the project.

Desired research and training priorities

JZ commented that it is imperative that we do not start by training people on speech technologies directly, but rather train them on speaking fluently and in complete sentences, on performing sight translation and other types of text-production tasks with their voice, and on explaining the pros and cons of the different approaches to SR and dictation tools in different contexts and situations. As a response to this, one member of the audience commented that it is necessary to train people to control their own stress levels to improve the recognition rate: the more natural and fluent the speech, the better the accuracy of the transcription by the SR system.

AS noted that the amount of SR research and training within audiovisual translation is significant, and that for future SS and SR research and training, combining interpreting studies with translation studies would make sense. For example, integrating voice coaching for translators using SR: like in interpreting, your voice becomes your tool. Students should also be exposed to SR and SS tools during their translation and interpreting training, so that they have an opportunity to practise and identify professional scenarios where using such tools would have a positive impact on their work. She added that considerations about the health and well-being of linguists should be a priority, and that researchers need to further explore the

potential of speech technologies to provide linguists with more ergonomic alternatives to the conventional keyboard-and-mouse computing devices.

For his part, MF noted that while they focus on recognition of somewhat spontaneous speech (not controlled dictation environments), the quality of the output depends on the quality of input signals. Their system has no major issues with accents or regional dialects, but issues could arise if the quality of the original signal is suboptimal. CT added that working in an office with other people might be a limitation to the use of dictation in translation, and suggested the possibility of developing a mouthpiece that would prevent the sound from being heard by other parties, e.g., when sensitive material is being translated.

Giving a voice to panel attendees

Towards the end of the time allocated for the panel, DC also opened the floor to questions or comments from the audience.

One audience member suggested further discussion of the benefits of these technologies and the need for dictation. As a response to this, CT reminded that translators have always had to adapt to the evolution of technologies, and that time and effort to adapt were also required with the introduction of new tools (Dictaphones, CAT, MT, etc.); it will be necessary to learn about these tools but also to dictate translations. He also pointed out, however, that dictation is not for everyone, nor for every task. For example, software or web localisation segments with tags might not be an ideal use case. On the other hand, in translation of marketing material or other types of creative-text translation, dictation would be ideal. In short, concludes CT, even for the same translator it may not always be useful, but it should be an option and should be used whenever it makes sense and is possible. DC mentioned that in questionnaire studies, participants noted that in texts that require a higher level of informality, using dictation could be a tool supporting this stylistic requirement. Moreover, in his studies it became clear that there was a difference between novice and expert translators in terms of productivity to be gained from dictation. The translators who reported the biggest change in productivity were those who were experts in a specific field and who also had years of experience gradually building up their use of SR tools. MF noted that research has observed that the average speaking rate is 120–150 words/min, whereas the average typing rate is 40–50 words/min. AS also agreed that speech technologies are a solution in contexts where speed is key, for example in the creation of live captioning, but that they do not work for every person or context.

Another panel attendee commented that a colleague from the European Parliament used to dictate in the English to French combination and confirmed that translating texts with tags using this method was challenging. However, the biggest challenge for a user without training on these technologies is to identify and address the typical errors by the tool, since “the tool made mistakes [my colleague] wouldn’t have made as a translator. Mistakes from dictation were completely different from human mistakes, and [my colleague] would miss them.” The same attendee then asked if there is any training to identify typical SR errors like we have now for neural MT errors. JZ noted that this is precisely one important aspect of the training under development: identifying common errors and learning how to avoid them, since most of the errors are a lack of dictation skills and efficient preparation, or poor understanding of how machines process speech.

Another member of the audience suggested that further research into the impact of using speech technologies on the translation performance would be useful. For example, one aspect which would be worth investigating is if the potential gains in creativity and the reported positive effect on productivity are commensurate with the emotional and cognitive effect on the translator. DC noted that such a hypothesis is currently being researched by his team. He also highlighted one area on which these tools can have a positive effect straight away, and

which should not be underestimated, is ergonomics. One can get up and move or look away from the screen when using speech tools. Another member of the audience agreed and added that in other fields the relationship between movement and creativity has been demonstrated.

In conclusion, it was agreed that the development and use of speech tools are rapidly growing, but that they do not offer a one-solution-fits-all. It is down to the individuals to assess their skills and suitability of applying these tools in their own context.

Conclusion

The history of speech technologies proves that, despite the significant improvements in SR and SS systems over the years, the integration of these technologies into professional translation has not experienced definite success. However, they are currently reaching such a level of performance that it will be necessary to grant them considerable importance in new research and tool-development efforts. Some trainers and researchers even see it as the future of translator-computer interaction. These technologies have proved to be effective in several professional fields and in various daily life situations and constitute a promising approach in current efforts to develop translation tools that are more efficient and ergonomic. Indeed, speech technologies introduce certain elements that other technological applications have ignored in the past, one of these being the consideration of the human factor, that is, of the translator and their professional needs.

The challenges are still very numerous and interdisciplinary research would be more than desirable. One thing seems sure and certain: with the arrival of touch screens, mobile devices and cloud computing, keyboard-and-mouse computers are beginning to gradually fade away. It is time to bring translation dictation back; to reinvent translation training based on oral translation techniques and emerging interactive technologies; to design technological tools with a human dimension. Speech is natural for humans. Typing is not.

Acknowledgments

We would like to thank the Office of the Dean of Arts, Toronto Metropolitan University, for their generous support through a travel grant provided to the first author, allowing him to lead the organisation of, and travel to, the panel reported in this paper. Our heartfelt thanks also to the co-panellists, as well as the TC44 organisers for their enthusiasm in supporting this panel and their hard work in making this conference possible.

References

Benis, Michael. 2002. “Softly Spoken or Hard of Hearing?” *Language International* 14 (3): 26–29.

Brousseau, Julie, Caroline Drouin, George Foster, Pierre Isabelle, Roland Kuhn, Yves Normandin, and Pierre Plamondon. 1995. “French Speech Recognition in an Automatic Dictation System for Translators: The TransTalk Project.” In *Proceedings of Eurospeech’95*. http://www.iro.umontreal.ca/~foster/papers/ttalk-eurospeech95.pdf.

Brown, Peter, Stanley Chen, Stephen Della Pietra, Vincent Della Pietra, Andrew Kehler, and Robert Mercer. 1994. “Automatic Speech Recognition in Machine-Aided Translation.” *Computer Speech and Language* 8 (3): 177–87.

Ciobanu, Dragoș. 2014. “Of Dragons and Speech Recognition Wizards and Apprentices.” *Revista Tradumàtica* 12: 524–38.

Ciobanu, Dragoş, 2016. “Automatic Speech Recognition in the Professonal Translation Process.” *Translation Spaces* 5 (1): 124–44.

Ciobanu, Dragoş, Valentina Ragni, and Alina Secară. 2019. “Speech Synthesis in the Translation Revision Process: Evidence from Error Analysis, Questionnaire, and Eye-Tracking.” *Informatics* 6 (4): 51. https://doi.org/10.3390/informatics6040051.

Ciobanu, Dragoș, and Alina Secară. 2019. “Speech Recognition and Synthesis Technologies in the Translation Workflow.” In *The Routledge Handbook of Translation and Technology*, edited by Minako O’Hagan, 91–106. Abingdon: Routledge. https://doi.org/10.4324/9781315311258-6.

Désilets, Alain, Marta Stojanovic, Jean-François Lapointe, Rick Rose, and Aarthi Reddy. 2008. “Evaluating Productivity Gains of Hybrid ASR-MT Systems for Translation Dictation.” http://www.mt-archive.info/IWSLT-2008-Desilets.pdf.

Dragsted, Barbara, Inge Gorm Hansen, and Henrik Selsøe Sørensen. 2009. “Experts Exposed.” *Copenhaguen Studies in Language* 38: 293–317.

Dymetman, Marc, Julie Brousseau, George Foster, Pierre Isabelle, Yves Normandin, and Pierre Plamondon. 1994. “Towards an Automatic Dictation System for Translators: The TransTalk Project.” In *Fourth European Conference on Speech Communication and Technology*, 4.

European Union. European Parliament and the Council of the European Union. *European Accessibility Act*. Directive 2019/882. 2019. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019L0882.

Garcia-Martinez, Mercedes, Karan Singla, Aniruddha Tammewar, Bartolomé Mesa-Lao, Ankita Thakur, Anusuya M. A., Michael Carl, and Srinivas Bangalore. 2014. “SEECAT: ASR & Eye-Tracking Enabled Computer-Assisted Translation.” In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, 81–88.

Gingold, Kurt. 1978. “The Use of Dictation Equipment in Translation.” In *La traduction, une profession. Actes du VIIIe Congrès mondial de la Fédération internationale des traducteurs*, edited by Paul A. Horguelin, 444–48. Ottawa: Conseil des traducteurs et interprètes du Canada.

Gouadec, Daniel. 2002. *Profession : Traducteur*. Paris: La Maison du dictionnaire.

Herbig, Nico, Tim Düwel, Santanu Pal, Kalliopi Meladaki, Mahsa Monshizadeh, Antonio Krüger, and Josef van Genabith. 2020. “MMPE: A Multi-Modal Interface for Post-Editing Machine Translation.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1691–1702. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.155.

Hétu, Marie-Pierre. 2012. “Le travail au dictaphone, une solution ergonomique?” *Circuit* 116 (été 2012): 23.

Jiménez Ivars, Amparo, and Amparo Hurtado Albir. 2003. “Variedades de Traducción a La Vista. Definición y Clasificación.” *Trans Revista de Traductología* 7: 47–57. http://www.trans.uma.es/trans_07.html.

Jurafsky, Daniel, and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2nd ed. Upper Saddle River, New Jersey: Pearson Prentice Hall.

Laroque-Divirgilio, Lise. 1981. “La Traduction Au Magnétophone.” *Meta* 26 (4): 398–403. http://www.erudit.org/revue/meta/1981/v26/n4/002573ar.pdf.

LITHME. 2021. “The Dawn of the Human-Machine Era. A Forecast of New and Emerging Language Technologies.” 2021. https://lithme.eu/2021/05/18/first-publication-forecast-of-future-language-technologies/.

Liyanapathirana, Jeevanthi, and Pierrette Bouillon. 2021. “Integrating Post-Editing with Dragon Speech Recognizer: A Use Case in an International Organization.” In *Translating and the Computer 43*, 55–67.

Liyanapathirana, Jeevanthi, Pierrette Bouillon, Jonathan David Mutal, and Lise Volkart. 2022. “Integrating Speech in Post-Editing (PE)-Comparison of Two PE Interfaces.” In *New Trends in Translation and Technology (NeTTT) Rhodes Island, Greece,* 120–23. Rhodes Island, Greece.

Llisterri, Joaquim. 2009. “Las Tecnologías Del Habla En Las Lenguas Románicas Ibéricas.” *Studies in Hispanic and Lusophone Linguistics* 2 (1): 133–80.

Mees, Inger M., Barbara Dragsted, Inge Gorm Hansen, and Arnt Lykke Jakobsen. 2013. “Sound Effects in Translation.” *Target* 25 (1): 140–54. http://openurl.ingenta.com/content/xref?genre=article&issn=0924-1884&volume=25&issue=1&spage=140.

Mesa-Lao, Bartolomé. 2014. “Speech-Enabled Computer-Aided Translation: A Satisfaction Survey with Post-Editor Trainees.” In *Workshop on Humans and Computer-Assisted Translation*, 99–103.

O’Brien, Sharon. 2012. “Translation as Human–Computer Interaction.” *Translation Spaces* 1 (1): 101–22. https://doi.org/10.1075/ts.1.05obr.

Reddy, Aarthi, and Richard C. Rose. 2010. “Integration of Statistical Models for Dictation of Document Translations in a Machine Aided Human Translation Task.” *IEEE Transactions on Audio, Speech and Language Processing* 18 (8): 1–11. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=05393062.

Seaman, Lee. 2002. “Voice Recognition for Translators, Or Why I Started Talking To My Computer.” *JLD Times. Newsletter of the Japanese Language Division of the American Translators Association*, 2002.

Stroman, John. 2002. “Translation and Voice Recognition Software.” *JLD Times. Newsletter of the Japanese Language Division of the American Translators Association*, 2002.

Teixeira, Carlos, Joss Moorkens, Daniel Turner, Joris Vreeke, and Andy Way. 2019. “Creating a Multimodal Translation Tool and Testing Machine Translation Integration Using Touch and Voice.” *Informatics* 6 (1): 13. https://doi.org/10.3390/informatics6010013.

Turing, Alan. 1950. “Computing Machinery and Intelligence.” *Mind* 25 (236): 433–60. http://mind.oxfordjournals.org/content/LIX/236/433.

Verástegui, Nelson. 2009. “Digital Recording Survey.”

Wiesinger, Claudia Karin, Justus Brockmann, Alina Secară, and Dragoș Ciobanu. 2022. “Speech-Enabled Machine Translation Post-Editing in the Context of Translator Training.” In *Contextuality in Translation and Interpreting: Selected Papers from the Łódź-ZHAW Duo Colloquium on Translation and Meaning 2020–2021. Łódź Studies in Language*, edited by Michał Kornacki and Garry Massey, 67–89.

Zapata, Julián. 2012. “Traduction dictée interactive : Intégrer la reconnaissance vocale à l’enseignement et à la pratique de la traduction professionnelle.” M.A. thesis, University of Ottawa. http://www.ruor.uottawa.ca/en/bitstream/handle/10393/23227/Zapata Rojas_Julian_2012_these.pdf?sequence=1.

Zapata, Julian, Sheila Castilho, and Joss Moorkens. 2017. “Translation Dictation vs. Post-Editing with Cloud-Based Voice Recognition: A Pilot Experiment.” In *Proceedings of MT Summit XVI, vol. 2: Users and Translators Track*, 123–36. https://www.researchgate.net/publication/319853140.

Zapata, Julián, and Andreas Søeborg Kirkedal. 2015. “Assessing the Performance of Automatic Speech Recognition Systems When Used by Native and Non-Native Speakers of Three Major Languages in Dictation Workflows.” In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, 201–10.

Section B: Machine translation and users

Impact of Domain-Adapted Multilingual Neural Machine Translation in the Medical Domain

Miguel Rios, Raluca-Maria Chereji, Alina Secară, Dragoș Ciobanu

Centre for Translation Studies, University of Vienna

{miguel.angel.rios.gaona, raluca-maria.chereji,
alina.secara, dragos.ioan.ciobanu}@univie.ac.at

Abstract

Multilingual Neural Machine Translation (MNMT) models leverage many language pairs during training to improve translation quality for low-resource languages by transferring knowledge from high-resource languages. We study the quality of a domain-adapted MNMT model in the medical domain for English-Romanian with automatic metrics and a human error typology annotation which includes terminology-specific error categories. We compare the out-of-domain MNMT with the in-domain adapted MNMT. The in-domain MNMT model outperforms the out-of-domain MNMT in all measured automatic metrics and produces fewer terminology errors.

1 Introduction

Current state-of-the-art Neural Machine Translation (NMT) models have shown promising results on low-resource language pairs, particularly for non-specialised domains (Araabi & Monz, 2020). However, in a high-risk and low-resource domain, like the medical domain, the accurate translation of terminology is crucial for exchanging information across international healthcare providers or users (Skianis et al., 2020). Multilingual NMT (MNMT) models leverage many language pairs and millions of segments during training (Johnson et al., 2017). The inclusion of many language pairs helps to improve the translation quality for low-resource languages by transferring knowledge from high-resource languages. Moreover, domain adaptation techniques are used to adapt MNMT models into new domains (Bérard et al., 2020). However, evaluation studies of MNMT models are focused on automatic metrics without providing insights into the quality of the translation of specialised terminology.

In this paper, we study the quality of a pre-trained MNMT model in the medical domain for a low-resource language pair (English-Romanian). Our goal is to compare an out-of-domain MNMT with a fine-tuned in-domain MNMT in terms of automatic metrics and terminology translation. We use a pre-trained model based on MBart (Liu et al., 2020) and fine-tune it with a medical in-domain parallel corpus.

We test the models on the English-Romanian language pair with a corpus of medical paper abstracts (Neves et al., 2018). We evaluate both models with automatic metrics, and a terminology error typology annotation performed by in-house human annotators (Haque et al., 2019). The fine-tuned MBart model outperforms MBart on the automatic metrics. In addition,

the error analysis based on a terminology-based error typology (Haque et al., 2019) shows that the fine-tuned model also produces fewer errors than the MBart model.

2 Background

MNMT models are based on transferring parameters or information across multiple languages, where low-resource languages benefit from the high-resource languages. The MNMT model shares a common word representation (i.e., arrays of numbers) across language pairs. During training, the MNMT model clusters words with similar contexts from the high- and low-resource segments (Johnson et al., 2017). The low-resource pairs learn meaningful word representations given the access to a large number of similar contexts from the high-resource language pairs. Moreover, an MNMT model allows to translate across multiple languages by using only one translation system. The multiple languages are processed jointly by indicating the target translation direction on each segment of the multilingual corpora in the input training data by using an artificial token (label *<2target>*). For example, an English-Romanian segment pair would be labelled as follows:

<2ro> *It is noted that in some cases increase of blood pressure was documented.* → *Se remarcă faptul că, în unele cazuri, s-a înregistrat creșterea tensiunii arteriale.*

MNMT models outperform standard bilingual baselines on translation quality for low-resource languages (Johnson et al., 2017), but they require a high amount of computational resources to process the millions of parallel multilingual segments. In particular, MBart is a sequence-to-sequence model pre-trained on monolingual data from 25 languages based on a text reconstruction learning objective for MNMT (Liu et al., 2020). MBart incorporates a monolingual training step before the multilingual MT training for a better initialisation of the translation model. In other words, MBart first learns an improved individual representation of each language with monolingual data. After that, MBart continues with the multilingual translation training based on parallel data. MBart shows a better translation quality compared to previous MNMT models.

However, most MNMT models are general-purpose systems trained with web crawled corpora (Liu et al., 2020; Verma et al., 2022), and they struggle with specialised domains (e.g. medical). Domain adaptation aims to improve the translation performance in specialised domains, where fine-tuning is a low-cost and common technique. Fine-tuning consists of resuming the training of an out-of-domain resource-rich MT model with a poor-resourced in-domain corpus (Chu & Wang, 2018). The resulting model is adapted to work with an in-domain language pair, instead of re-training the MNMT model from scratch (Verma et al., 2022).

MT models are usually evaluated with automatic metrics that take into account fluency and adequacy, by comparing the machine translation output against one or more human reference translations (Papineni et al., 2002). Metrics produce a corpus-level score or a segment-level score for a given MT model (Rei et al., 2020). However, automatic metrics are not designed to identify translation errors in MT outputs, for example, errors in terminology (Haque et al., 2019). On the other hand, error typology evaluation frameworks, such as the Multidimensional

Quality Metrics (MQM) (Lommel et al., 2013), are based on manually classifying and annotating errors using predefined categories. Haque et al. (2019) propose an error typology with a focus on terminology: human evaluators identify an error in the MT output, select a category out of the eight available, and assign a severity score.

3 Experiments

For fine-tuning, we use the English-Romanian section from the EMEA parallel corpus (CLARIN:EL, 2015). The EMEA corpus consists of PDF documents from the European Medicines Agency. We split the corpus into 775,904 training, and 7,837 validation segments. We evaluate the MNMT models with the test dataset of abstracts from scientific publications from Medline (Neves et al., 2018) which contains 291 segments.

We use BLEU (Papineni et al., 2002; Post, 2018), chrF (Popović, 2015), and COMET (Rei et al., 2020) for automatic evaluation. For human evaluation, we use Haque et al. (2019) which contains eight terminology-related error categories - Partial error, Source term copied, Inflectional error, Reorder error, Disambiguation issue in target, Incorrect lexical selection, Term drop, and Other error -, and three severity levels - Minor, Major and Critical.

We continue training MBart with the EMEA corpus to adapt it into the medical domain, and we perform model selection using BLEU on the validation split. We performed our experiments with Fairseq (Ott et al., 2019) using an open-source pre-trained model for MBart.[1] The settings for the fine-tuned MBart are as follows: Adam with learning rate 3e-5, inverse square root scheduler, 2,500 warm-up updates, 40,000 updates, dropout 0.3, attention dropout 0.1, label smoothing 0.2, batch size 2048 tokens (256 maximum tokens per batch, and 8 batches for gradient accumulation), and memory efficient fp16 training. We used a 16GB Tesla T4 GPU from the Google Cloud platform for training.[2] The fine-tuning process took 38 hours to complete.

3.1 Results

We define general MBart (out-of-domain data), and fine-tuned MBart (in-domain medical data). Table 1. shows the automatic metrics scores for both models. Fine-tuned MBart outperforms the general model on all the metrics. The BLEU score is statistically significant p=0.001 based on bootstrap resampling with 1,000 iterations.

	BLEU ↑	chrF ↑	COMET ↑
MBart	21.9	51.5	0.556
fine-tuned MBart	**25.8**	**54.9**	**0.663**

Table 1. Automatic metrics for MBart and fine-tuned MBart.

[1] https://dl.fbaipublicfiles.com/fairseq/models/mbart/mbart.cc25.ft.enro.tar.gz

[2] The scripts for our experiments are available at: https://github.com/mriosb08/medical-NMT-HAITrans

Furthermore, we performed an analysis of the COMET segment level scores. We use MT-Telescope (Rei et al., 2021) to compare both systems. Figure 1. shows the percentage of segments divided into four quality bins. Each bin is defined by a default threshold from the COMET scores, from green (residual errors) to red (critical errors). The red bin has translations lower than 0.10 score, the yellow bin has translations between 0.10 and 0.30 score, the light green has translations between 0.30 and 0.70 score, and the dark green has translations greater than 0.7 score. MBart is *System X* and fine-tuned MBart is *System Y*. The fine-tuned MBart has the highest number of high scores compared to the original general model.

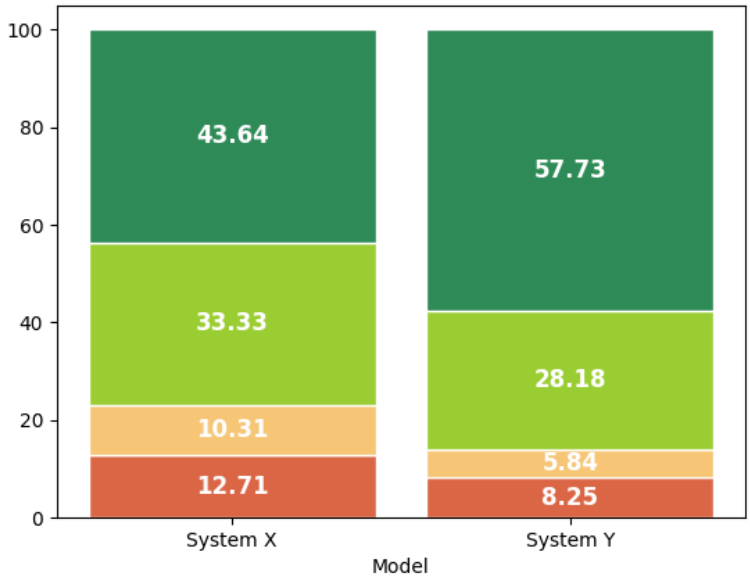


Figure 1. COMET scores for segments divided into quality bins for MBart (System X) and fine-tuned MBart (System Y).

Figure 2. shows visually the difference of COMET scores between the two systems for each segment. The size and colour of a bubble is the difference in the COMET score between systems for the same segment. Moreover, systems are different when the bubbles are far from each other along the axis (*x_score* MBart, and *y_score* fine-tuned MBart), and from the centre of the plot. Both systems MBart and fine-tuned MBart are different in terms of COMET scores, and fine-tuned MBart has a higher COMET score. If both models produce different translations, in this case, it means that the fine-tuned model is learning to generate MT outputs close to the medical domain.

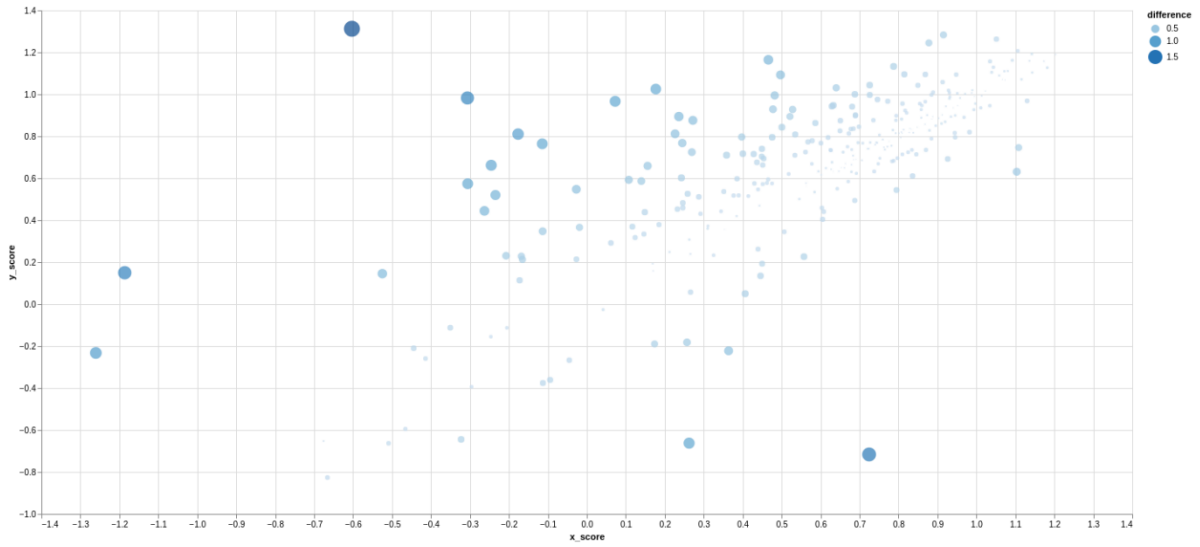


Figure 2. Bubble plot of segment level COMET scores for MBart (x_score) and fine-tuned MBart (y_score).

3.2 Error Analysis

To gain insights into the specific terminology errors produced by the two models, we show a sample of 12 abstracts with a total of 75 segments to three annotators. The annotators are native Romanian speakers with in-house and freelance translation experience; moreover, one of the annotators also has in-house and freelance medical translation experience. The annotators had access to the source, the reference, and the output of the two MT systems to annotate each MT segment with error categories (Klubička et al., 2017) using (Haque et al., 2019). The annotators annotated the abstracts collaboratively (Esperança-Rodier et al., 2019) – the motivation for the joint in-person annotation is to increase agreement for identifying possible terms and errors.

To perform the annotation, we set up a translation project in Trados Studio 2021[3] and import the source, reference and MT output files as bilingual .xlsx files. We install the freely-available Qualitivity[4] plugin integrated into Studio using an API key; this serves as the environment in which the annotators record any identified errors, their severity level and proposed corrections, along with explanatory comments. At the end of the annotation process, we export a report from Qualitivity containing the full annotation data for the reference texts, and MBart and fine-tuned MBart outputs.

The total number of terminology-related errors for general-model MBart and fine-tuned MBart are 98 and 64 respectively, demonstrating the improvement brought about by the fine-

[3] https://www.trados.com/products/trados-studio/
[4] https://community.rws.com/product-groups/trados-portfolio/rws-appstore/w/wiki/2251/qualitivity

tuning process with in-domain data. Table 2. shows the number of errors for each category present in the abstracts.

Error Type	MBart ↓	fine-tuned MBart ↓
Partial error	41	**23**
Source term copied	22	**9**
Inflectional error	**2**	4
Reorder error	**1**	3
Disambiguation issue in target	14	**6**
Incorrect lexical selection	9	**6**
Term drop	0	0
Other error	**9**	13

Table 2. Total errors for each terminology-related category.

The fine-tuned MBart model produces fewer errors than the general model on most of the categories. However, the fine-tuned model fails in the following categories: *Inflectional error*, *Reorder error*, and *Other*. Moreover, we show annotated examples of random segments for each error category to further understand the cause of the errors. In Table 3. we show a random selection of source and MT output for each error category, except *Other*, and highlight the annotated errors for each category for fine-tuned MBart.

Category	Source	Target (fine-tuned MBart)
Partial error	The DX-OSA score may be useful for identifying obese patients with significant OSA who require CPAP (continuous positive airway pressure) treatment, and CPAP could be commenced without the need for polysomnography, therefore, without delaying surgery.	Scorul DX-OSA poate fi util pentru identificarea pacienţilor obezi cu OSA semnificativă care necesită tratament cu CPAP (**tensiune arterială continuă pozitivă** [instead of **presiune pozitivă continuă în căile aeriene**]), iar CPAP poate fi început fără a fi necesară polisomnografie, prin urmare, fără a întârzia intervenţia chirurgicală.

Source term copied	The objectives of this study were to reveal possible relations between antioxidant therapy and a number of serum biochemical variables (ALT, AST, APPT, LDH, urea, leukocytes, platelets), the length of mechanical ventilation, the time spent in the ICU, and the mortality rate in major trauma patients.	Obiectivul acestui studiu a fost să evidențieze posibilele relații dintre tratamentul cu antioxidanti și o serie de variabile biochimice serice (ALT, AST, **APPT** [instead of **APTT**], LDH, uree, leucocite, trombocite), durata ventilației mecanice, timpul petrecut în ICU și rata mortalității la pacienții cu traumatisme majore.
Inflectional error	Two of these drugs, duloxetine and venlafaxine, are used also in chronic pain management.	Două dintre aceste medicamente, **duloxetină** şi **venlafaxină** [instead of **duloxetina** și **venlafaxina**], sunt utilizate şi în tratamentul durerii cronice.
Reorder error	Although not statistically significant, MODS and ARDS incidences were higher in the DCO shock group: MODS (41.7% versus 22.6% and 20%; p = 0.08/0.17), ARDS (29.2% versus 17% and 20%; p = 0.22/0.53).	Deşi nu au fost semnificative statistic, incidenţele MODS şi ARDS au fost mai mari în **grupul cu şoc DCO** [instead of **grupul DCO cu şoc**]: MODS (41,7% faţă de 22,6% și 20%; p = 0,08/0,07), ARDS (29,2% faţă de 17% și 20%; p = 0,22/0,53).
Disambiguation issue in target	The drug's efficacy results from its modulating effect on the descending inhibitory pain pathways and the inhibition of the nociceptive input.	Eficacitatea medicamentului rezultă din efectul său de modulare asupra **căilor de durere inhibatoare descendente** [instead of **căilor descendente inhibitorii ale durerii**] și inhibarea contribuției nociceptive.
Incorrect lexical selection	These results correlate with a higher trauma score in these patients, more serious lesions requiring several damage control procedures.	Aceste rezultate sunt corelate cu un **scor traumatic** [instead of **gravitatea traumatismelor**] mai mare la acești pacienți, leziunile mai grave necesitând mai multe proceduri de control al leziunilor.

Table 3. Fine-tuned MBart annotated examples for each error category (except *Other error*).

Table 4. shows all the examples for the *Other error* category for the fine-tuned MBart. As the fine-tuned model underperformed in terms of *Other* errors - 13 to 9 -, we investigate this further and list all the annotated errors within the *Other* category in Table 4. We identify two phenomena regarding the treatment of English borrowings and acronyms, and evidence of hallucination. The first phenomenon observed is that source terms are translated, even where a borrowing from English would be the correct translation strategy. For example, *Early Total Care* and *Damage Control Orthopaedics* lead to translations based on erroneous lexical selection: *metode de control al daunelor*, and *principii de îngrijire în primii ani de viaţă*, respectively, instead of retaining the original source terms in English. Moreover, for *burst* and

burst (suppression), the fine-tuned model produces the translations *arsură* and (*supresie) pulmonară* belonging to the lexical fields of *burn* and *bust*, pointing to challenges with the setup of the Byte pair encoding (BPE) vocabulary in NMT (Araabi et al., 2022; Lignos et al., 2019). Secondly, when acronyms should have been maintained as per the EN source, for instance *MODS, DCO, ARDS,* and *OS*, they are instead randomly recomposed as *SMO, COD, SRA,* and *SSO*. Acronyms corresponding to terms with a translation into Romanian are also randomly recomposed, for example *FR* is translated as *RF* rather than *RL*. Finally, there is also an example of a hallucination, the English *intramedullary (nailing)* is erroneously translated by adding a Romanian inflection at the end: *(nailing) intramedullar*, instead of *tijă centromedulară*.

Source	Target (fine-tuned MBart)
The aim of this study was to evaluate the frontal intracortical connectivity during deep anaesthesia (burst-suppression).	Scopul acestui studiu a fost să evalueze conectivitatea intracorticală frontală în timpul anesteziei profunde (**supresie pulmonară**).
Rats were maintained in deep level anaesthesia (burst-suppression).	Ratii s-au menţinut în anestezie profundă (**supresie pulmonară**).
The global cortical connectivity (0.5-100 Hz) was 0.61 ± 0.078 during the burst periods compared to 0.55 ± 0.032.	Conectivitatea corticală globală (0,5-100 Hz) a fost de 0,61 ± 0,078 în timpul perioadelor de **arsură** comparativ cu 0,55 ± 0,032.
The global cortical connectivity increased during the burst periods.	Conectivitatea corticală globală a crescut în timpul perioadelor de **arsură**.
Once the "two event model" was accepted, it became clear that patients although initially resuscitated, but in a vulnerable condition, have a high risk that a secondary aggression (for example, surgical interventions) would precipitate a state of hyperinflammation and early multiple organ dysfunction syndrome (MODS).	Odată ce „modelul celor două evenimente” a fost acceptat, a devenit clar că pacienţii, deşi iniţial resuscitaţi, dar aflaţi într-o stare vulnerabilă, prezintă un risc crescut ca o agresivitate secundară (de exemplu intervenţii chirurgicale) să precipite o stare de hiper inflamaţie şi sindrom de disfuncţie multiplă precoce (**SMO**).
This is a retrospective study of severe polytrauma patients with femoral shaft fractures admitted to the intensive care unit of the Emergency clinical Hospital of Bucharest and treated from an orthopaedic point of view by either Damage Control Orthopaedics (DCO) or Early Total Care (ETC) principles.	Acesta este un studiu retrospectiv la pacienţi cu politraum sever, cu fracturi ale căilor femurale, internaţi în unitatea de terapie intensivă a Spitalului clinic de urgenţă din Bucureşti şi trataţi din punct de vedere ortopedic, fie conform principiilor de **control al deteriorării (DCO)**, fie conform principiilor de îngrijire totală precoce (ETC).

This is a retrospective study of severe polytrauma patients with femoral shaft fractures admitted to the intensive care unit of the Emergency clinical Hospital of Bucharest and treated from an orthopaedic point of view by either Damage Control Orthopaedics (DCO) or Early Total Care (ETC) principles.	Acesta este un studiu retrospectiv la pacienţi cu politraum sever, cu fracturi ale căilor femurale, internaţi în unitatea de terapie intensivă a Spitalului clinic de urgenţă din Bucureşti şi trataţi din punct de vedere ortopedic, fie conform principiilor de control al deteriorării (DCO), fie conform principiilor de **îngrijire totală precoce (ETC)**.
Using patients files we recorded the following data: 30 day mortality, development of acute respiratory distress syndrome (ARDS) and MODS, local infectious complications (LIC), intensive care unit length of stay (ICU LOS), days of mechanical ventilation (MV), units of red blood cells units/48 h (RBC).	Utilizând dosarele pacienţilor, am înregistrat următoarele date: mortalitate cu durata de 30 zile, apariţia sindromului de detresă respiratorie acută (**SRA**) şi MODS, complicaţii infecţioase locale (LIC), durata de şedere la unitatea de terapie intensivă (ICU LOS), zile de ventilaţie mecanică (MV), unităţi de celule roşii în sânge/48 ore (RBC).
We decided to analyze results in three groups - DCO group with shock on admission, DCO group without shock and ETC group.	Am hotărât să analizăm rezultatele în trei grupuri - grupul cu **COD** cu șoc la admitere, grupul cu COD fără șoc și grupul cu ETC.
In the other two groups (DCO without shock and ETC) all outcomes were similar.	În celelalte două grupuri (**COD** fără şoc şi ETC), toate rezultatele au fost similare.
In patients who are not in a very severe condition (shock), the choice for femoral shaft stabilization by intramedullary nailing represents a safe option.	La pacienţii care nu sunt într-o afecţiune foarte severă (şoc), opţiunea stabilizării căilor femurale prin **nailing intramedullar** reprezintă o opţiune sigură.
The biochemical processes of bioproduction of free radicals (FR) are significantly increasing in polytrauma patients.	Procesele biochimice de bioproducţie a radicalilor liberi (**RF**) cresc semnificativ la pacienţii cu politrauma.
Decreased plasma concentrations of antioxidants, correlated with a disturbance of the redox balance are responsible for the installation of the phenomenon called oxidative stress (OS).	Scăderea concentraţiilor plasmatice de antioxidanti, corelată cu o tulburare a echilibrului redox, este responsabilă de instalarea fenomenului numit stres oxidativ (**SSO**).

Table 4. Fine-tuned MBart annotated examples for the *Other error* category. The additional errors present in these examples have not been highlighted in this table.

4 Conclusions and Future Work

We quantified the impact of domain adaptation on MBart in the medical domain for English-Romanian. The fine-tuned MBart outperforms the general model with automatic metrics and produces fewer errors (↓ 34.7%) related to terminology in the relatively small sample (75 segments belonging to 12 medical article abstracts) annotated by our annotators. While lower

numbers of errors were recorded in the *Partial error*, *Source term copied*, *Disambiguation issue in target*, *Incorrect lexical selection*, and *Term drop*, in the three remaining categories the fine-tuned MBart engine actually contained more errors than general MBart: *Inflectional error*, *Reorder error*, and *Other error*.

Of these three categories, the *Inflectional error*, and *Other error* items present in the fine-tuned MBart output we evaluated are related to the Byte pair encoding (BPE) vocabulary. In future work, we plan to extend the BPE vocabulary in MBart (Berard, 2021) to cope with in-domain terminology, and to quantify the impact of the fine-tuning on other error types present in the MQM Core. Moreover, we noticed further examples of hallucinations, but they were not within the area of terminology translation, and we will leave them as future work, alongside the additional types of errors noticed in the general MBart and fine-tuned MBart outputs, but also in the reference translations, which were by no means error-free.

More generally, it is essential to raise the awareness of machine translation post-editors, as well as clients, regarding how these error categories are still manifested in MT output even after fine-tuning. NMT output errors remain difficult to identify due to the apparent fluency of the output, and even subject-matter experts can miss some of them. The alert *revision* and *correction* of MT output (which has been misleadingly called "postediting" for the past 60 years (Pierce & Carroll, 1966) as if it were a monolingual task, not a bilingual one) carries important risks in some settings if assigned to only one person working under high time pressure and using the same text-based revision environments created in the 1990s to accommodate translation memories.

Acknowledgments

The GPU used for this research was sponsored by the Google Cloud Research Credits Program.

References

Araabi, A., & Monz, C. (2020). *Optimizing Transformer for Low-Resource Neural Machine Translation* (arXiv:2011.02266). arXiv. http://arxiv.org/abs/2011.02266

Araabi, A., Monz, C., & Niculae, V. (2022). *How Effective is Byte Pair Encoding for Out-Of-Vocabulary Words in Neural Machine Translation?* (arXiv:2208.05225). arXiv. https://doi.org/10.48550/arXiv.2208.05225

Berard, A. (2021). *Continual Learning in Multilingual NMT via Language-Specific Embeddings* (arXiv:2110.10478). arXiv. https://doi.org/10.48550/arXiv.2110.10478

Bérard, A., Kim, Z. M., Nikoulina, V., Park, E. L., & Gallé, M. (2020). *A Multilingual Neural Machine Translation Model for Biomedical Data* (arXiv:2008.02878). arXiv. http://arxiv.org/abs/2008.02878

Chu, C., & Wang, R. (2018). A Survey of Domain Adaptation for Neural Machine Translation. *Proceedings of the 27th International Conference on Computational Linguistics*, 1304–1319. https://aclanthology.org/C18-1111

CLARIN:EL. (2015). *EMEA Corpus*. CLARIN:EL. http://hdl.handle.net/11500/ATHENA-0000-0000-25DB-0

Esperança-Rodier, E., Brunet-Manquat, F., & Eady, S. (2019, November). ACCOLÉ: A Collaborative Platform of Error Annotation for Aligned Corpora. *Translating and the Computer 41*. https://hal.archives-ouvertes.fr/hal-02363208

Haque, R., Hasanuzzaman, M., & Way, A. (2019). Investigating Terminology Translation in Statistical and Neural Machine Translation: A Case Study on English-to-Hindi and Hindi-to-English. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 437–446. https://doi.org/10.26615/978-954-452-056-4_052

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., & Dean, J. (2017). Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, *5*, 339–351. https://doi.org/10.1162/tacl_a_00065

Klubička, F., Toral, A., & Sánchez-Cartagena, V. M. (2017). Fine-grained human evaluation of neural versus phrase-based machine translation. *The Prague Bulletin of Mathematical Linguistics*, *108*(1), 121–132. https://doi.org/10.1515/pralin-2017-0014

Lignos, C., Cohen, D., Lien, Y.-C., Mehta, P., Croft, W. B., & Miller, S. (2019). The Challenges of Optimizing Machine Translation for Low Resource Cross-Language Information Retrieval. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3497–3502. https://doi.org/10.18653/v1/D19-1353

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, *8*, 726–742. https://doi.org/10.1162/tacl_a_00343

Lommel, A. R., Burchardt, A., & Uszkoreit, H. (2013). *Multidimensional quality metrics: A flexible system for assessing translation quality*. 7.

Neves, M., Jimeno Yepes, A., Névéol, A., Grozea, C., Siu, A., Kittner, M., & Verspoor, K. (2018). Findings of the WMT 2018 Biomedical Translation Shared Task: Evaluation on Medline test sets. *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 324–339. https://doi.org/10.18653/v1/W18-6403

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., & Auli, M. (2019). *fairseq: A Fast, Extensible Toolkit for Sequence Modeling* (arXiv:1904.01038). arXiv. http://arxiv.org/abs/1904.01038

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. https://doi.org/10.3115/1073083.1073135

Pierce, J. R., & Carroll, J. B. (1966). *Language and Machines: Computers in Translation and Linguistics*. National Academy of Sciences/National Research Council.

Popović, M. (2015). chrF: Character n-gram F-score for automatic MT evaluation. *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 392–395. https://doi.org/10.18653/v1/W15-3049

Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. *Proceedings of the Third Conference on Machine Translation: Research Papers*, 186–191. https://doi.org/10.18653/v1/W18-6319

Rei, R., Farinha, A. C., Stewart, C., Coheur, L., & Lavie, A. (2021). MT-Telescope: An interactive platform for contrastive evaluation of MT systems. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, 73–80. https://doi.org/10.18653/v1/2021.acl-demo.9

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A Neural Framework for MT Evaluation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702. https://doi.org/10.18653/v1/2020.emnlp-main.213

Skianis, K., Briand, Y., & Desgrippes, F. (2020). Evaluation of Machine Translation Methods applied to Medical Terminologies. *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, 59–69. https://doi.org/10.18653/v1/2020.louhi-1.7

Verma, N., Murray, K., & Duh, K. (2022). Strategies for Adapting Multilingual Pre-training for Domain-Specific Machine Translation. *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, 31–44. https://aclanthology.org/2022.amta-research.3

A Study Towards a Standardized Typology of Machine Translation Post-Editing Guidelines: A Suggested Template for Language Professionals

Viveta Gene
PhD Candidate, Ionian University (Corfu)
f20gken@ionio.gr

Lucía Guerrero
Machine Translation Manager
lucguerrero@expediagroup.com

Abstract

It is commonly agreed that Machine Translation Post-Editing (MTPE) guidelines are key for a successful outcome in MTPE tasks. In spite of this fact, the current lack of an international standard or recommendations on how to draft such guidelines and what type of information they should include risks creating confusion amongst post-editors. This paper aims to use as a basis the findings of the authors' previous workshop on the knowhow of drafting MTPE guidelines (Gene & Guerrero, 2021) and reports on a qualitative study that investigates how professional users and writers of MTPE guidelines expect them to be drafted in terms of content, length, and format. The findings of the research lead to the creation of an example template of MTPE guidelines and a questionnaire for guidance on when to write them, having the practical perspective of covering the gap of a common MTPE guidelines template for the Language Service Providers and clients.

Introduction

MTPE is the focus of Language Service Providers (LSPs) in recent years with great interest of research as a result of productivity gains (Guerberof, 2009; Federico et al., 2012; WEB, a). However, it has been noted that there are no widely accepted general or standard MTPE guidelines (DePalma, 2013; Hu. & Cadwell, 2016; TAUS, 2016). More specifically, the existing guidelines entail the difficulty of interpretation by the user, primarily due to users' competency gaps, but also due to the wording of the guidelines (Flanagan & Christensen, 2014). MTPE is defined as 'the task of editing, modifying and/or correcting pre-translated text that has been processed by an MT system from a source language into a target language' (Allen 2003, 297). De Almeida and O'Brien (2010) argue that there is currently a lack of MTPE training and of clear and consistent guidelines for professional post-editors.

With that in mind, the authors, members of the MTPE Training Special Interest Group steering committee, which is an initiative from the Globalization and Localization Association (GALA), in which stakeholders from academia, clients, LSPs and post-editors collaborate towards drafting a common protocol on MTPE training, conducted this study expanding their previous work at the AsLing TC43 Conference (Gene, Guerrero 2021) in order to propose best practices in the field of designing and drafting MTPE guidelines.

The paper is structured as follows: Section 2 discusses previous research on MTPE guidelines and gaps. Section 3 introduces the previous work of the authors for the AsLing TC43 workshop on the topic. Section 4 discusses the expansion of the previous work with the current study of the authors presented in this paper presenting its methodology, Section 5 presents the data analysis and findings, and Section 6 discusses the findings and presents our MTPE guidelines template for the use by LSPs. We draw conclusions and propose future research in Section 7.

The need for MTPE Guidelines Research

The Scope of MTPE Guidelines

According to academic research and the best practices described by several organisations (to be explored later), MTPE guidelines can be defined as a set of specific instructions that the requesters of a MTPE service, such as machine translation (MT) buyers, LSPs or researchers, prepare for post-editors so that they know exactly what is expected of them in terms of quality, tools, the areas to focus on and other important aspects that define how the MTPE task is to be carried out and determine the final output.

As MTPE workflows become increasingly common, there is a growing need for both organisation-specific and more general MTPE guidelines (Guerberof 2009, 2010). A set of such guidelines was drafted by TAUS (2010), with the intention of helping clients and LSPs to set clear expectations and instruct post-editors. These guidelines include general recommendations for decreasing the amount of MTPE needed, as well as basic guidelines for carrying out MTPE to two defined quality levels (Koponen, 2016).

Based on this, we examine to what extent MTPE guidelines are deemed relevant in any MTPE workflow. According to Díaz and Rico (2012), the specification of MTPE guidelines is founded on a decision-making process regarding the text quality acceptance, which is determined by client expectations, the turnaround time or document life cycle, among others. The importance of drafting MTPE guidelines is also highlighted by Hu and Cadwell (2016) as this is the vehicle for clients and LSPs to set clear expectations and reduce the effort for post-editors.

Overall, we can conclude that the two main objectives of MTPE guidelines are to transform the customer's expectations into clear specifications and save time and effort for the post-editors.

In this paper we will consider MTPE guidelines as part of an MTPE assignment which can happen in two different scenarios: a) Post-editing in an isolated environment involving an interface with source/MTPE output without any interaction with a computer-assisted translation (CAT) tool — a research environment, for example, and b) Post-editing in a multimodal interface involving CAT tools, meaning that the MTPE guidelines are part of an assignment containing instructions for the use of translation memories, glossaries, style guides, client instructions, delivery instructions and more.

The Need for MTPE Guidelines

Although MTPE guidelines have been drafted (e.g., ISO/DIS 18587, 2016; TAUS, 2010), their practical implementation is not necessarily clear to post-editors (Koponen, Salmi, 2017). The fact that post-editors may have differing interpretations of the guidelines (Flanagan & Christensen, 2014, pp. 264–265) reveals the need for research in the field of the "how-to" when drafting MTPE guidelines from the perspective of content, length, and format. To support the idea that MTPE guidelines should be part of any MTPE assignment, the findings during our collaboration work in GALA will be presented in the Section 3.1.

Clear guidance on how to perform the MTPE task, indeed any translation task, is a very basic need especially where there are assumptions and grey areas. To support this argument, in ISO 18587, the standard for MTPE, it is stated that full MTPE describes the "process of post-editing to obtain a product comparable to a product obtained by human translation". The word 'comparable' sets the question of: a) how comparable is defined and b) in what terms this is comparable. Style is also one of the most controversial aspects in MTPE specifications, and often the main subject of arbitration, asking for guidelines making clear the definition between a preferential/stylistic and non-preferential/non-stylistic change.

Another point making the drafting of MTPE guidelines a critical need is the differentiation of quality levels in MTPE with the main quality levels being "light" and "full". The topic was examined in the light of Nunziatini and Marg (2020), who propose a better definition of a MTPE level which is between full and light ('medium post-editing'). They present comparison tables showing that standards and industry practices generally agree on the existence of two MTPE levels — light and full — but they have different views on what these levels entail. They suggest the idea that there is room to offer flexible MTPE services between those two, depending on the aspects to be focused on (e.g., technical terms, brands/product names, etc.).

Based on the work of Nunziatini and Marg (2020), asking for a quality level of MTPE without giving

instructions to post-editors can lead to the following implications: i) Misunderstanding scope. There is a widespread assumption that light MTPE is a level which should be applied when the quality of the MT engine is so good that only a few tweaks in the target text are needed. This conception is wrong, as the distinction between full and light MTPE is not based on the amount of editing effort, but on the purpose of the translation and the quality requirements. As defined in ISO 18587, "[light post-editing is] normally used when the final text is not intended for publication and is mainly needed for information gisting, i.e., for rendering the main idea or point of the text. In this level of post-editing, the output shall be comprehensible and accurate but need not be stylistically adequate"; ii) Misunderstanding about the quantity of corrections. Post-editors, in particular very experienced translators who usually deal with high quality requirements and have less experience with MTPE, tend to engage in full MTPE especially when light MTPE instructions are not clear. iii) Misunderstanding about the stylistic corrections. This is due to the lack of agreement between existing standards and guidelines (Nunziatini and Marg, 2020). According to the TAUS guidelines, the style "may not be as good as that achieved by a native-speaker human translator". Sharon O'Brien (2010) recommends that stylistic and textuality problems are ignored while ISO 18587 recommends that client's stylistic guidelines are followed, and highlights that the style should be appropriate for the text type. The ability to focus on the correction of specific error categories only while not fixing the rest, even if they are detected, certainly requires clear instructions just as much as training.

Finally, it makes sense to conclude that insofar as there are many different parameters that define a translation assignment, the same should apply to MTPE. Different quality levels of MTPE are acceptable, which leads us again to the need to develop clear guidelines covering the specific aspects to focus on at each of these levels.

The Gaps in MTPE Guidelines

As already pointed out, individuals who are in charge of preparing MTPE assignments and/or would like to write instructions face several challenges: i) The lack of internationally agreed standards on how to draft MTPE guidelines, or the inconsistency of existing guidelines; ii) The restricted availability of existing MTPE guidelines, limited exclusively to the organisations in charge of writing these guidelines, and thus the significant variations among them; iii) The absence of a clear distinction between MTPE guidelines, which are supposed to be very specific to a MTPE task, and general assignment instructions, resulting in an overlap of instructions.

More specifically, existing academic research on how to draft MTPE guidelines is very scarce, and most of the papers can be classified under any of the following categories:[1] i) language-dependent; ii) domain-based; iii) outdated (only apply to rule-based or statistical MT); iv) comparative studies; v) exclusively focused on the two levels of MTPE (light and full).

This means that based on the current research of the existing studies, for a potential writer of MTPE guidelines, the know-how is limited as this does not cover all scenarios. All these shortcomings and challenges mentioned in Section 2.2, if not addressed properly, can result in a negative perception of the guidelines by the post-editors, who may consider them too elaborate, too dense, too long, complicated, repetitive, or inconsistent.

Considering the existing gaps, the different and sometimes contradictory assumptions that exist regarding MTPE levels, and the fact that each task has its specific requirements in terms of quality, it may be concluded that it is extremely important to summarize all these quality expectations in a document that the post-editors can understand, and which effectively can guide them towards meeting the quality expectations, while at the same time saving time and effort, which is the goal of this study.

The Prototype for the Current Study

The GALA MTPE Training Special Interest Group (SIG)

The MTPE Training GALA SIG is a collaborative initiative aiming to develop best practices in the training and preparation of professionals who handle the post-editing of machine translated content. In the context of this SIG, the authors

[1] See the References section for a list of related literature.

being the founders and moderators of this group, have examined in one of their meetings (October 2021) with the members who fall under the groups of Academia, Post-Editors, LSPs and Clients, the topic of drafting guidelines for setting a basic MTPE workflow.[2]

The authors asked the participants what the guidelines are, i.e., aspects that the requesters of the MTPE service should consider when preparing a MTPE assignment. Based on the ideas suggested, the voting process revealed as the three most voted aspects of MTPE guidelines: i) the expected effort and the type of corrections, ii) the client expectations for MTPE and cost savings, and the iii) Error typology list based on the MT system behaviour.

The AsLing TC43 Conference Workshop

The authors conducted a workshop at the AsLing TC43 conference on how to design MTPE guidelines proposing best practices (Gene & Guerrero, 2021). The objective of the AsLing TC43 conference workshop was twofold:

- Create awareness of the importance of sending accurate instructions to the post-editors so that they may carry out the MTPE task according to the customer's expectations, focusing exclusively on the aspects that are required.
- Collect feedback to create a flexible and granular template for writing effective MTPE guidelines in the context of NMT including a) content, b) length, and c) format.

The findings of this workshop revealed the need to produce meaningful MTPE guidelines and allowed the authors to draw some conclusions in the form of recommendations: a) define use cases for writing MTPE guidelines, b) support the understanding and clarity of the MTPE guidelines in the form of a supplementary session with the post-editors upon communication of such guidelines, and c) ensure the relevance and accuracy of their content. Based on the participants input the sections of the MTPE guidelines should include, but not be limited to i) subject area, ii) content type, iii) purpose of translation, iv) level of MTPE quality expected (light/medium/full MTPE), v) type of MT system, vi) examples of errors, and vii) tips on how to address errors.

However, due to the structure of the workshop and the time restrictions regarding its organization and duration, some of the questions that were raised and remained unanswered during the discussion due to lack of time were referring to: a) the aspects that writers take into consideration when deciding whether to write MTPE guidelines, b) the content that should appear in the guidelines, and c) additional sections that users and writers consider that they should be part of the MTPE guidelines.

Additionally, the number of participants representing users and writers of MTPE guidelines in this workshop was not representative to make the findings valid.

Based on these preliminary conclusions, the authors have agreed to continue the above-mentioned work and relaunched the research in the form of an online questionnaire targeting users and writers of MTPE guidelines.

Expansion of Work and Current Study of the Authors: Methodology

A questionnaire designed on Microsoft Forms was made available for about 3 weeks — from 22nd February until 10th March EOD CET, and promoted on social media (LinkedIn, Twitter), the GALA MTPE Training Special Interest Group, and via an emailing campaign targeted to external collaborators from companies where the authors worked at the time of writing this paper.

Following the consent statement, branching logic options split the respondents into three question paths depending on whether they were users or potential users of MTPE guidelines e.g., freelance or inhouse translators/post-editors, students, etc., writers or potential writers of MTPE guidelines e.g., translation/localization project managers, researchers, etc., or none of these two. The questions from each path were identical, or almost identical but from a different point of view. All the replies were downloaded in a spreadsheet and transformed into pivot tables and charts.

[2] https://www.gala-global.org/events/events-calendar/gala-connected-2021-bounce-forward-gala-mtpe-training-sig-joining-forces

Data Analysis and Findings

Profile of the participants

From a total of 229 respondents, 4 did not accept the consent form and 225 accepted and took the questionnaire.

The majority of participants were users rather than writers of MTPE guidelines (73.78% vs 10.22%), which goes in line with the most voted job profile: 74.32% of respondents were freelance translators or post-editors, compared to a 10.36% of LSP staff members, which, based on the answers, the authors assume accounts for the writers. The rest of the job roles are much less represented: academia staff: 5.41%, inhouse translator/post-editor in the public sector: 4.05%, inhouse translator/post-editor in LSP: 2.7%, public sector staff: 1.35%, private sector (other than LSP) staff: 0.9%, inhouse translators/post-editor in the private sector (other than LSPs): 1.8%.

Frequency

Both writers and users were asked to choose from a list of default answers about how often they create (writers) or receive (users) MTPE guidelines. The majority of both writers (39.13%) and users (23.49%) create or receive them, respectively, only in specific cases. In the next section we present more details about such cases.

As to the rest of the options, to our surprise, most writers (43.49%) chose positive answers ('always', 'very often', 'often') whereas most users (43.37%) chose negative answers ('never', 'rarely', 'very rarely'). Such figures lead us to the assumption that the questionnaire was responded only by writers who have experience in drafting MTPE guidelines or at least with machine translation post-editing workflows. Up to what extent the profiles of both writers and users can affect the writing and reception of the MTPE guidelines is, in fact, a topic which deserves future work on its own.

Motivations

In an open question targeted for the writers, the authors enquired them about the aspects considered when deciding whether to write or not MTPE guidelines. We are providing a summary of the reported motivations: i) skill level and experience of the post-editor in the subject area, ii) customer needs and requirements awareness (purpose, quality level, iii) content, iv) atypical errors, v) complexity (length of project and training level of the linguists), vi) time available for writing, vii) impact of a non-well performed MTPE, and viii) the existence of standard MTPE guidelines.

Difficulty/Usefulness

The following question was different for writers and users: writers were enquired whether they find writing MTPE guidelines a difficult task, and why, whereas users were asked whether they find them useful or not, and why. Both were presented a rating scale from 1 to 5 as shown in the figures below.

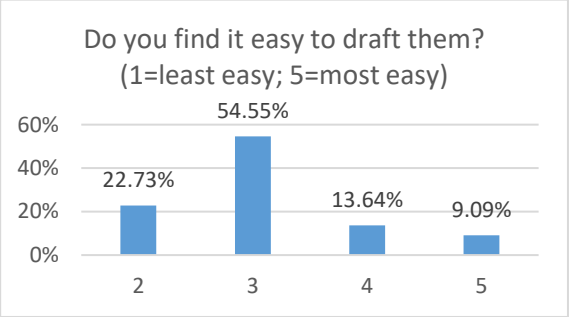


Figure 1: Question for the writers: 'Do you find it easy to draft them?'

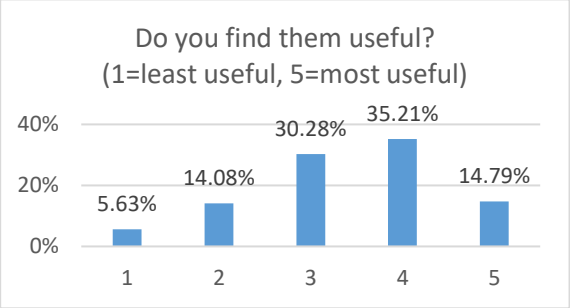


Figure 2: Question for the users: 'Do you find them useful?'

The fact that 50% of the users chose a positive score (4 or 5) as opposed to a 19.71% who chose a negative score (1 or 2) supports the notion that MTPE guidelines are perceived as a valuable material by most post-editors. The overload of information exceeding the compensation for the MTPE task itself was reported as the main reason why some users do not consider the MTPE guidelines useful. Other reasons were the lack of practical aspects or the contents being too superfluous or vague.

As to the writers, who also gave a majority of positive scores, the reasons highlighted by those

who consider that drafting guidelines is not easy were: i) worthless task due to users not reading the guidelines, ii) lack of information about the contents, and iii) task perceived as time-consuming. Finally, two main assumptions pointed out by writers regarding the reasons why users do not read the guidelines are the lack of observance and importance attached to the guidelines, and the lack of experience with the observance of guidelines. Exploring such assumptions could be a topic for future work.

Length

The figures below suggest that MTPE guidelines have an average length of 1-2 pages.

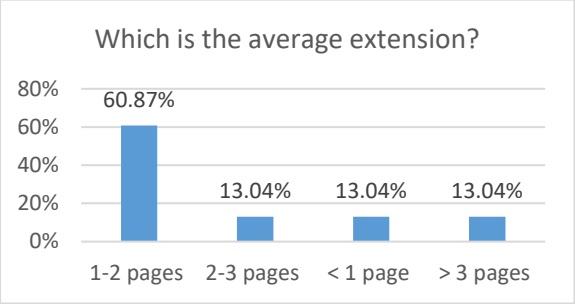


Figure 3: Question for the writers: 'Which is the average extension?'

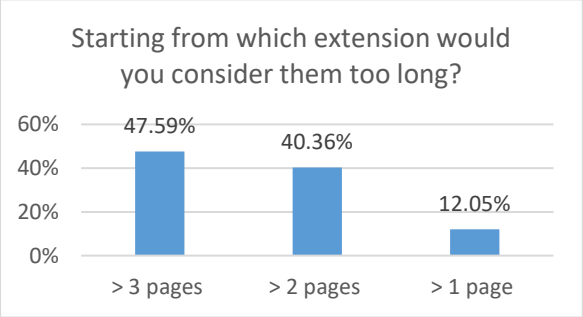


Figure 4: Question for the users: 'Starting from which extension would you consider them too long'?

Those users who considered that MTPE guidelines should not exceed one single page were required to indicate which are the very basic sections that should be included in them: i) target audience, ii) level of quality and/or MTPE, iii) glossaries, TM, style guides, register, iv) CAT and other tools to be used, and v) main errors, localization conventions (e.g. target language variant and date format).

Contributions

Due to the fact that the needs in the different language industry settings vary, MTPE guidelines will never be general or standard. (Hu & Cadwell, 2016). Also, as a type of translation instructions, it seems reasonable to think of the MTPE guidelines as a 'work-in-progress' document that can be updated with examples of MT errors once the assignment is completed for future reference, as described by Díaz and Rico (2012).

The questionnaire reveals that the majority of writers (60.78%) ask post-editors to share examples of MT errors or tips on how to detect or fix them, whereas most of the users (84.94%) said they are not asked to do so. This implies the possibility that most of the writers who participated in the questionnaire have an advanced knowledge of MTPE workflows, whereas the users might be receiving assignments from translation project managers with a variety of profiles, which is similar to what was observed in the question about frequency.

Format

According to both writers (92.45%) and users (87.35%), Word/PDF is the best format for MTPE guidelines for reasons of readability, editability, transferability, practicality, usability and clarity (being printable and searchable are especially mentioned by users).

It is worth mentioning that the kick-off meeting/call as a complement to the PDF was preferred by 8.7% of the writers due to the fact that it is i) well understood, ii) an open format for feedback and questions, and iii) a better environment for clarifications about budget and technical set up.

Finally, none of the writers voted for 'recorded video'. Likewise, it was the second-least voted by writers.

Contents

The next two questions, targeted to all participants, explore in more detail which sections are currently included in the MTPE guidelines and which should be part of them.

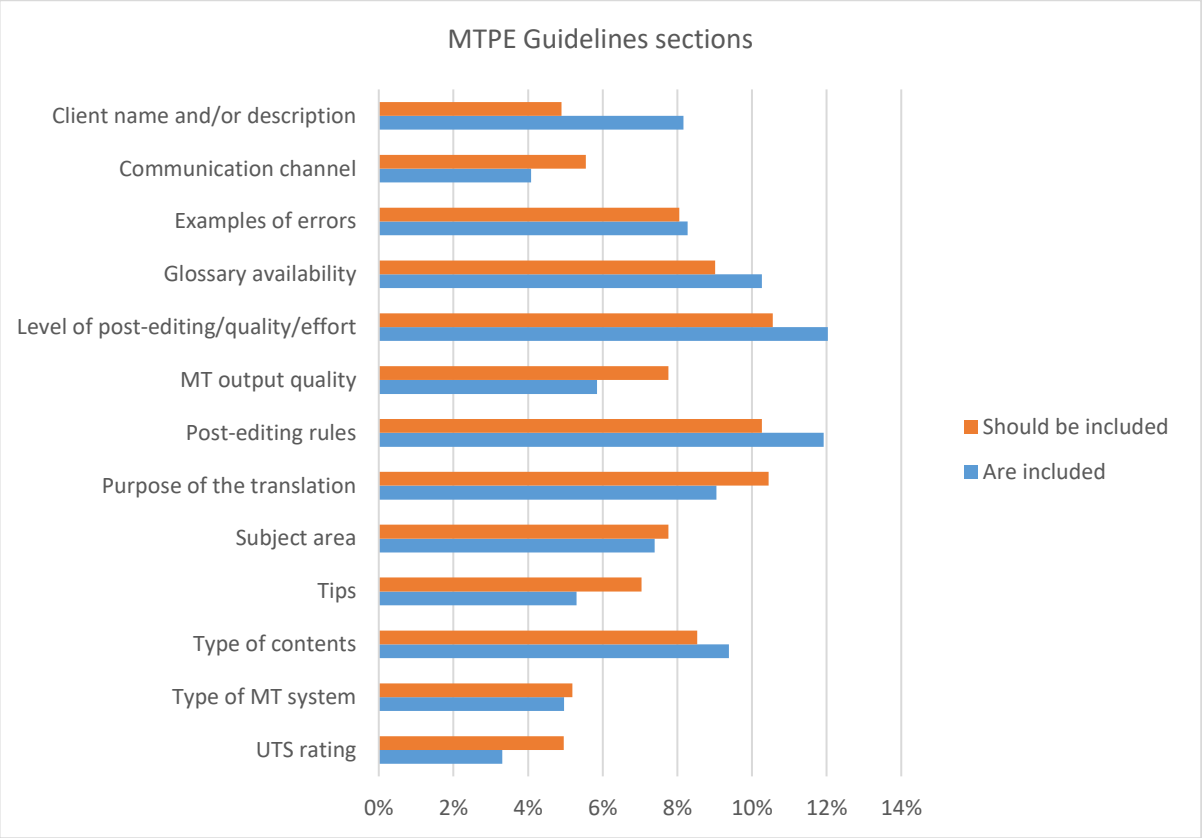


Figure 5: Questions for all participants: 'Which of these areas are covered in the guidelines?' and 'Which of these areas you think must be covered in the guidelines?'[3]

The answers to both questions show a similar distribution. If we look at which sections should be included, the results indicate a preference for those aspects that are specific to the MTPE project: level of MTPE, purpose of the translation, MTPE rules and glossary availability, suggesting that these are the sections to be included in all MTPE guidelines. These aspects are followed by those related to the nature of the source document: type of contents and subject area. The next most voted sections are aspects directly related to the MT engine: MT output quality and examples of errors.

Finally, the least voted sections are type of MT system, tips, details about the contents (communication channel and UTS rating) and client name and or description. From an industry-oriented perspective, the authors are inclined to think that the reasons for which those aspects are not considered essential could be the fact that nowadays the use of NMT is assumed. Likewise, the tips could be part of a basic training on MTPE, and the details about the contents and the client could be inferred from the source text.

Additionally to the default answers, participants contributed with the following sections: i) locale-specific aspects (e.g. product capitalization, measurement units, dates, numbers, cultural aspects affecting the MTPE task), ii) general feedback, iii) editing in a CAT tool, iv) stylistic aspects (e.g. active voice, present tense, third person), v) links to resources, vi) client specific requests/instructions, vii) identification of MT segments in the CAT tool/TMS, viii) links to additional resources, and ix) target audience.

Some of these suggested sections, which are not specific to MTPE, support the notion that MTPE guidelines are, in fact, a type of translation instructions expanded with aspects oriented towards the MTPE task. Combined with the rest of the findings, this research reveals that there is

[3] Communication channel means internal/external. UTS rating refers to utility (functionality of the translation), time (speed at which the MTPE output is to be handed), and sentiment (importance of impact on brand image), and it is usually scored as low, medium or high.

solid ground for developing a template to guide the writers towards drafting a document which is useful to post-editors.

A Suggested Template for MTPE Guidelines

Drafting one single instruction document covering all kinds of MTPE jobs and scenarios is neither impossible, nor it is recommendable. Nevertheless, the authors use the previous research of the EDI-TA project[4]combined with their findings from this study to propose a flexible template and a questionnaire in an attempt to cover the need to guide writers when drafting MTPE guidelines (Appendix A-I and II).

The aim of EDI-TA was to define a methodology for MTPE and included a strategy for inserting metadata into the bilingual files for later use, as well as a dynamic MTPE tool containing all the aspects which must be considered when writing MTPE guidelines and inserting the above-mentioned metadata. Even if it was based on a few language pairs (English-Spanish, Spanish-English, Spanish-Catalan, Spanish-Basque), and the rules and patterns that it describes are very specific to an MT technology that is hardly used nowadays (rule-based and statistic MT) the authors found in EDI-TA a source of inspiration for designing the questionnaire and the MTPE guidelines template. The basic sections suggested by the EDI-TA project were PE Project Information, Communication Channel, UTS Rating, Content Profile, MTPE Rules and Example Patterns.

The proposed template transforms the sections of the EDI-TA project (Fig. 7) using the findings of this study and giving an industrial and practical perspective as depicted in Appendix A. The template is complemented with a list of questions targeted for potential writers of MTPE guidelines, to assist them in their decision-making process. These questions are based on the list of aspects highlighted by the writers who participated in the questionnaire.

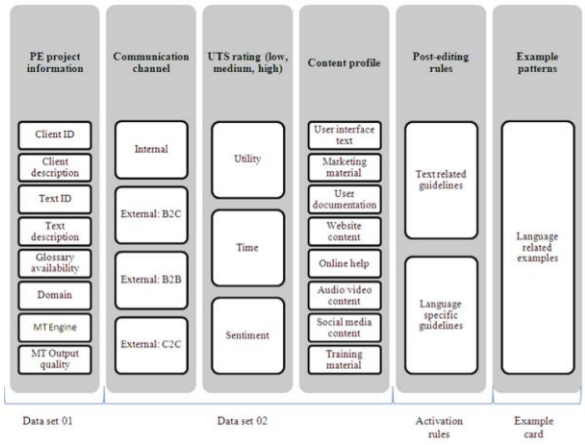


Figure 6: Elements to be considered when designing MTPE guidelines (EDI-TA project)

Conclusions and Future Work

Based on the findings of this research, we can conclude that there is agreement as to using MTPE guidelines only in specific cases based on several factors, mainly the post-editors' skills and experience, as well as the complexity or specificity of the job and the time available. The most recommended format is Word or PDF, though a kick-off meeting was pointed out as a complementary format to address questions. The ideal length is between 1 and 2 pages, which should contain at least instructions about the level of MTPE, the MTPE rules to follow, and information about the purpose of the translation and glossary availability. The template and the questionnaire presented on Annex A to guide potential writers of MTPE guidelines are based on these conclusions.

For future work, the authors suggest exploring the following topics: i) the method and frequency of contribution of the post-editors to the MTPE guidelines with examples and/or tips, , ii) the implementation or not of such contributions to the MTPE guidelines by the writers, iii) how the profile of the (potential) writers and users of MTPE guidelines (experience with MTPE workflows, attitude towards MT) can influence the drafting of such guidelines, in the case of the writers, and its reception, in the case of the users, and iv) the preconceived ideas behind writers' assumptions regarding the reasons why users do not read the guidelines (as mentioned, lack of observance and importance attached to the

[4] EDI-TA was a R&D project funded by the European Commission as part of the MultilingualWeb-LT (Language Technology on the Web) group and conducted by Linguaserve and Universidad Europea de Madrid in 2012.

guidelines, and the lack of experience with the observance of guidelines) and up to what extent they correlate with the real reasons.

References

Allen, J. (2003). Post-editing. Computers and Translations: A Translator's Guide. 35, 297-317.

Díaz, P., Rico, C. (2012). D4.1.4-Annex I: EDI-TA: Post-editing methodology for machine translation. Accessed October 2021. Available at www.w3.org/International/multilingualweb/lt/wiki/images/1/1f/D4.1.4.Annex_I_EDI-TA_Methology.pdf

DePalma, D. (2013). Post-editing in practice, available at http://www.tcworld.info/e-magazine/translation-and-localization/article/post-editing-inpractice/

Flanagan, M., Christensen, T.P. (2014). Testing post-editing guidelines: how translation trainees interpret them and how to tailor them for translator training purposes. The Interpreter and Translator Trainer, vol. 8, no. 2, pp. 257-275.

Guerberof, A. (2009). Productivity and quality in MT post-editing, In: Proceedings of MT Summit XII-Workshop: Beyond Translation Memories: New Tools for Translators MT, AMTA 2009 (26-30 Aug. 2009, Ottawa, Canada).

Hu, K., Cadwell, P. (2016). A Comparative Study of Post-editing Guidelines. Baltic J. Modern Computing, Vol. 4 (2016), No. 2, 346-353.

International Organization for Standardization. 2017. Translation services — Post-editing of machine translation output — Requirements (ISO Standard No. 18587).

ISO, ISO 17100:2015: Translation services – Requirements for translation services, available at http://www.iso.org/iso/catalogue_detail.htm?csnumber=59149

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. 131-148.

Koponen, M., Salmi, L. (2017). Post-editing quality: Analysing the correctness and necessity of post-editor corrections. Linguistica Antverpiensia.16. 137-148.

Nunziatini, M., Marg, L. (2020). Machine Translation Post-Editing Levels: Breaking Away from the Tradition and Delivering a Tailored Service. Accessed October 2021. Available at https://aclanthology.org/2020.eamt-1.33.pdf

O'Brien, S. (2010). Introduction to Post-Editing: Who, What, How and Where to Next? Accessed January 2020. Available at https://aclanthology.org/2010.amta-tutorials.1.pdf

Rico-Pérez, C., Ariano-Gahn, M. (2014). Defining language dependent post-editing rules: the case of the language pair English-Spanish. In O'Brien, S., Balling, L.M., Carl, M., Simard, M., Specia, L., (eds.), Post-editing of machine translation: processes and applications. Newcastle, pp. 299-322.

TAUS. (2010). MT Post-editing Guidelines. Accessed October 2021. Available at https://www.taus.net/academy/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines

TAUS. (2016). TAUS Post-Editing Guidelines, available at https://www.taus.net/think-tank/articles/postedit-articles/taus-post-editing-guidelines

WEB (a). Post-Edited Machine Translation. http://iconictranslation.com/solutions/post_edited_mt/

WEB (b). Post Editing Guidelines for GALE Machine Translation Evaluation. http://projects.ldc.upenn.edu/gale/Translation/Editors/GALEpostedit_guidelines-3.0.2.pdf

WEB (c). Post Editing Guidelines for BOLT Machine Translation Evaluation. http://www.nist.gov/itl/iad/mig/upload/BOLT_P3_PostEditingGuidelinesV1_3_3.pdf

Appendix A

I. MTPE guidelines template

MTPE project information	
Level of MTPE	*E.g.: light/full/medium*
Purpose of the translation	*E.g.: for publishing/understanding*
Target audience*	
MTPE rules	*E.g.: types of errors to be/not to be corrected*
Glossary	
MT output quality	*E.g.: score/expected words per hour*
Type of MT system*	*E.g.: generic/domain-based, RBMT/SMT/NMT*
Locale specific aspects*	*E.g. product capitalization, measurement units, dates, numbers, cultural aspects affecting the MTPE task*
Stylistic aspects*	*E.g.: active voice, present tense, third person*
CAT tool considerations*	*E.g.: pretranslation with TM/MT threshold*
Links to resources*	
Content profile	
Content type	*E.g.: user's guide, UI*
Subject area	*E.g.: legal, IT, pharma*
Client name*	
Communication channel*	*E.g.: internal/external*
UTS rating*	*E.g.: low/medium/high, see Footnote 3*
Examples of errors of the raw MT	**Tips to fix them***

* Optional

II. Questionnaire for potential writers of MTPE guidelines

1. **Related to time and resources**:

- Do we already have MTPE guidelines? Consider simply adapting them
- Do we have the time to write them?

2. **Related to the characteristics of the project:**

- Is the project worth producing guidelines? Consider volume, duration, complexity
- What is the impact of potential errors in this MTPE? Consider purpose and quality level required

3. **Related to the team:**

    - Have the post-editors experience as such and, especially, in this type of contents and subject area?

- Have they received MTPE training?

4. **Related to the MT engine:**

- Does the MT engine produce atypical errors?

Do translators use machine translation and if so, how? Results of a survey held among professional translators

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

Abstract

The author conducted an anonymous online survey between 23 July and 21 October 2022 to gain insight into the proportion of translators that use machine translation (MT) in their translation workflow and the various ways they do. The results show that translators with more experience are less likely to accept MT post-editing (MTPE) assignments than their less experienced colleagues but are equally likely to use MT themselves in their translation work. Translators who deal with lower-resource languages are also less likely to accept MTPE jobs, but there is no such relationship regarding the use of MT in their own workflow. When left to their own devices, only 18.57% of the 69.54% of respondents that declared that they use MT while translating always or usually use it in the way the pioneers of MT envisaged, i.e., MTPE. Most either usually or always prefer to use MT in a whole range of other ways, including enabling MT functions in CAT tools and doing *hybrid post-editing*; using MT engines as if they were dictionaries; and using MT for *inspiration*. The vast majority of MT users see MT as just another tool that their clients do not necessarily need to be informed about.

Introduction

Right from the early days of machine translation (MT), it was apparent that totally replacing humans with machines for all kinds of translation was not a realistic goal since, as Warren Weaver put it in his ground-breaking memorandum, "perfect translation is almost surely unattainable" (Weaver, 1949). This was further underlined by Yehoshua Bar-Hillel, organizer of the first Conference on Mechanical Translation in 1952, who reasoned that fully automatic high-quality machine translation was not feasible. In his theoretical demonstration, Bar-Hillel described the need for post-editing "not only for polishing up purposes" but also to deal with ambiguity which is "resolvable only on the basis of extra-linguistic knowledge" (Bar-Hillel, 1960).

From these beginnings, it looked as if MT post-editing (MTPE) was shortly destined to become the predominant approach to translation, at least for technical and scientific texts. However, a few years later, the 1966 report published by the Automatic Language Processing Advisory Committee (ALPAC, 1966) cast doubt on its economic viability. The Committee concluded that, at the time, human translation could be done "faster and for less than half the price". The ALPAC report did on the other hand promote the use of *machine-aided translation*, later known as *computer-aided translation* (CAT), which in 1966 consisted of using *text-related glossaries* compiled with the help of a computer.

After the ALPAC report, MTPE underwent a period of what Garcia (2012) defines as latency. Post-editing was still used in various projects throughout the world, but attention gradually shifted towards CAT tools, which "grew out of MT developers' frustration at being unable to design a product which could truly assist in producing faster, cheaper and yet still useable translation" (Garcia, 2014). Initially, MT systems and CAT tools followed two separate paths of development although some attempts were made at integrating CAT tools with MT in the early 1990s. However, it was not until Lingotek produced a web-based CAT tool with MT integration in 2006 (Garcia, 2014) that the barrier between the two approaches began to break down.

CAT-MT integration makes what this paper terms as *hybrid post-editing* possible, i.e., a process whereby part of the translation is done through the post-editing of MT output and part through the editing of translation memory matches. Several CAT tools today offer even more complex features such as the automatic *repair* of translation memory matches using MT output and MT-output-based predictive typing, which make it hard to determine which type of editing the human translator is doing. Moreover, some recent studies on the two types of editing, particularly Sánchez-Gijón (2019) and do Carmo and Moorkens (2020), have noted the blurring lines between the two processes caused by improvements in the quality of MT output.

This paper presents the results of an anonymous online survey designed to gain insight into the proportion of translators that use MT during their work and the various ways they do so. Several surveys have already been conducted on the use of technology in the translation industry, and some of them also set out to measure the degree of use of MT among translators, notably the QTLaunchPad survey (Doherty et al., 2013), the Use of Machine Translation among Professional Translators survey (Zaretskaya, 2015) and the annual European Language Industry Surveys published by ELIA, et al. However, to this author's knowledge, there have been no surveys designed to obtain details of precisely how freelance translators choose to include MT in their workflow from among the whole host of options available to them. This paper intends to fill that gap.

Methods

The anonymous online survey was drawn up in English, due to its international nature. The questions were inspired by an informal discussion the author launched in a private Facebook group (Translators in Italy) in February 2022, which was a *de facto* brainstorming session on how professional translators use MT during their work. The various techniques that emerged from the discussion allowed closed-ended survey questions to be designed, with the advantage of making result analysis simpler and the survey less time-consuming to take. In any case, additional *other (please specify)* options were provided so that answers that did not emerge during the brainstorming session could still be given.

Since Zaretskaya (2015) reports that translators with advanced knowledge of IT tend to use MT more than others, it was initially decided *not* to post the survey on public websites or social media but to ask professional translators' associations to share it with their members in the hopes of reaching people with a broad range of IT skills.

The survey link was sent to 97 associations on 23 July and 2 others on 23 August 2022, ninety-five of which were members of the International Federation of Translators. With a large population, it is commonly estimated that 385 replies are sufficient to reach a confidence level of 95%, assuming the sample is truly random. This amounts to responses from fewer than four members of each association contacted.

However, in early September, it became apparent that very few associations were willing to take part in the research: only 11 had written to say they had shared the link and one large one had replied that the survey did not align with their mission. Since the total number of responses stood at 249 on 8 September, including some incomplete ones, and the response rate was beginning to flag badly, the author decided to share the survey on Facebook, LinkedIn, Twitter, and ProZ.com using a different link (collector). Moreover, when the abstract of the presentation of this paper was published on the *Translating and the Computer 44* website, an additional question was added to identify any responses from the new channel. The data from the two populations (survey received through an association vs. survey found in a *technological* way) could therefore be analysed separately.

Most of the variables measured in the survey are non-numeric, non-parametric, categorical variables which can only take on a limited number of values, and several of the continuous, numerical variables, such as years of experience, were analysed in bands of values and therefore

transformed into categorical variables. For this reason, the widely used chi-square (χ^2) test was chosen for the statistical analysis. The significance level was set to .05, as per convention, to ensure a 95% confidence level, and the online chi-square test calculator provided by Dr Jeremy Stangroom was used.[1] The results are reported in the format required by the American Psychological Association (APA): χ^2 (degrees of freedom, N = sample size) = chi-square statistic value, p = p value.

Results and discussion

1.1 Survey population

The survey closed as scheduled on 21 October 2022. A total of 12 of the 99 professional associations contacted had written to say they had shared the survey link with their members, although it was discovered by chance that at least 3 others had also done so without informing the author. One had written to say they would *not* share the link and none of the other associations replied at all. Survey responses were received from 452 people: 6 were disqualified since they answered that they were *not* professional translators; 301 were sent the survey link by a professional association or a member thereof (group A); 145 received the survey link from social media or a website, or from someone who found it that way (group B). Two responses were so incomplete they could not be used; other incomplete responses were used up to the question they reached.

The first step in the analysis is to see if the two groups of respondents gave significantly different replies regarding the key questions: use of MT and willingness to accept MTPE assignments.

	Never MTPE	**MTPE**
Group A	136	148
Group B	60	77
(χ^2 (1, N = 421) = 0.62, p = .430).		

Table 5: MTPE contingency table

	Use MT	**Never use MT**
Group A	197	83
Group B	93	44
(χ^2 (1, N = 417) = 0.27, p = .606).		

Table 2: Use of MT contingency table

In both cases the answers to the questions were independent of the group the respondent belonged to (p > .05). This may be because what Zaretskaya observed in 2015 no longer holds, or because frequenting social media and the internet is not indicative of a particularly high level of IT skill, or because predominantly tech-savvy association members tend to reply to online surveys. Whatever the explanation, there is no reason to keep the data separate from hereon in.

1.2 Respondents

The first questions aimed at getting a picture of how much experience the respondents have, the languages they work with and the way they work (freelance, in-house, etc.) to see if these factors affect their attitude towards MT. The mean professional experience was calculated at

[1] https://www.socscistatistics.com/tests/chisquare2/default2.aspx

21.00±12.38 years.[2] Table 3 shows willingness to accept MTPE jobs according to years of experience. The bands were chosen so that there is approximately the same number of respondents in each.

Years of experience	Never MTPE	Sometimes MTPE	Often MTPE
0-12	30	49	24
13-19	44	45	9
20-28	57	41	16
29-70	64	33	7

Table 3: Acceptance of MTPE jobs according to experience

As experience grows, the likelihood of accepting post-editing assignments falls in a statistically significant way (χ^2 (6, N = 419) = 29.01, p < .01). Table 4 shows the number of respondents that reported they use MT as an aid at some point in their translation workflow according to years of experience for the same ranges. Perhaps surprisingly, there is no statistically significant difference (χ^2 (3, N = 415) = 0.39, p = .941). *Young* and *old* translators are just as likely to use MT in their personal translation process.

Years of experience	Use MT	Never use MT
0-12	70	32
13-19	66	31
20-28	81	32
29-70	72	31

Table 4: Use of MT according to experience

Table 5 shows how willing the respondents are to accept MTPE assignments according to how much of their work consists of translation, expressed as a percentage of all the language services (LSs) the translator provides. The bands were chosen so that there is approximately the same number of respondents in each. No statistically significant relationship was found (χ^2 (6, N = 421) = 10.20, p = .116).

Translation as % of LSs	Never MTPE	Sometimes MTPE	Often MTPE
1-60	45	42	18
61-80	45	52	12
81-95	49	43	9
96-100	57	32	17

Table 5: MTPE according to translation as a percentage of all the language services provided

Translation as % of LSs	Use MT	Never use MT
1-60	73	30
61-80	85	24
81-95	60	39
96-100	72	34

Table 6: Use of MT according to translation as a percentage of all the language services provided

[2]Two respondents indicated that they had 100 years of professional experience. Their data were not considered plausible and discarded.

No statistically significant relationship was found regarding the use of MT in the workflow either (χ^2 (3, N = 417) = 7.62, p = .055).

91.90% of respondents were freelance translators, 6.71% were in-house employees, 5.56% were employees working from home, 6.25% were volunteer translators and 1.16% had a different working relationship. Multiple answers were allowed since translators may work part-time in different ways.

Employees might have been expected to accept more MTPE jobs and use MT more often than freelancers, but the survey data shows that these two variables are independent of the way the profession is practiced (χ^2 (4, N = 471) = 4.07, p = .396 and χ^2 (4, N = 466) = 5.99, p = .200).

Employees were asked if the organization they worked for dictated the way they could use MT in their workflow. Only one answer was allowed. 64.86% of respondents said no rules were imposed, 5.41% said they were obliged to use MT and 29.73% are allowed to use MT in certain circumstances.

The circumstances mentioned amounted to not being allowed to use MT for specific jobs where privacy was an issue (1), being allowed to use MT within CAT tools (3), and being obliged to use MT if explicitly requested by the end client (7).

1.3 Translation languages

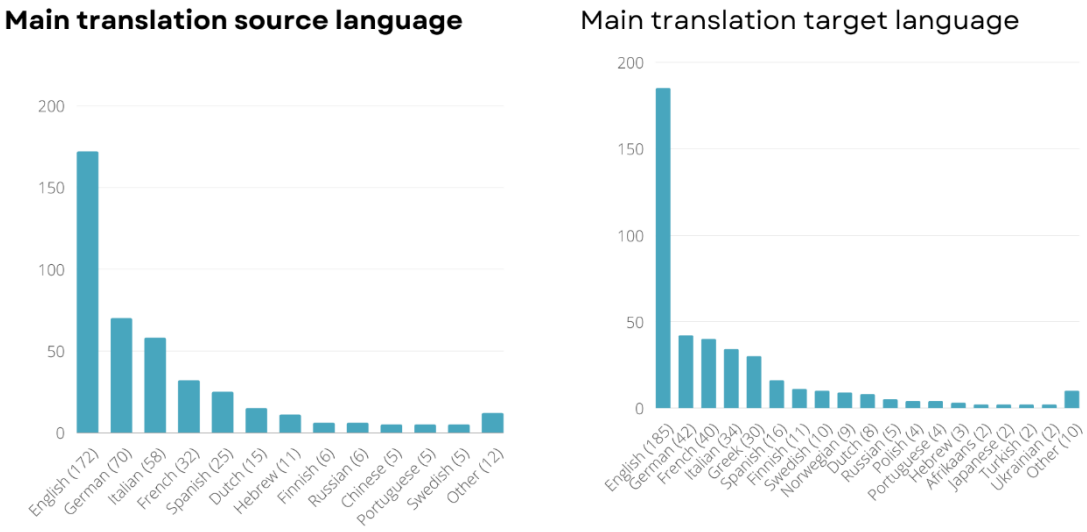


Chart 1: Main translation source and target languages

Professional translators might be expected to be more likely to consider post-editing assignments and use MT in their workflows if they work with higher-resource languages, for which the quality of MT output is normally better. To verify this hypothesis, the Digital Language Equality Metric (technological factors) was used as a measure of language resource richness (Gaspari et al., 2022). Only 22 of the 31 languages reported by respondents are rated on the European Language Grid Dashboard[3], but – fortunately – those languages account for 94% of the overall source language data and 96% of the overall target language data gathered in this survey.

Upon analysis, it was found that there seems to be a threshold under which professional translators are less likely to accept MTPE jobs (source language TDLE score of somewhere

[3] Consulted on 25 October 2022.

between 13807 and 14765 and target language TDLE score of somewhere between 14765 and 15414). However, as Zaretskaya (2015) observed, there is no such threshold as regards using MT in the workflow.

Respondents did an average of 81.98±18.76% of their work in their main language pair. Table 7 shows willingness to accept MTPE jobs according to the amount of translation work the translator does in their main language pair. The ranges were chosen so that there is approximately the same number of respondents in each group. No significant relationship was found (χ^2 (6, N = 421) = 2.02, p = .918).

Amount of work in main language pair (%)	Never MTPE	Sometimes MTPE	Often MTPE
100	56	41	14
90-99	47	43	13
70-89	54	47	19
10-69	39	38	10

Table 7: Acceptance of MTPE according to the proportion of work the respondent does in their main language pair

A similar contingency table was drawn up between the amount of work a professional translator does in their main language pair and whether they use MT in their workflow, again without finding any significant relationship (χ^2 (3, N = 417) = 2.06, p = .561).

1.4 Acceptance of MTPE assignments

46.56% of respondents said they never accept MTPE jobs. They were allowed to give multiple answers to explain why: “I refuse to do them” (48.21%), “I have never been offered one” (28.72%), “the rates offered are too low” (44.62%), and “other” (36.41%). The four most frequent open-ended *other* answers given amounted to (in decreasing order of frequency) a dislike for or little satisfaction from post-editing (one respondent used the expression “soul destroying”), post-editing requiring as much or more time than translation from scratch, MT giving poor results in the translator’s field of specialization, and MT output being a bad influence on the translator or leading to bad translation habits.

Some translators reported that they suspected or were sure that some of the translations they were given to revise were in reality MT output or MTPE done by non-native speakers of the target language even though they were told they were human translations or texts written by non-native speakers. These might be described as *stealth monolingual post-editing assignments*.

40.14% of respondents said they sometimes accept MTPE jobs. They were allowed to leave multiple closed-ended comments to add detail to their answer: “but I prefer to avoid them” (51.79%), “but I do not actively seek them” (60.71%), and “I am not often asked to do them” (32.14%). Respondents could also leave an open-ended comment (10.12%). The main two amounted to – from most to least common – “only if the rate is right” and “maybe I am doing them without being told”, as discussed above.

13.30% of respondents said they often accept MTPE jobs. Again, they were allowed to leave multiple closed-ended comments. 33.93% said they preferred to avoid them, and 16.07% said that they actively seek them. Respondents could also make another comment (51.79%) not included among the closed-ended answers. The vast majority of those who wrote something said that post-editing is simply another language service, and several comments seemed tinged with melancholic resignation: “because - while I don't love them - I cannot turn a blind eye to MT and pretend it's not there.”

One reason why so many translators seem to dislike post-editing may be that the rewarding part of the translation process lies in the sense of achievement attained when you elegantly

express the same concept in the target language. Post-editing mostly removes this task leaving the translator the chore of dotting the i's and crossing the t's, which is felt to be less satisfying.

1.5 Use of MT at some point during the translation workflow

69.54% of respondents use MT at some point in their translation workflow (MT users). This figure is virtually the same as the *slightly more than 70% of independent professionals* reported in the 2022 European Language Industry Survey (ELIA, et al., 2022). No significant relationship was found between willingness to accept MTPE jobs from clients and using MT as an aid while translating (χ^2 (2, N = 417) = 1.45, p = .485).

The respondents that said they never use MT at any point in their translation workflow gave the following reasons (multiple answers were allowed): "because the kinds of texts I translate do not lend themselves to machine translation" (51.64%); "because it harms the quality of the final translation" (42.62%); "because of GDPR/privacy issues" (34.43%); "I have experimented with it but do not find it useful" (31.97%); "I have never tried to integrate it into my workflow" (29.51%); "because my employer/client(s) specifically ask(s) me not to use it" (18.85%); "because it is *unprofessional*" (16.39%), and "other" (20.49%).

Among the *other* open-ended answers given, three respondents said that MT quality was not good enough in the languages they worked with, two said they could not afford good MT output, two did not want to provide the engines with training data and put their jobs at risk, one said it harms their language skills and one only translates handwritten documents.

1.6 MT engines

81.40% of MT users said they use one or more cloud or web-based MT engines, as shown in Chart 2 (multiple answers were allowed).

MyMemory is a large public translation memory and not an MT engine. However, the service also provides machine translations from Google Translate and Microsoft Translator.[4] 0.85% of web-based MT engine users said they pay to use the following MT engines (multiple answers were allowed): DeepL (102), Google Translate (20), ModernMT (9), Microsoft Translator (4), and other engines (7). The others use the free versions.

18.59% of MT users use custom MT engines (multiple answers were allowed): 37 of these use engines provided by employers/clients, 17 use their own engine and 3 use other engines. ModernMT, mentioned in the question about web-based engines, utilizes user-uploaded corpora (translation memories) and adds translated segments to its training data on the fly (Germann et al., 2016). It should therefore be regarded as a custom MT engine built by the translator. However, 8 of the 9 respondents that stated they use ModernMT said that they did *not* use custom MT engines, possibly because they did not know what a custom MT engine is. The respondent that answered that they use KantanMT, on the other hand, replied correctly. With hindsight, perhaps the survey question should have provided a definition of the term. The data given above has been adjusted to include the ModernMT users, but it would be reasonable to assume the true figures might be higher.

1.07% of MT users said they use one or more non-web-based MT engines, not including custom MT engines. Only one person named a non-web-based MT engine: OPUS-CAT. One translator said their clients use a non-web-based MT engine without stating which. And one respondent said that their client provides a penalized translation memory containing MT output. This working method is also suggested in a training manual on using MT with the CAT tool memoQ (Pawelec, 2021).

[4]https://site.matecat.com/support/managing-language-resources/machine-translation-engines/

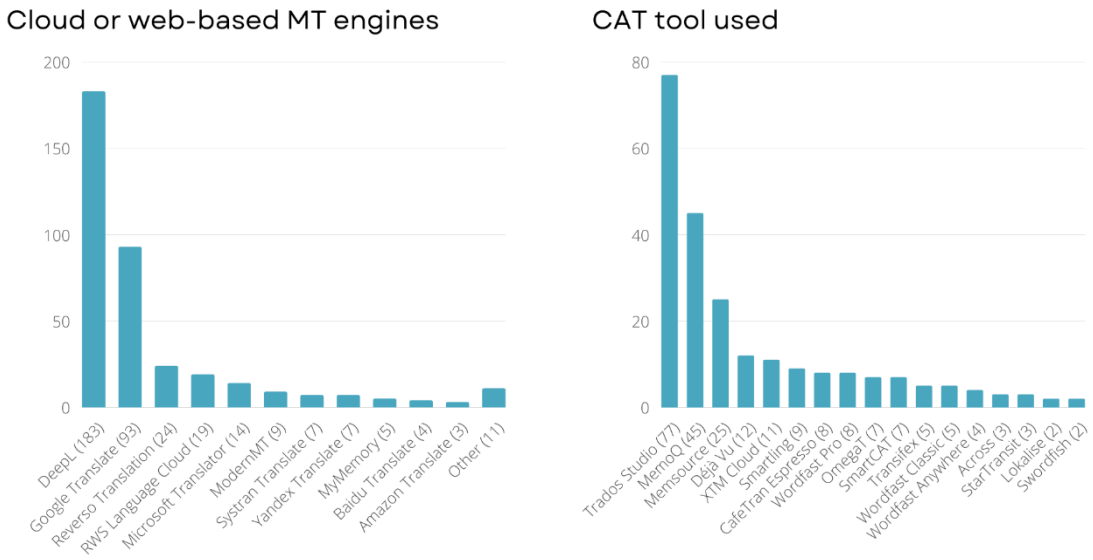


Chart 2: Cloud or web-based MT engines and CAT tools used

1.7 *Pure post-editing*

51.79% of MT users reported they do *pure post-editing* (always = 5.36%, usually = 13.21%, sometimes = 16.43%, rarely = 16.79%), which is when the translator decides to deal with their own translation project as if it were a post-editing assignment. In other words, they receive a source text to translate from their client, machine-translate the entire text, and then carry out a full post-editing on the output. This can be done in a CAT tool or by feeding the source text to an MT engine and post-editing the output file in a word processor. Perhaps unsurprisingly translators who do not accept MTPE assignments from clients are also less likely to do *pure post-editing* for themselves (χ^2 (1, N = 406) = 7.31, $p < 0.05$).

	All MT users	**Those who never accept MTPE assignments from clients**
Pure MTPE	145	47
No pure MTPE	135	79

Table 8: Translators who do *pure post-editing*, all MT users vs. those who do not accept MTPE assignments

1.8 *Hybrid post-editing*

33.21% of MT users do *not* use or enable MT functions in their CAT tools and 13.72% do not use CAT tools at all. The remaining 53.07% enable or use MT in the ways shown in Table 9 (multiple answers were allowed). By enabling MT functions, many of the translators are effectively doing *hybrid post-editing*, in other words, a process whereby part of the translation is done through the post-editing of machine translation output and part through the editing of translation memory matches. The CAT tools which respondents reported they enabled MT functions in are shown in Chart 2 (again multiple answers were allowed).

Machine translation when there is no exact match	55.78%
Machine translation when there is no good fuzzy match	43.54%
Machine translation to integrate or *repair* fuzzy matches	16.33%
Machine translation through predictive typing	20.41%
Other way (please specify)	23.13%

Table 9: How MT is enabled in CAT tools

Six respondents used the *other way* reply to specify that they keep the MT output in a side CAT tool window and only copy it into the translation if they think it is useful. Three others machine-translate whole paragraphs or the whole document and keep the output as reference, in one case in the form of a translation memory.

1.9 MT as a dictionary

In this use, the translator takes a single word, expression (phrase), or whole sentence and feeds it to an MT engine. This can be done with a specific function inside a CAT tool by selecting a segment or part thereof. It can also be done when using a word processor to do a translation with add-ons, such as GT4T[5] or IntelliWebSearch[6], which can even be used as alternatives to enhance the built-in MT functions in CAT tools. A less sophisticated technique entails the translator simply opening an online MT engine in a browser window and copy-pasting parts of the text.

77.93% of MT users use MT engines as if they were dictionaries in the following ways (multiple answers were allowed): by feeding in whole sentences to find the translation of an expression in context (67.70%); by feeding in whole sentences to find the translation of a single word in context (65.93%); by feeding in expressions on their own (63.72%); by feeding in single words on their own (46.46%); by feeding in lists of related terms, e.g. nations, species of plants, names of pharmaceuticals, etc. (21.24%) and *other similar ways* (8.41%).

18.58% of respondents reported they prefer to use an MT engine for the purposes described above rather than using a traditional dictionary. One respondent specified that they fed their queries to two different MT engines to have a "range of options". It should be noted that the web interface of all the top eight engines shown in Chart 2, excluding ModernMT, give dictionary-like results if a single term is input, complete with definitions and alternative translations. The DeepL web interface also gives alternative translations for whole segments and, together with Systran, allow the user to click on any word in a segment (source or target) to see a definition of that word.

1.10 MT for *inspiration*

This use also regards individual sentences, words, or expressions (phrases), much as described for the dictionary-like uses, but this time the aim is not to solve a vocabulary problem, but to be *inspired*. One respondent clarified how this can be done: "I translate passages or sentences myself and then use the MT on the source text to see what it comes up with, and I may adjust my translation on that basis or indeed completely ignore the MT text. The MT never takes the lead but can sometimes be useful as a supplement."

A total of 86.21% of MT users use MT this way.

74.80% of MT users use MT to overcome what Michael Cronin defines as blockage, when – as he puts it – the "word or the expression or the equivalent allusion will not come, the textual

[5]https://gt4t.net

[6]https://www.intelliwebsearch.com/version-5/api/

whole does not seem the right fit and try as you might, there seems to be no way out, the words refuse to come to your rescue" (Cronin, 2003).

86.40% of MT users use MT for a second opinion when they are not entirely happy with their own translation of a word, phrase, or sentence.

The concept of using MT to escape from one's idiolect and add variety to a text, mentioned by 59.20% of MT users, seems to contradict the findings of some authors that report that MT leads to lexical impoverishment (Farrell, 2018; Volkart, 2022). However, if the translator already has a solution in mind or has previously translated a word or expression in a certain way elsewhere in the same text and is looking for a synonym, then using an MT proposal instead of their own idea has the effect of adding variety, which can be an important factor in the quality of the translation of creative texts.

In the *other similar way* box (9.60%), one respondent wrote "I feed larger chunks of text into DeepL, sometimes paragraphs […] This enhances the quality of the output since the MT has more context." In December 2020, the author carried out a series of experiments which revealed small differences in the translation DeepL provides when fed whole paragraphs rather than the single sentences that make up the same paragraphs. This feature is however not documented on the DeepL website (last consulted on 23 September 2022).

1.11 MT for comic relief

25.86% of MT users reported that they use MT for an occasional giggle to brighten up their working day. However, several translators used the *other similar way* box (completed by 22.67% of respondents) to clarify that they do not intentionally use it this way but enjoy the odd chuckle when MT happens to produce entertaining output.

1.12 Other uses of MT

The only other uses of MT in the translation workflow that truly do not fit into one of the previous categories (3.7 to 3.11) were the back translation of incomprehensible parts of source text written by non-native speakers into their native language (3 respondents), and as a sort of double-check to prevent omissions or mistakes during the revision process (1 respondent). All the other replies could be reclassified as answers to other questions.

1.13 Transparency

Respondents were asked if they tell their employer/client(s) that they use MT in their workflow.

Always	8.49%
Sometimes	25.83%
Never	65.68%

Table 10: Answer to "do you tell your employer/client(s) you use MT in your workflow?"

Those who answered *sometimes* also specified when. The most common replies - in descending order of frequency – are "if asked", "when the client has specifically asked for MT to be used", "when the translator decides to do *pure MTPE*", "when I think they should know" and "when they know already".

Respondents were then asked if they explained precisely how they use MT when they tell their employer/client(s) that they use it.

Always	25.81%
Usually	17.20%
Sometimes	20.43%
Rarely	10.75%
Never	25.81%

Table 11: Precise explanation of the use of MT

1.14 Other language pairs

86.62% of MT users who work with more than one language pair reported that there are no significant differences in the way they use MT in pairs other than their main one. The reasons given by the respondents who use MT in a different way according to language pair can mainly be categorized as (from most to least common): "MT output is better/worse for the other language(s)" and "because my knowledge of the other language(s) is weaker".

Conclusion

This paper reports the results of an anonymous online survey conducted between 23 July and 21 October 2022 designed to establish the proportion of translators that use MT in their translation workflow and the various ways in which they do.

Although it was found that translators with more experience are less likely to accept MTPE assignments than their less experienced colleagues, it was seen that they are equally likely to use MT themselves in their own translation work.

As might be expected, translators who work with lower-resource languages are less likely to accept MTPE jobs, but – perhaps surprisingly – there is no such relationship regarding the use of MT in their workflow.

Attitude towards using MT and accepting MTPE jobs was also found *not* to depend on how much of a professional translator's work consists of translation compared with the other language services they provide, the way the translator works (freelancer, in-house, etc.) or the proportion of translation work the translator does in their main language pair.

When left to their own devices, only 18.57% of the translators who use MT in their workflow (69.54%) always or usually use it in the way the pioneers of MT envisaged, i.e., MTPE. Most either usually or always prefer to use MT in a wide range of other ways. These may be classified as using or enabling MT functions in CAT tools and doing *hybrid post-editing*; using MT engines as if they were dictionaries; using MT for *inspiration*; and even using it for comic relief, although this seems more likely to be incidental rather than deliberate.

The vast majority of MT users (91.51%) do not feel that it is always necessary to inform their employer/client(s) that they use MT in their workflow and 65.68% never do so. The impression is that translators today see MT as just one of the many tools they have available to them and not so special as to need pointing out.

References

Automatic Language Processing Advisory Committee. 1966. Language and machines - Computers In translation and linguistics. National Academy of Sciences, Washington DC.

Cronin, Michael. 2003. Translation and globalization. Routledge.

Bar-Hillel, Yehoshua. 1960. The present status of automatic translation of languages. In Advances in Computers 1, pages 91-163.

do Carmo, Félix and Joss Moorkens. 2020. Differentiating Editing, Post-Editing and Revision. In Translation Revision and Post-Editing, Routledge, pages 35-49.

Doherty, Stephen, Federico Gaspari, Declan Groves, Josef van Genabith, Lucia Specia, Aljoscha Burchardt, Arle Lommel and Hans Uszkoreit. 2013. QTLaunchPad – Mapping the Industry I: Findings on Translation Technologies and Quality Assessment. European Comission Report. Technical report.

ELIA, et al., European Language industry Survey 2022. Retrieved from https://ec.europa.eu/info/sites/default/files/about_the_european_commission/service_standards_and_principles/documents/elis2022-report.pdf

Farrell, Michael. 2018. Machine Translation Markers in Post-Edited Machine Translation Output. In Proceedings of the 40th Conference Translating and the Computer, pages 50–59.

Garcia, Ignacio. 2012. A Brief History of Postediting and of Research on Postediting. In Revista Anglo Saxonica, pages 291-310.

Garcia, Ignacio. 2014. Computer-Aided Translation. In Routledge Encyclopedia of Translation Technology, pages 68-87.

Gaspari, Federico, Owen Gallagher, Georg Rehm, Maria Giagkou, Stelios Piperidis, Jane Dunne and Andy Way. 2022. Introducing the Digital Language Equality Metric: Technological Factors. In LREC 2022 Workshop Language Resources and Evaluation Conference, page 1.

Germann, Ulrich, E. Barbu, L. Bentivogli, N. Bertoldi, N. Bogoychev, C. Buck, D. Caroselli, L. Carvalho, A. Cattelan, R. Cattoni, M. Cettolo, M. Federico, B. Haddow, D. Madl, L. Mastrostefano, P. Mathur, A. Ruopp, A. Samiotou, V. Sudharshan, M. Trombetti, J. Van der Meer. 2016. Modern MT: a new open-source machine translation platform for the translation industry. Baltic Journal of Modern Computing, 4(2), pages 397-397.

Pawelec, Marek. 2021. memoQ Inside Out: Machine Translation. Wasaty, page 1.

Sánchez-Gijón, Pilar, Joss Moorkens and Andy Way. 2019. Post-Editing Neural Machine Translation versus Translation Memory Segments. Machine Translation 33(1), pages 31–59.

Volkart, Lise and Pierrette Bouillon. 2022. Studying Post-Editese in a Professional Context: A Pilot Study. In Proceedings of the 23rd Annual Conference of the European Association for Machine Translation 2022, pages 71-79.

Weaver, Warren. 1949. Translation. In Proceedings of the Conference on Mechanical Translation (1952).

Zaretskaya, Anna. 2015. The use of machine translation among professional translators. In Proceedings of the EXPERT Scientific and Technological Workshop, pages 1–12.

Peculiarities of Polish academic legal writing in English translation: field experts vs. algorithms

Anna Setkowicz-Ryszka *University of Lodz*

Email: anna.setkowicz.ryszka@edu.uni.lodz.pl *ORCiD 0000-0003-0057-3477*

Keywords: *machine translation; post-editing; revision; English as a Lingua Franca; academic legal writing*

1. Introduction

This paper discusses lexical problems encountered in academic legal writing (Section 2), the challenges of using English as a Lingua Franca (Section 3), and the challenges, but also benefits of post-editing machine-translated texts, including legal ones (Section 4). Section 5 provides examples of peculiar features of Polish academic legal language, which cause problems in translation, compares their renderings by legal scholars and machine translation, and suggests preferred translation solutions. Section 6 concludes that neither mode of translation is fully successful in dealing with such peculiarities.

2. Academic legal writing

According to Šarčević, "[u]nlike medicine, chemistry, computer science, and other disciplines of the exact sciences, law remains first and foremost a national phenomenon" (1997: 13). Therefore, differences between legal systems, especially lack of correspondence (asymmetry) between legal concepts, make legal translation a demanding kind of specialized translation (Scott, 2017). Legal translators need knowledge of the law, including comparative law skills (Doczekalska, 2013; Engberg, 2013; Jopek-Bosiacka, 2013), but also familiarity with legal writing conventions, including in particular legal genres (Orlando, 2015: 117). It is sometimes debated who makes better legal translators: bilingual lawyers or linguists who choose to specialize in legal translation (Cao, 2007: 5; Orlando, 2015: 76-77, 101-103).

The focus in legal translation studies has been on legislative texts, including inter- or supranational law (Biel, 2014; Cao, 2007; Šarčević, 1997) and judgments (Gościński, 2019; Pontrandolfo, 2015, 2018; Prieto Ramos, 2014a). Academic texts are rarely discussed in legal translation literature. Cao mentions 'scholarly legal texts', such as commentaries, whose legal status varies depending on the legal system (2007: 9, 83), Prieto Ramos explains that they include journal articles and textbooks, with heterogenous stylistic features (2014b: 263), while Alcaraz Varó & Hughes discuss features of law reports (2002: 108-112).

Due to requirements of internationalization of academia, legal scholars in Poland wish to communicate their findings to international audience. This is also required by legal journals that wish to appear in international databases: Scopus or Web of Science. Nowadays, at least titles, keywords and abstracts ,us be in English. Authors who are not fluent in academic legal English themselves and prefer not to hire a professional translator often translate themselves or use (generic) MT engines.

3. Challenges posed by English as a Lingua Franca

The first case, a do-it-yourself translation, can be expected to lead to unsatisfactory results. A comparison of translation processes and products of bilingual lawyers and final-year translation students translating into their native language (Orlando, 2015) revealed that lawyers translated faster, in a more word-for-word manner, consulted mainly bilingual dictionaries and Google Translate, and spent less time on revision. Despite better phraseology, lawyers' translations contained more errors than those of translation students and were mostly assessed as "borderline" quality. However, Lesznyák and Balogh found that translation trainees with humanities background made more errors

and focused more on surface structures than trainees with legal background (2019). When offered the possibility to consult experts, trainees lacking legal background asked questions that were irrelevant from the legal point of view (Lesznyák & Balogh, 2022).

Lawyers' translations can be seen as samples of English as a Lingua Franca (EFL) by untrained multilinguals (Albl-Mikasa, 2017: 369). ELF texts often display features such as transfers from the writers' first language, unusual choices of words or strange syntax (Albl-Mikasa et al., 2017: 372). Use of "non-English rhetorical patterns" can cause misunderstandings (Bennett & Queiroz de Barros, 2017: 366). Research found that ELF communication between non-native speakers is successful, with speakers managing to employ various repair and co-construction strategies when faced with odd structures (Kecskes & Kirner-Ludwig, 2019) or to create ad hoc rules of interaction (Kecskes, 2008). Therefore, ELF is seen as "not a defective, but a fully functional means of communication" (House, 2013: 286). But the fact that non-native English speakers are no longer expected to achieve native-like standard (Albl-Mikasa & Ehrensberger-Dow, 2019: 49) is frustrating for language mediators. EFL poses additional demands on their processing, especially in simultaneous interpreting ("*brain stoppers*"), but also in translation ("*flow blockers*") (Albl-Mikasa, 2017: 381, emphasis in original; Albl-Mikasa et al., 2020: 272-273).

A study on the process and product of translation of edited and non-edited ELF texts found that unedited texts were more likely to lead to mistranslations, with the most persistent translation problems caused by "non-standard lexical choices, such as cognates, L1-influenced expressions, calques, unusual collocations, etc.", much less so by "non-conventional use of articles and prepositions" (Albl-Mikasa et al., 2017: 375). Despite longer processing time, translations did not always reflect the writers' intentions. Greater effort was needed to disambiguate or normalize what was expressed vaguely in source texts (Albl-Mikasa et al., 2017: 380-385).

ELF is also challenging from a reviser's point of view, especially in monolingual revision. Despite the "shared language benefit" (Albl-Mikasa, 2017: 375), it may be difficult to understand the message, which has then to be expressed correctly (publishable quality), preferably also in a way that caters to the needs of international readers, presumably unfamiliar with the Polish legal system. Since the texts are written by law experts, they tend to be highly technical. Profound differences between the legal systems of English-speaking countries (common law) and the Polish legal system (civil law) make the asymmetry of legal concepts that usually accompanies legal translation particularly acute. Therefore, the use of ELF in legal scholarship is fraught with difficulties.

4. Challenges and benefits of MT post-editing, including in legal texts

Translation scholars are well aware that MT is more successful with controlled and routine texts rather than creative and unconventional ones (Biel, 2021: 23). Compared to revising good human translations, post-editing MT output (PEMT) is more challenging and greater cognitive effort involved in PEMT compared to revision is generally recognized (Biel, 2021: 24; O'Brien, 2022: 115-117), including the strong priming effect (Nitzke & Hansen-Schirra, 2021: 71). Do Carmo & Moorkens (2020) even argue that PEMT is closer to translation than revision.

There are certain categories of typical errors found in MT output. The Multidimensional Quality Metrics (http://themqm.info/typology/) contain eight main categories: Terminology, Accuracy, Linguistic conventions, Style, Locale conventions, Audience appropriateness, Design and markup, and Custom, with a number of sub-categories (over 80 issue/error types). TAUS guidelines suggest limiting the number of error categories, the most frequent being language, terminology, accuracy and style (2017: 5). Mossop identifies five items that need fixing most often: grammar errors, extra/missing words, term inconsistencies, proper names and stylistic problems (2020: 218). Biel's 13-item list of

frequent error types (2021: 25) includes mistranslations caused by polysemy or limited context, inability to deal with neologisms and slang, and lack of 'understanding' of context and cultural references.

It has been found that even error-free, i.e., acceptable translations may not always be adequate in the given context (Voita et al., 2019: 1199). Adequacy (fluency) errors were found to cause greater cognitive effort for post-editors than acceptability errors (Daems et al., 2017: 7, 12). And Koponen points out various levels of error severity or amounts of effort in terms of cognitive processing needed to detect errors and rewriting needed to correct them (2016: 26).

As demonstrated by EU eTranslation, MT systems can handle legislative, judicial and administrative texts when trained on such texts. Still, in the English-Polish pair, a study found modest productivity gains of PEMT compared to translation of EU texts, with greater gains for legislative texts than non-legislative ones. There were problems with terminological consistency, textual coherence, erroneous titles or quotes, wrong pronouns, etc., and errors were hard to detect in the fluent output (Stefaniak, 2020). Many errors in terminology and terminological inconsistencies were also reported in a study on English-Slovene eTranslation output. However, there were also errors caused by polysemous words, "neural neologisms" and other mistranslations (Arnejšek & Unk, 2020). More promising findings were reported in studies comparing PEMT and translation of court documents from English into Spanish with MT output obtained from generic phrase-based statistical Google Translate engine (Killman & Rodríguez-Castro, 2022) and PEMT and translation of normative texts from Greek into English using a custom-built NMT engine (Sosoni et al., 2022). Productivity and quality gains in PEMT scenario were reported in both studies, though cited literature pointed to problems. But although academic legal texts contain references to legislation or judgments, they often present novel approaches to issues where doubts or controversies arise.

5. Examples of problematic features of Polish legal academic language in ELF and MT

Analyses conducted by Polish linguists and lawyers focused on legislative texts and legal terminology, with academic genres mostly passing below the radar. The non-legislative legal language is described as hermetic and formalized, often unintelligible to non-lawyers (Kurek, 2015: 304-305). In addition, legal academics rarely realize the need for pre-editing (Biel, 2021:29-30). The articles need not be written in plain language because they are intended to be read by experts, but they might need adjusting to be understandable to international readers, unfamiliar with Polish law and peculiarities of Polish lawyers' language (Setkowicz-Ryszka, 2022).

In addition to the usual challenges of legal translation – system-bound terms, frequent abbreviations, "formulaic and elliptical usage", "deviations from normal language use" (Sosoni et al., 2022: 95) – the main difficulties posed by academic texts include complex syntax: sentences with multiple subordinate or embedded clauses, passive or impersonal constructions, and stilted language: deverbal nouns, Latinisms, low-frequency items or archaic expressions. It is mainly the stilted language that causes problems in MT output now, as severe acceptability errors have almost disappeared in neural MT. ELF texts can contain both acceptability and adequacy errors, even basic ones (poor grammar, orthography, unconventional collocations, wrong translation equivalents). The examples of ELF renditions used in this section come from English titles and abstracts submitted by legal academics together with articles for publication in legal journals. The publishing house sends them for revision to legal translators, including myself. It is likely that authors use MT when preparing English versions, though they should be considered to endorse the submitted English versions, regardless of how they were produced. Moreover, a study found no significant differences between ELF and self-PEMT versions created by academics from various disciplines (O'Brien et al., 2018), but

From my experience, MT (DeepL and eTranslation) output is more even in terms of quality than ELF texts, which vary considerably depending on the authors' proficiency levels. Table 1 below provides ten examples of Polish words or phrases whose renderings often need fixing in both ELF and MT texts. They all concern lexis: Latinisms and certain peculiarities of academic legal language. The resulting errors can be included in "Audience appropriateness", "Style/stylistic problems," or "Inability to deal with slang" (here, legal jargon) and "Mistranslations caused by polysemy" categories of the classifications mentioned in section 4. These lexical items are problematic because some ELF/MT equivalents are not used in English or are used sporadically (Latinisms 1-4 are not found in English-language corpora or Black's Law Dictionary), while others are used less often and/or in a different sense (items 5-10). 'Glossary' or 'gloss' are false friends of *glosa*, which does not mean an explanation, but a critical or approving commentary on a court judgment, while 'voice' or 'speech' must have appeared due to the similarity between *glosa* and *głos* (voice).

ST item	**ELF**	**MT**		**Suggested translations**
		DeepL (free version)	**eTranslation (general text)**	
(1) de lege lata	de lege lata	de lege lata	de lege lata *de lege years [omission]	as the law stands lex lata
(2) de lege ferenda	de lege ferenda			as the law should be lex ferenda
(3) vacatio legis	vacatio legis			[explanation]
(4) expressis verbis	-	expressis verbis	expressis verbis expressly	expressly
(5) glosa	*gloss *glossa commentary	*gloss *glossary *voice *glosa	*gloss *speech commentary	commentary case comment
(6) glosator	*glossator commentator	*glossator	*Glosator	commentator author
(7) doktryna	doctrine			legal scholarship legal literature
(8) instytucja prawna	legal institution			legal institution mechanism [concretization]
(9) ustawodawca	legislator			legislator/lawmaker parliament/legislature law/legislation
(10) przesłanka	*premise prerequisite	prerequisite condition *premise	condition	condition grounds prerequisite criterion principle

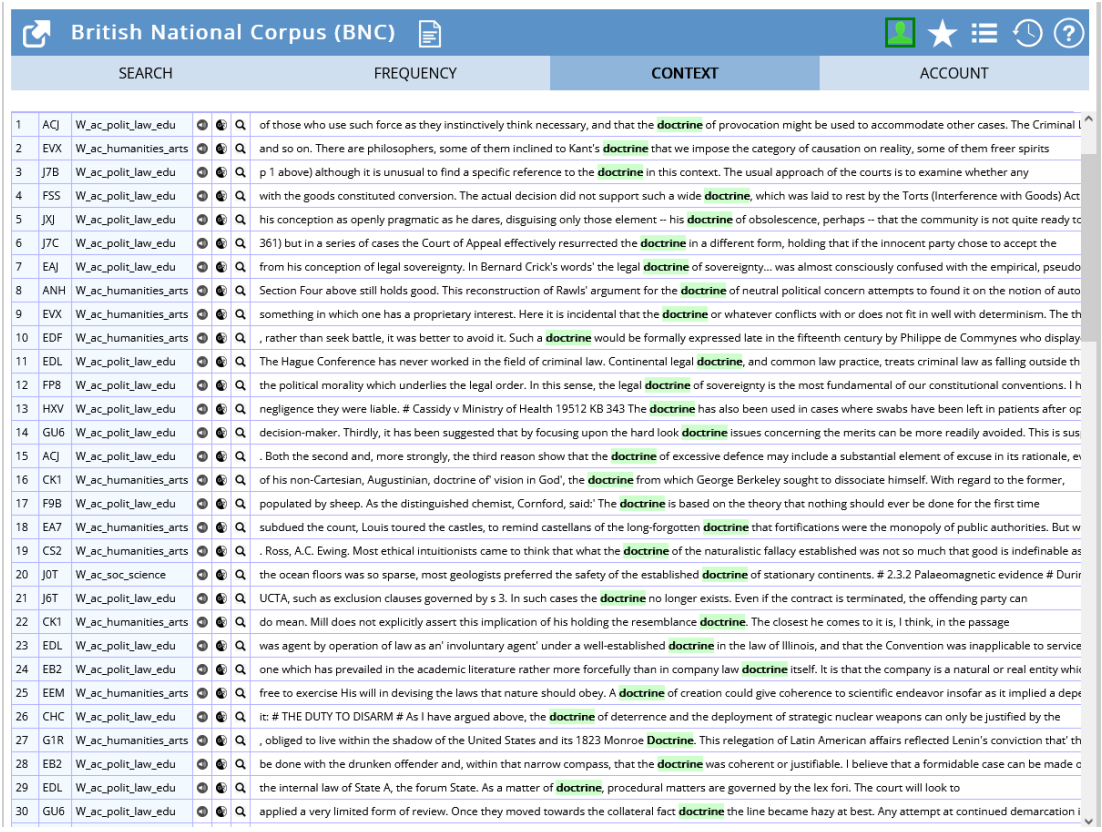
* accuracy errors

Table 6. Comparison of ELF and MT renderings of ten Polish problematic lexical items with suggestions of correct translations

There is little difference between ELF and MT renderings of the above items, especially 1-3 and 7-9. In order to spot and fix these problems, corpora or Internet resources have to be consulted to determine whether and how often literal equivalents are used. Often, particularly in case of items 8-10, contextualization is needed. Some eTranslation renderings of items 4, 5 and 10 are better than ELF or DeepL versions. Legal scholars writing in ELF seem not to realize that their use of Latinisms and certain words is particular or that they used them differently from their English-speaking colleagues. The broad range of meanings covered by item 10, *przesłanka*, rather than the basic meaning of 'premise' (according to the entry in Ling.pl dictionary), is particularly telling.

The English word *doctrine* can mean ‘teaching’, but nowadays this applies mostly to the religious context. Meanwhile, *(legal) institution* and *legislator* are possible translations of the Polish *instytucja (prawna)* and *ustawodawca*, though not the only ones. Interestingly, *legal institution* often serves to avoid repeating the name of that institution, while *ustawodawca* often appears as an entity that acts and succeeds or fails, when a reference to what the law now provides could be enough. The following concordance lines from academic tests for items 7-9 illustrate that in English:

- *doctrine* refers to a ‘principle’, rather than ‘legal scholarship’ (**Error! Reference source not found.**);
- *legal institution* does not only refer to abstract notions, as the Polish term does (*Figure 4*);
- *legislator* often denotes individuals, unlike in Polish, where it means ‘parliament’ (*Figure 5*).



1	ACJ	W_ac_polit_law_edu	of those who use such force as they instinctively think necessary, and that the **doctrine** of provocation might be used to accommodate other cases. The Criminal L
2	EVX	W_ac_humanities_arts	and so on. There are philosophers, some of them inclined to Kant's **doctrine** that we impose the category of causation on reality, some of them freer spirits
3	J7B	W_ac_polit_law_edu	p 1 above) although it is unusual to find a specific reference to the **doctrine** in this context. The usual approach of the courts is to examine whether any
4	FSS	W_ac_polit_law_edu	with the goods constituted conversion. The actual decision did not support such a wide **doctrine**, which was laid to rest by the Torts (Interference with Goods) Act
5	JXJ	W_ac_polit_law_edu	his conception as openly pragmatic as he dares, disguising only those element -- his **doctrine** of obsolescence, perhaps -- that the community is not quite ready to
6	J7C	W_ac_polit_law_edu	361) but in a series of cases the Court of Appeal effectively resurrected the **doctrine** in a different form, holding that if the innocent party chose to accept the
7	EAJ	W_ac_polit_law_edu	from his conception of legal sovereignty. In Bernard Crick's words' the legal **doctrine** of sovereignty... was almost consciously confused with the empirical, pseudo
8	ANH	W_ac_humanities_arts	Section Four above still holds good. This reconstruction of Rawls' argument for the **doctrine** of neutral political concern attempts to found it on the notion of auto
9	EVX	W_ac_humanities_arts	something in which one has a proprietary interest. Here it is incidental that the **doctrine** or whatever conflicts with or does not fit in well with determinism. The th
10	EDF	W_ac_humanities_arts	, rather than seek battle, it was better to avoid it. Such a **doctrine** would be formally expressed late in the fifteenth century by Philippe de Commynes who display
11	EDL	W_ac_polit_law_edu	The Hague Conference has never worked in the field of criminal law. Continental legal **doctrine**, and common law practice, treats criminal law as falling outside th
12	FP8	W_ac_polit_law_edu	the political morality which underlies the legal order. In this sense, the legal **doctrine** of sovereignty is the most fundamental of our constitutional conventions. I h
13	HXV	W_ac_polit_law_edu	negligence they were liable. # Cassidy v Ministry of Health 19512 KB 343 The **doctrine** has also been used in cases where swabs have been left in patients after op
14	GU6	W_ac_polit_law_edu	decision-maker. Thirdly, it has been suggested that by focusing upon the hard look **doctrine** issues concerning the merits can be more readily avoided. This is sus
15	ACJ	W_ac_polit_law_edu	. Both the second and, more strongly, the third reason show that the **doctrine** of excessive defence may include a substantial element of excuse in its rationale, ev
16	CK1	W_ac_humanities_arts	of his non-Cartesian, Augustinian, doctrine of' vision in God', the **doctrine** from which George Berkeley sought to dissociate himself. With regard to the former,
17	F9B	W_ac_polit_law_edu	populated by sheep. As the distinguished chemist, Cornford, said:' The **doctrine** is based on the theory that nothing should ever be done for the first time
18	EA7	W_ac_humanities_arts	subdued the count, Louis toured the castles, to remind castellans of the long-forgotten **doctrine** that fortifications were the monopoly of public authorities. But w
19	CS2	W_ac_humanities_arts	. Ross, A.C. Ewing. Most ethical intuitionists came to think that what the **doctrine** of the naturalistic fallacy established was not so much that good is indefinable as
20	J0T	W_ac_soc_science	the ocean floors was so sparse, most geologists preferred the safety of the established **doctrine** of stationary continents. # 2.3.2 Palaeomagnetic evidence # Durir
21	J6T	W_ac_polit_law_edu	UCTA, such as exclusion clauses governed by s 3. In such cases the **doctrine** no longer exists. Even if the contract is terminated, the offending party can
22	CK1	W_ac_humanities_arts	do mean. Mill does not explicitly assert this implication of his holding the resemblance **doctrine**. The closest he comes to it is, I think, in the passage
23	EDL	W_ac_polit_law_edu	was agent by operation of law as an' involuntary agent' under a well-established **doctrine** in the law of Illinois, and that the Convention was inapplicable to service
24	EB2	W_ac_polit_law_edu	one which has prevailed in the academic literature rather more forcefully than in company law **doctrine** itself. It is that the company is a natural or real entity whi
25	EEM	W_ac_humanities_arts	free to exercise His will in devising the laws that nature should obey. A **doctrine** of creation could give coherence to scientific endeavor insofar as it implied a depe
26	CHC	W_ac_polit_law_edu	it: # THE DUTY TO DISARM # As I have argued above, the **doctrine** of deterrence and the deployment of strategic nuclear weapons can only be justified by the
27	G1R	W_ac_humanities_arts	, obliged to live within the shadow of the United States and its 1823 Monroe **Doctrine**. This relegation of Latin American affairs reflected Lenin's conviction that' th
28	EB2	W_ac_polit_law_edu	be done with the drunken offender and, within that narrow compass, that the **doctrine** was coherent or justifiable. I believe that a formidable case can be made c
29	EDL	W_ac_polit_law_edu	the internal law of State A, the forum State. As a matter of **doctrine**, procedural matters are governed by the lex fori. The court will look to
30	GU6	W_ac_polit_law_edu	applied a very limited form of review. Once they moved towards the collateral fact **doctrine** the line became hazy at best. Any attempt at continued demarcation i

Figure 3. Random sample of concordance lines for “doctrine” in BNC 1994 (https://www.english-corpora.org/bnc/)

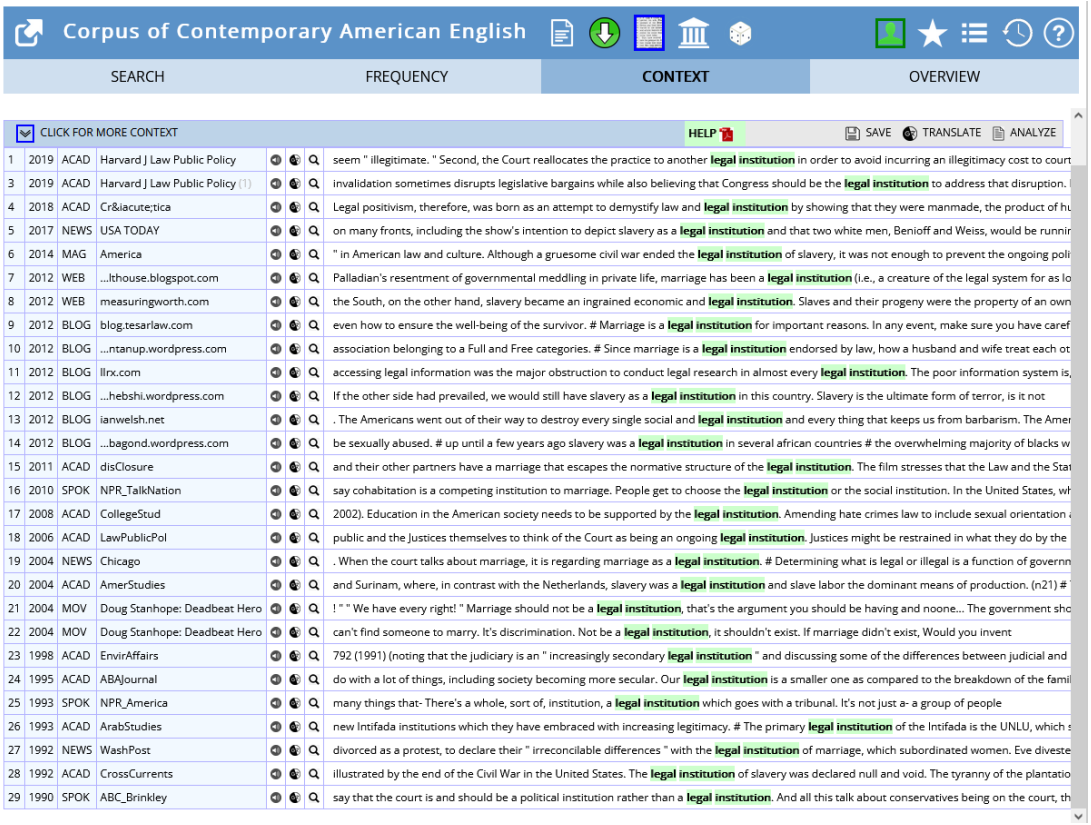


Figure 4. Concordance lines for "legal institution" in COCA (https://www.english-corpora.org/coca/)

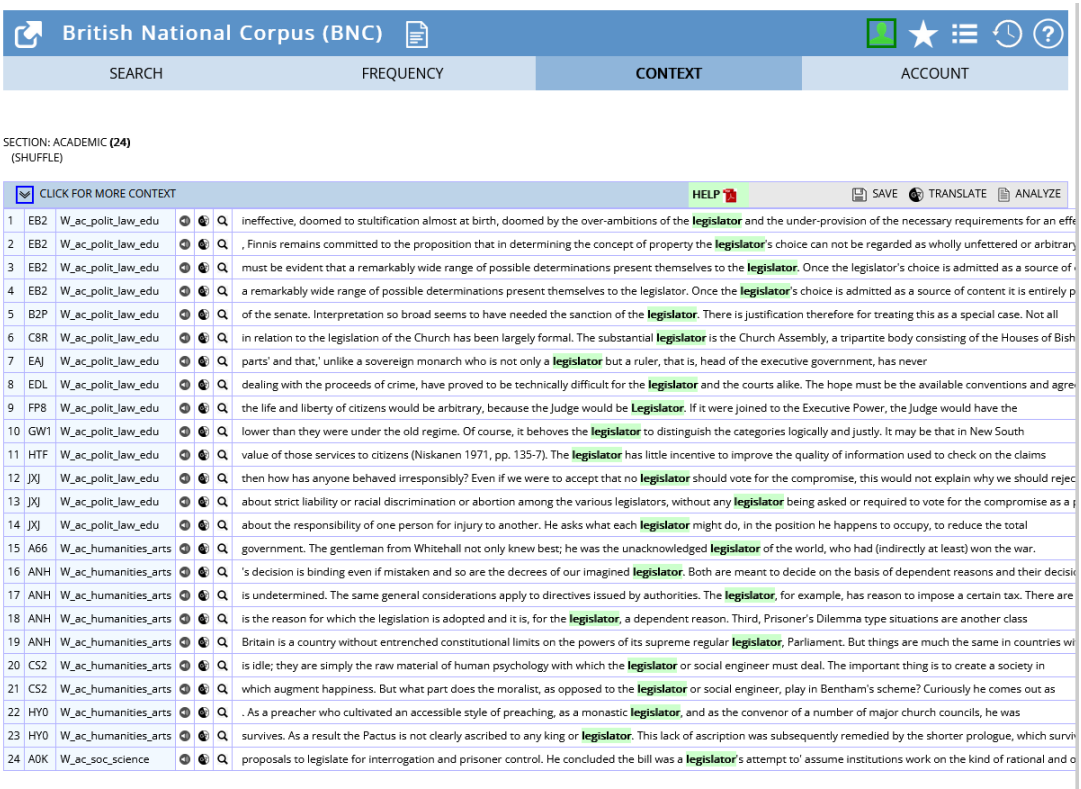


Figure 5. Concordance lines for "legislator" in academic texts of BNC 1994 (https://www.english-corpora.org/bnc/)

6. Conclusions

An analysis of ten problematic lexical items from Polish academic legal texts translated into English by bilingual lawyers and two MT engines shows that in both cases the target language conventions are

disregarded. As a result, target readers may struggle to understand the message. Legal scholars' own translations of lexical peculiarities of Polish legal language are often literal, sometimes less felicitous than MT output. In case of academic legal texts, both ELF revision and PEMT are cognitively challenging because considerable research effort is often required to determine if foreign readers will understand a given lexical item and to maintain target language conventions in these genres. When legal scholars wish to use MT for dissemination purposes, they need greater MT literacy (Bowker & Buitrago Ciro, 2019), both in terms of pre-editing or writing in an MT-friendly way, and in terms of awareness of risks and typical errors in MT output in the legal field.

REFERENCES

Albl-Mikasa, M. (2017). ELF in translation/interpreting. In J. Jenkins, W. Baker, & M. Dewey (Eds.), *The Routledge Handbook of English as a Lingua Franca* (pp. 369–383). Routledge.

Albl-Mikasa, M., & Ehrensberger-Dow, M. (2019). ITELF: (E)merging Interests in Interpreting and Translation Studies. In E. Dal Fovo & P. Gentile (Eds.), *Translation and Interpreting. Convergence, Contrast and Interaction* (pp. 45–62). Peter Lang.

Albl-Mikasa, M., Ehrensberger-Dow, M., Hunziker Heeb, A., Lehr, C., Boos, M., Kobi, M., Jäncke, L., & Elmer, S. (2020). Cognitive load in relation to non-standard language input: Insights from interpreting, translationand neuropsychology. *Translation, Cognition & Behavior*, *3*(2), 263–286. https://doi.org/https://doi.org/10.1075/tcb.00044.alb

Albl-Mikasa, M., Fontana, G., Fuchs, L. M., Stüdeli, L. M., & Zaugg, A. (2017). Professional translations of non-native English: 'before and after' texts from the European Parliament's Editing Unit. *The Translator*, *23*(4), 371–387. https://doi.org/10.1080/13556509.2017.1385940

Alcaraz Varó, E., & Hughes, B. (2002). *Legal Translation Explained*. St.Jerome Publishing.

Arnejšek, M., & Unk, A. (2020). Multidimensional assessment of the eTranslation output for English-Slovene. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 383–392. https://aclanthology.org/2020.eamt-1.41

Bennett, K., & Queiroz de Barros, R. (2017). International English its current status and implications for translation. *The Translator*, *23*(4), 363–370.

Biel, Ł. (2014). The textual fit of translated EU law: a corpus-based study of deontic modality. *The Textual Fit of Translated EU Law: A Corpus-Based Study of Deontic Modality*, *20*(3), 332–355. https://doi.org/10.1080/13556509.2014.909675

Biel, Ł. (2021). Postedycja tłumaczeń maszynowych. *Lingua Legis*, *29*(1), 11–34.

Bowker, L., & Buitrago Ciro, J. (2019). *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Emerald Publishing. https://doi.org/doi:10.1108/978-1-78756-721-420191006

Cao, D. (1962-). (2007). *Translating law*. Multilingual Matters.

Daems, J., Vandepitte, S., Hartsuiker, R. J., & Macken, L. (2017). Identifying the machine translation error types with the greatest impact on post-editing effort. *Frontiers in Psychology*, *8*(AUG), 1–15. https://doi.org/10.3389/fpsyg.2017.01282

do Carmo, F., & Moorkens, J. (2020). Differentiating Editing, Post-Editing and Revision. In M. Koponen, B. Mossop, I. S. Robert, & G. Scocchera (Eds.), *Translation Revision and Post-Editing: Industry Practices and Cognitive Processes* (pp. 35–49). Routledge.

Doczekalska, A. (2013). Comparative law and legal translation in the search for functional equivalents - Separate or intertwined domains. *Comparative Legilinguistics*, *16*, 63–75.

Engberg, J. (2013). Comparative law for translation: The key to successful mediation between legal systems. In A. Borja Albi & F. Prieto Ramos (Eds.), *Legal Translation in Context. Professional Issues and Prospects* (pp. 9–25). Peter Lang.

Gościński, J. (2019). *Egzamin na tłumacza przysięgłego. Angielskie orzeczenia w sprawach karnych*. Wydawnictwo C.H.Beck.

House, J. (2013). English as a lingua franca and translation. *The Interpreter and Translator Trainer*, *7*(2), 279–298.

Jopek-Bosiacka, A. (2013). Comparative law and equivalence assessment of system-bound terms in EU legal translation. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, 110–146.

Kecskes, I. (2008). Formulaic language in English Lingua Franca. In I. Kecskes & L. R. Horn (Eds.), *Explorations in Pragmatics: Linguistic, Cognitive and Intercultural Aspects* (pp. 191–218). De Gruyter Mouton.

Kecskes, I., & Kirner-Ludwig, M. (2019). "Odd structures" in English as a lingua franca discourse. *Journal of Pragmatics*, *151*, 76–90.

Killman, J., & Rodríguez-Castro, M. (2022). Post-editing vs. translating in the legal context: Quality and time effects from English to Spanish. *Revista de Llengua i Dret, Journal of Language and Law*, *78*, 56–72. https://doi.org/https://www. doi.org/10.2436/rld.i78.2022.3831

Koponen, M. (2016). *Machine Translation Post-editing and Effort Empirical Studies on the Post-editing Process*. University of Helsinki.

Kurek, H. (2015). Język prawny i prawniczy na przełomie wieków (perspektywy badawcze i zagrożenia). *Zagadnienia Naukoznawstwa*, *3 (205)*, 303–310.

Lesznyák, M., & Balogh, D. (2019). Comparative Analysis of Translations Prepared by Students with and Without Legal Qualifications. *Comparative Legilinguistics*, *37*, 85–115. https://doi.org/DOI: http://dx.doi.org/10.14746/cl.2019.37.3

Lesznyák, M., & Balogh, D. (2022). *Cooperation between legal professionals and translation MA students within translator training*.

Mossop, B. (2020). *Revising and Editing for Translators: Vol. Fourth ed.* Routledge.

Nitzke, J., & Hansen-Schirra, S. (2021). *A short guide to post-editing*. Language Science Press. http://langsci-press.org/catalog/book/319

Nunes Vieira, L. (2016). *Cognitive effort in post-editing of machine translation: evidence from eye movements, subjective ratings, and think-aloud protocols* [Newcastle University]. http://hdl.handle.net/10443/3130

O'Brien, S. (2022). How to deal with errors in machine translation: Post-editing. In D. Kenny (Ed.), *Machine translation for everyone: Empowering users in the age of artificial intelligence* (pp. 105–120). Language Science Press.

O'Brien, S., Simard, M., & Goulet, M.-J. (2018). Machine Translation and Self-post-editing for Academic Writing Support. In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), *Translation Quality Assessment. From Principles to Practice* (pp. 237–262). Springer. https://doi.org/https://doi.org/10.1007/978-3-319-91241-7

Orlando, D. (2015). *The trials of legal translation competence: Triangulating processes and products of translators vs. lawyers* [UNIVERSITÀ DEGLI STUDI DI TRIESTE]. https://arts.units.it/retrieve/handle/11368/2908045/187217/ORLANDO_PhD_Thesis.pdf

Pontrandolfo, G. (2015). Investigating Judicial Phraseology with COSPE: A contrastive Corpus-based Study. In *New Directions in Corpus-Based Translation Studies* (pp. 137–159).

Pontrandolfo, G. (2018). Sensibly and appropriately the judge considered… A corpus-based study of sentence adverbs in judicial language. *Linguistik Online*, *92*(5), 215–233. https://doi.org/10.13092/lo.92.4511

Prieto Ramos, F. (2014a). International and supranational law in translation: from multilingual lawmaking to

adjudication. *The Translator*, *20*(3), 313–331. https://doi.org/10.1080/13556509.2014.904080

Prieto Ramos, F. (2014b). Legal Translation Studies as Interdiscipline: Scope and Evolution. *Meta*, *59*(2), 260–277. https://doi.org/10.7202/1027475ar

Šarčević, S. (1997). *New Approach to Legal Translation*. Kluwer Law International.

Scott, J. (2017). Legal translation – A multidimensional endeavour. *Comparative Legilinguistics*, *32*, 37–66.

Setkowicz-Ryszka, A. (2022). *Academic legal writing in Polish-English translation. Corpus analysis of lexical items that cause difficulties titles and abstracts of legal papers*.

Sosoni, V., O'Shea, J., & Stasimioti, M. (2022). Translating law: A comparison of human and post-edited translations from Greek to English. *Revista de Llengua i Dret, Journal of Language and Law*, *78*, 92–120. https://doi.org/https://doi. org/10.2436/rld.i78.2022.3704

Stefaniak, K. (2020). Evaluating the usefulness of neural machine translation for the Polish translators in the European Commission. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 263–269. https://aclanthology.org/2020.eamt-1.28

TAUS. (2017). *Quality Evaluation Using an Error Typology Approach* (p. 7). TAUS Signature Editions.

Voita, E., Sennrich, R., & Titov, I. (2019). When a Good Translation is Wrong in Context : Context-Aware Machine Translation Improves on Deixis , Ellipsis , and Lexical Cohesion. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1198–1212. https://doi.org/10.18653/v1/P19-1116

Why are generic MT engines of limited assistance to legal academics wishing to communicate in English as a *Lingua Franca*? A reviser's and post-editor's perspective

Anna Setkowicz-Ryszka *University of Lodz*

Email: anna.setkowicz.ryszka@edu.uni.lodz.pl *ORCiD 0000-0003-0057-3477*

1. Introduction

This paper is based on the author's experience as a reviser of academic legal texts, including around a decade of post-editing machine translation (PEMT) of such texts, with gradual transition from frustration with MT output to acceptance of DeepL as a productivity-boosting tool. However, this acceptance only came about after several years of willy-nilly fixing machine-translated texts submitted by legal academics and after the recent noticeable improvements in the quality of MT output.

The differences between texts translated by the authors themselves (or by incompetent translators) and machine-translated texts are not clear-cut, as authors can use MT engines as dictionaries or try post-editing MT output. Since the assignments come from a publishing house and not directly from the authors, it is sometimes only possible to compare the submitted texts with the output from generic MT engines (mainly Google Translate and free DeepL Translator). When even MT output appears less problematic than the text at hand a conclusion has to be drawn about it being a poor human translation, but there is never full certainty.

This paper focuses on unsatisfactory translations by humans or MT algorithms, leaving aside many translations submitted for revision that are reasonably or very good. Moreover, it should be stressed that many authors (especially those specializing in commercial law) submit only Polish texts, which means they ask the publishing house to commission translations. It is the bad or appalling ones, which need a lot of effort to fix, that inspire reflections about the challenges of translating academic legal writing. The upside of the usually problematic lack of direct contact with authors is that they might behave more naturally, inadvertently revealing that they perceive legal translation to be a relatively simple activity.

The following sections will look at the language used by legal academics in abstracts and titles of their articles (section 2), the typical problems of English used as a Lingua Franca (section 3) and examples of problems from academic legal texts (section 4). Next, a discussion of post-editing of machine translation in general (section 5) will be followed by a presentation of three major groups of errors that often appear in academic legal writing (section 6), while the concluding section (section 7) will sum up use the problems of English as the *lingua franca* of legal academia and offer recommendations for source texts that might improve the quality of both ELF translation and MT output.

2. Features of legal language

Academic legal texts constitute one of three major groups of legal texts, in addition to normative and judicial texts (Prieto Ramos, 2014b: 262). According to Prieto Ramos,

> Legal texts constitute or apply instruments governing public or private legal relations (including codified law, case-law and contracts), or give formal expression to specialized knowledge on legal aspects of such instruments and relations (2014b: 264).

Clearly, academic writing is mainly concerned with the third function, i.e., expressing expert legal knowledge. Polish academic legal texts include essays, articles about legislation (including suggestions of desirable amendments) or case comments (Polish: *glosa*), which do not merely report facts of the cases and court determinations, but evaluate such determinations in terms of correctness of reasoning, axiology or methodology, engaging in a dialogue with the judiciary (Łętowska, 2005).

Generally speaking, features of legal texts include precision, logic of argumentation, matter-of-factness, directive character, standardization, and presence of terminology. Additionally, these texts are often impersonal and concise, even excessively concise (Jopek-Bosiacka, 2010: 21). On the other hand, their precision is often limited, intentionally or not, by the use of vague wording, indeterminate terms, and even euphemisms (Alcaraz Varó & Hughes, 2002: 11-13, 30-42; Jopek-Bosiacka, 2010: 32-33).

In legal translation studies, the focus has been mainly on normative and judicial texts, academic genres being only briefly mentioned. Cao stresses that the legal status of 'scholarly legal texts', such as commentaries, depends on the particular legal system (2007: 9, 83). Prieto Ramos mentions that these texts – including journal articles, textbooks or press reports – are not always addressed to legal experts. This is why they display heterogenous stylistic features, but all of them fulfil "descriptive and argumentative functions" to some extent (Prieto Ramos, 2014b: 263). Alcaraz Varó & Hughes discuss the structure of law reports and professional articles (2002: 108-112; 146-149).

Legal terminology and asymmetry between legal concepts have been studied extensively (e.g., Gościński, 2019: 164-169; Šarčević, 1997: 237-239). Legal terms are often system-bound, while lack of equivalence between them forces translators to use various techniques to compensate for terminological incongruency (Gościński, 2019: 164-169; Šarčević, 1997: 250-264). Legal translators are often recommended to use comparative law methods (Prieto Ramos, 2014: 267-268; Engberg, 2013: 10-18; Šarčević, 1997: 114, 235).

However, terminology is only one of the challenges. As Kjaer points out, "[t]he functioning of a legal system is dependent on constant processes of stabilization and specialization of words *and phrases* that accompany the construction, deconstruction or reconstruction of legal concepts" (2007: 508, my emphasis). She stresses the tendency to "reproduce words and phrases in exactly the same form" in legal texts (Kjaer, 2007: 510). Some routine phrases are directly or indirectly prescribed by law, some are *implicit quotations* from other texts (intertextuality), but some are simply habitual (Kjaer, 2007: 512, my emphasis). All of them are found in academic texts.

Moreover, legal writing is often syntactically complex, with long sentences, passive constructions or embeddings, complex noun phrases, strings of nouns, complex prepositions, qualificational insertions or (multiple) negatives (Jopek-Bosiacka, 2010: 63-72). Kurek's list of common problems or errors that make Polish non-normative legal texts (i.e., texts from the fields of legal practice and academia) difficult to understand includes complex syntax, especially syntactic homonymy or secondary syntactic connections, both of which lead to ambiguity, poor punctuation, and excessive use of words of foreign origin (2015: 305-309). A study on comprehension and recall of legal texts (mainly contracts), found centre-embedded clauses and low-frequency vocabulary to be the two top features making them difficult (Martínez et al., 2022: 6).

For the above reasons, legal translation (LT) competence includes additional elements compared to general translation competence. Without a detailed analysis, LT competence comprises knowledge of the source and target legal systems, including various branches of law, awareness of differences between these systems and legal concepts from them, ability to address such differences, familiarity with the conventions of legal language and different genres of legal texts, as well as the ability to find and critically evaluate information sources (for a summary of LT competence models, *see* Popiołek, 2020).

3. ELF translations

Many academic legal texts are written in English as a Lingua Franca (ELF) by authors themselves or other persons whose command of English or LT competence is insufficient to produce texts of high

quality (revision of the latter texts is beyond the scope of this paper). These texts, usually abstracts, less often full articles, need to be revised, sometimes monolingually, that is, without the source text. With the number of non-native speakers of English five times higher than that of native speakers (Albl-Mikasa & Ehrensberger-Dow, 2019: 46) and "claims of the effective and successful nature of ELF in international communication" (Albl-Mikasa, 2017: 372), English might appear the perfect language of international communication. It was found that, in face-to-face communication, non-native speakers manage to repair or co-construct messages when faced with odd structures (Kecskes & Kirner-Ludwig, 2019) or to create ad hoc rules of interaction (Kecskes, 2008).

However, there are some risks and costs of using ELF in professional or academic settings. For instance, the use of ELF was found to lead to unsuccessful discussions when engineers who were not native speakers of English were limited to using the word "error" to express a number of concepts: "quality defect", "oversight", "incorrect planning", "mismanagement", and "deficiencies". Similarly, in academia, lack of familiarity with English rhetorical conventions can undermine the success of researchers, especially in the humanities or social sciences (Albl-Mikasa & Ehrensberger-Dow, 2019: 48), sometimes leading to misunderstandings (Bennett & Queiroz de Barros, 2017: 366).

ELF is alternatively referred to as International English or English as International Language, but language mediators use more emotionally-tinged names: "globish", "bad simple English", "Lego English" or "desesperanto" (Albl-Mikasa, 2017: 372). Interpreters express their concerns about poor efficiency of communication or rather miscommunications they witness when not asked to interpret: "most users are not aware of the enormous cultural gaps that exist", "speakers end up being unclear and sometimes even saying the opposite of what they mean" or "[p]eople think they speak the same language, but it is rarely the case" (Gentile & Albl-Mikasa, 2017: 58-63).

Research finds that language mediators face extra challenges with understanding ELF, especially when multiple negative features of ELF appear in combination. In terms of text organization, non-native speakers of English do not always use connectives or certainty markers in a conventional way, which can obscure the logic of their argument (Albl-Mikasa, 2017: 376), which is crucial in legal texts. In a study on translators, it was found that unedited ELF texts resulted in more mistranslations and longer processing time, which not always led to decoding the writers' intentions (Albl-Mikasa et al., 2017). Interpreters, who cannot spend more time on their task, report they need to be more focused and pay extra attention to ELF input to be able to disambiguate it or derive intended meaning from non-standard English, incomplete structures and unusual word combinations (Albl-Mikasa et al., 2020: 272), some of which may be word-for-word translations of routine phrases used in legal texts.

In my experience ELF texts can also be rather taxing to revise, even though in the case of texts discussed here the cognitive load is often lessened by the so-called shared language benefit (Albl-Mikasa & Ehrensberger-Dow, 2019: 53). Both for authors and for myself English is not the native language, so I accept that a native-speaker reviser might find problems, hopefully of lesser severity, in my translations. Still, in the case of Polish, L2 or inverse translation into English is the norm. Inverse translation as such is neither unusual (Pokorn, 2005), not necessarily inferior in terms of quality (Whyatt, 2019), but since Polish law students are not taught translation, it is not surprising that their work might exhibit problems caused by lesser familiarity with other legal systems or genre conventions in other legal systems.

4. What a reviser corrects most often in ELF translations?

Problems in ELF translations often appear as early as in the titles, many of which contain the word *glosa* translated as *glossary* or *gloss*, rather than *commentary* or *case comment.* The latter term closely corresponds to the Polish one, but is difficult to use as the standard title format is *glosa do wyroku*

[*commentary on judgment*] followed by the court name, date of issue, and case number (Setkowicz-Ryszka, 2022). The next item is the very word *abstrakt*. Even though the first impulse might be to render it as *abstract*, the word *summary* appears surprisingly often, perhaps because of the synonymous Polish word - *streszczenie*. The abstract sometimes includes content that is written so badly that a reviser has to guess what the author intended to say or refer to the Polish version. For instance:

It is the artists who owe the informational function about the past times.

It is worth for a law adept to become acquaintanced with the mechanisms of a courtroom and be able to use them.

Would seem to be able that pointing out elements consists on essence of mystery shoudn't cause more troubles because each person in own life gave the other person own secret or was a confidant of mystery. [*original spelling*]

In less problematic texts, grammatical or orthographical errors rarely affect comprehension, though unintended meaning shifts may appear, like in examples 1 and 3:

	Source text (ST)	**ELF**	**Revised version (RV)**
1	zmniejszanie świadczeń emerytalno-rentowych	reduction retirement and disability benefits	reduction **of** retirement and disability pension
2	Glosa ma charakter aprobujący.	The commentary is approving.	This commentary is an approving one.
3	adresat prawa powinien móc także rozumieć siebie jako twórcę prawa	the addressee of law should also be able to understand himself as **the** author of law	the addressee of law should also be able to understand themselves as **an** author of law
4	ochrona życia i bezpieczeństwa	protection of **heart** and safety	protection of **health** and safety

The most frequent category of problems in ELF texts is word-for-word translations from Polish. This applies to phrases and syntactic structures, which can sound awkward.

	ST	**ELF**	**RV**
5	uchwała **pełnego składu** Naczelnego Sądu Administracyjnego	resolution of the **Full Composition** of the Supreme Administrative Court	resolution of the Supreme Administrative Court **sitting as full court**
6	-	tax group **would have been extended by** another company	tax grouping **would have gained** another member
7	…jest **objęty** zakresem stosowania…	…is **grasped with** scope of…	**…falls within** the scope of application of…
8	…interwencji legislacyjnej ustawodawcy	…legislative intervention by the legislator	…legislative intervention

Many problems seem to result from the choice of a wrong equivalent of a polysemous word or term from a bilingual dictionary.

	ST	**ELF**	**RV**
9	zasadę **wolności wyborów** i zasadę uczciwych wyborów	the principle of **freedom of choice** and the principle of fair elections	the principle of **free** and fair **elections**
10	obowiązek **poufności**	**discretion** obligation	obligation of **confidentiality**
11	**kontrola** sądowa	judicial **supervision**	judicial **review** (UK sense)
12	**kwalifikowanej zdrady** małżeństwa	**qualified treason** in marriage	**aggravated infidelity** in marriage

Finally, there can be translations that are problematic from the point of view of semantic prosody or register.

	ST	ELF	RV
13	-	the Polish King […] founded (in 1364) the Cracow University, as a **breeding ground** for qualified legal staff	the Polish King […] founded (in 1364) the Cracow University as an **institution preparing** qualified legal staff
14	[teza x] musi być oceniona krytycznie	[statement x] seems **not to be quite right** in regard	[statement x] **deserves criticism**
15	kary za nieterminową dostawę towarów	penalties for **untimely** delivery of goods	penalties for **late** delivery of goods
16	-	identity in the sense of 'being yourself' (Latin *ipse*, **English *selfhood***, German *Selbstheit*)	identity as selfhood (Latin *ipse*, German *Selbstheit*, French *ipseite*)

5. Machine-translated texts

It seems that some legal academics in Poland embraced MT engines around 2012/2013. It was then that I started receiving machine-translated abstracts for "revision". I still do. The authors never mention they use MT engines. Initially I found dealing with these texts very frustrating, they often seemed to completely miss the point. It was not until 2015 that I learned that what I was doing was post-editing (PE) and that those texts were not human translations. Now, I consider MT output more predictable in terms of cognitive effort required for PE than for revising some ELF texts, even if still far from revising a good human translation. This is mainly due to the impressive progress of MT engines over the years, especially in terms of dealing with the complex syntax and stilted language often used by lawyers.

MT is generally known to perform better if the source text is a routine one, written in a controlled language (Biel, 2021: 23), so the formulaic language and the presence of routine phrases should be an advantage. eTranslation system used by EU translation services is able to handle legislative, judicial and administrative texts, because the engine has been trained on such texts. However, despite certain repeated lexical items, academic legal texts often present new content: suggestions of legislative amendments or attempts to dispel doubts or resolve controversies. Some features of the legal language described in section 2 belong to negative translatability indicators (O'Brien, 2006: E-1). Moreover, legal texts are highly intertextual, with references to legislation and instruments of international law, judgments of national courts, but also foreign or European ones, and earlier scholarship. Since legislation and judgments are not protected by copyright, the quotations are rarely marked and need to be researched.

Post-editing is usually found to be faster than translation from scratch and involve less keyboard and mouse activity, but "sometimes reported as being more demanding a task than translation without MT as an aid", so the productivity gain comes with at the expense of high cognitive effort (O'Brien, 2022: 115-116). Do Carmo & Moorkens (2020) claim that post-editing resembles translation more than revision, despite the obvious similarities (two texts, more reading than typing). In terms of skills, there seems to be a consensus that a post-editor needs translation skills (though not all translators will be good post-editors) plus some extra skills (O'Brien, 2022: 118), more specifically: error handling, MT engineering, consulting and PE soft skills (Nitzke & Hansen-Schirra, 2021: 71-73), and that special training is needed (O'Brien, 2002). Obviously, considering the severity of errors discussed in the previous section, legal academics writing in ELF at this level are unable to perform good PEMT.

Nitzke and Hansen-Schirra warn that neural MT generates less obvious errors, which are easy to overlook as the output reads fluently and seems correct. Therefore, post-editors need training to be able to "spot exactly these more fine-grained mistakes" and recognize the potential strong priming effect (2021: 71). It is important to bear in mind various levels of error severity, though academic writing must undergo full PE (Massardo et al., 2016), but also various levels of effort in terms of cognitive processing needed to detect them and rewriting needed to correct them (Koponen, 2016: 26). Fluency errors, such as coherence and meaning shifts, were found to cause greater cognitive effort, measured in duration and number of fixations, than acceptability errors (Daems et al., 2017: 7, 12). A study on English-Polish translation in the EU context confirmed that accuracy errors were difficult to spot in the fluent NMT output. It also found that maintaining terminological consistency and textual coherence was challenging. In addition, frequent errors appeared in titles, quotes and pronouns (Stefaniak, 2020). Moreover, even error-free (i.e., acceptable or plausible) translations may not always be adequate in context, e.g., that of the preceding sentence (Voita et al., 2019: 1199).

From the full set of error categories in the Multidimensional Quality Metrics, the most frequent categories that I deal with in academic legal texts nowadays include Style, Audience appropriateness, Terminology (inconsistency), and Accuracy. From Mossop's top five items that need fixing during PE these would be term inconsistencies, proper names and stylistic problems, rather than grammar errors, extra/missing words (2020: 218). Finally, from Biel's 13-item list of frequent MT error types, I might identify mistranslations caused by polysemy or limited context, lack of terminological consistency, lack of textual cohesion, inability to deal with slang (legal jargon in this case), and lack of understanding of context and cultural references (2021: 25, my translation).

6. What a post-editor corrects most often in machine-translated academic legal texts?

There are still many problems in machine-translated legal texts. I now rarely find the most severe errors I remember from the past, such as true "word salad" translations, serious omissions or completely misguided translations of polysemous words. MT output is also less word-for-word than many ELF translations. Still, post-editors need to be aware of and alert to certain peculiar errors if PEMT is to lead to greater efficiency or higher quality. As a legal translator trainer, I have been sharing my experience of post-editing with trainees, which has forced me to categorize MT errors.

The frequent errors that are currently typical of academic texts can be grouped in three broad categories of: (1) too liberal renditions; (2) too literal renditions, and (3) problems with inter- and intra-textual references, which are divided into subcategories and exemplified below (all examples come from DeepL pro plug-in in memoQ).

1. Too liberal renditions include:

(1a) the use of multiple alternative terms ("thesaurus effect") within a short text when consistent use of terminology is desirable; there are often multiple English alternatives of Polish terms, because various strategies can be used in legal translation and several target legal orders may be considered (English law, US law, international law, EU law):

ST	MT	PEMT
środek płatniczy	legal tender, means of payment	[*either one, consistently*]
wskazanie […] osoby […] do rady nadzorczej	**designation, indication** of a person to the supervisory board	**designation (nomination)** of a person as a candidate for a supervisory board member
lokal mieszkalny	residential unit, flat, dwelling, apartment	[*any equivalent, consistently*]

(1b) solutions that make the translation fluent, but at the expense of accuracy:

ST	MT	PEMT
…w dwustopniowej procedurze, której pierwszym etapem jest […] **Kolejnym**, z którego może skorzystać wspólnik, jest…	…in a two-stage procedure, the first stage of which is […] **Another avenue** which may be used by a shareholder is …	…in a two-stage procedure, the first stage of which is […] **The second one** is for a shareholder to use…
W glosowanym wyroku przyjęto **godne aprobaty zapatrywania** dotyczące…	The judgment under review adopts **a favourable opinion** on…	The commented judgment is based on a **commendable assumption** about…
Przedmiotowa kwestia problemowa ma istotne przełożenie…	The issue **at stake** has important implications …	The issue **in point** has important implications…

(1c) hallucinations, some of which seem likely in the context, others rather surprising:

ST	MT	PEMT
Skarb Państwa Chińskiej Republiki Ludowej	**People's Treasury** of the People's Republic of China	**State Treasury** of the PRCh
W artykule przedstawiono **węzłowe** problemy prawne…	The article outlines the **knotty** legal issues…	The article outlines the **crucial** legal issues…
art. 12 ust. 2 w zw. **z art.** 2 specustawy covidowej	Art. 12 sec. 2 in connection with **joke.** 2 of the special covid act	Art. 12(2) read in conjunction **with Art.** 2 of the Special COVID Act

2. The category of too literal renditions comprises:

(2a) literally translated typos that a human reader may not even detect:

ST	MT	PEMT
Dopiero **traki** [*correct word: taki*] oryginalny dokument…	Only the **traki** original document…	Only **such** an original document…
negatywnymi **stukami** [*correct word: skutkami*]	negative **knocks**	negative **consequences**
wykładni **przypisów** jej dotyczących [*correct word: przepisów*]	interpretation of **footnotes** concerning it	interpretation of **provisions** concerning it

(2b) contextually wrong translations of polysemous words:

ST	MT	PEMT
[wyrok] podejmuje problematykę dwóch zagadnień […] Oba **wątki**…	[the judgment] tackles two issues […] Both **strands**…	[the judgment] tackles two issues […] Both **themes**…
możliwość zbycia wierzytelności	**marketabilit**y of this claim	**transferability** of this claim
…z uwagi na zasadę jawności ksiąg wieczystych **określaną** […] **jako** zasada powszechności…	…due to the principle of openness of mortgage registers, **defined as** universality …	…because of the principle of open access to land and mortgage registers - **referred to as** the principle of common access

(2c) retained Latin expressions not used in the target legal language (Setkowicz-Ryszka, 2022):

ST	MT	PEMT
postulaty **de lege ferenda**	postulates **de lege ferenda**	suggestions of legislative amendments [*English explanation*]

Bezpośrednie określenie […] zostało **expressis verbis** wyrażone w art. 5	The direct definition […] is **expressis verbis** expressed in Article 5	…is [*omission*] expressed in Article 5
Ratio legis przepisu art. 552§1	**Ratio legis** of the provision of Art. 552§1	**The reason for the adoption** of Art. 552(1)

3. The third category, problems relating to references, groups together:
(3a) wrong intertextual references, namely quotes from case law and legislation, names, titles of legislative instruments, including their abbreviations:

ST	MT	PEMT
Krajowego Planu Odbudowy	National Reconstruction Plan	National Recovery and Resilience Plan
art. 120 § 2 **u.p.e.a.**	Article 120 § 2 **of the A.P.A.**	Article 120(2) **of the Act on Enforcement Proceedings in Administration**
…właścicielem sprzedawanego towaru jest **osoba fizyczna**…	…the owner of the goods sold is the **natural person**…	…the owner of the goods sold is a **private individual**… [*quote from CJEU judgment*]

(3b) wrong intra-textual references:

ST	MT	PEMT
Trybunał [Europejski Trybunał Praw Człowieka]uznał, że wolność…	The **Tribunal** held that the freedom…	The **Court** [European Court of Human Rights] held that the freedom…
Naczelny Sąd Administracyjny zwraca uwagę na… Wykładnia poczyniona przez **Sąd** nakazuje…	The Supreme Administrative Court draws attention to... The interpretation made by the **Court of First Instance** requires...	The Supreme Administrative Court draws attention to... The interpretation made by this **Court** requires...
Jedną z **przesłanek** warunkujących […] **Przesłanka** ta…	One of the **prerequisites** for […] This **premise** …	One of the **conditions** for […] This **condition**…

(3c) wrong deictic references, which are particularly troublesome when a text is translated between a language that has grammatical gender, such as Polish, and one which does not, such as English:

ST	MT	PEMT
Ocena **ich** dopuszczalności może być dokonywana jedynie… [*reference to "dowody", a plural noun*]	**Their** admissibility can only be assessed…	**Its** admissibility can only be assessed [*reference to "evidence", a singular noun*]
Stanowisko wyrażone w glosowanym orzeczeniu dotyczy odpowiedzialności komplementariusza […] Kształtują **ją** przepisy… [*feminine pronoun refering to "odpowiedzialność"*]	The position expressed in the judgment under review concerns the liability of a general partner […] **It** is shaped by the provisions… [*confusing reference: position, judgment or liability*]	The position expressed in the commented judgment concerns the liability of a general partner […] **Such liability** is shaped by the provisions…
Niemcy wyraźnie **uznają** […] **Uważają** również, że… [*Niemcy means both "Germany" and "Germans"*]	Germany explicitly **recognises** […] **They** also **consider**…	Germany explicitly **recognizes** […] **It** also **considers**…

7. Conclusion. Recommendations

Summarizing, both ELF revision and PEMT of academic legal texts pose certain challenges. In ELF, many errors result from excessive literality, word-for-word translations of complex syntactic structures, and system-bound legal terminology. In MT output, many more errors belong to the "too liberal" group. They are often caused by segmentation, which may remove the context needed to disambiguate a polysemous word or to maintain textual cohesion. Intertextuality is a major source of difficulties, especially when quotations are implicit. Both ELF writers and MT tend to translate anew titles and quotes which should be copied from source documents (which have to be found first).

It is possible that some ELF texts are based on MT output, hence the common categories of errors. However, ELF errors discussed in section 4 demonstrate that some legal academics underestimate the complexity of legal translation, while overestimating their own command of English. It can be presumed that the fact that they fail to fix problems in MT output (if they use it) means they do not see them, so they would not have prepared a better translation from scratch. However, they should realize that titles and abstracts in English are often the only traces of their work on Web of Science or Scopus. Generally speaking, legal academics need a more realistic assessment of their command of legal and academic English and greater awareness of the pitfalls of MT. As I learn from translation trainees, they usually discover MT output contains many more errors than they first realized and they are glad to have been warned, considering the inevitability of PEMT in the profession. Lawyers need greater MT literacy, too (Bowker & Buitrago Ciro, 2019).

As for the feasibility of MT in legal scholarship, the quality of the output depends both on factors such as specific system, language pair and training data, as well as on the quality of STs: whether they contain errors and they are written in controlled language. The current neural MT systems benefit less from controlled language, so it is ST quality that matters more (Nitzke & Hansen-Schirra, 2021: 62-64). Some of the problems described above might be avoidable if authors removed errors and followed recommendations for texts to be (machine) translated, including writing with greater clarity and using plain words, avoiding syntactically complex or ambiguous sentences, abbreviations, and even too many pronouns (Translation Centre for Bodies of the EU, 2019, 2021). Translation-friendlier STs could also make their own translations better. Dealing with ST defects, complex syntax and specialist vocabulary are tricks of the trade that legal translators learn over time. Meanwhile bilingual lawyers were found to make more errors in translation of legal texts (into their L1) than advanced translation students (Orlando, 2015).

Finally, even though English is nowadays the *lingua franca* of academia, its usefulness in the field of law is reduced by the differences between legal systems, especially between civil and common law systems. Yet, when English is used as a *lingua franca* of law, one should not automatically assume that the legal system of the target readers is one of common law systems. English is often used for communication between lawyers from various civil law countries, so the potential for information loss or distortion seems greater than if German or French were the *lingua franca*. Many gaps need bridging in legal translation. System-bound terms may require descriptive translations, context-dependent translations or even explanatory notes. Also, the academic legal culture of common law seems more focused on the concrete, while in Polish law studies there seems to be more theorizing and abstract categories.

It would be helpful if lawyers considered the needs of foreign readers, who are unfamiliar with the details of national law that the ST concerns. Writing in a more explicit way, without assuming certain shared knowledge on the part of recipients, while avoiding the local legal jargon/particular expressions

would help both foreign readers and language mediators working on those texts. This means that in legal scholarship abstracts for local and international readers may need to be different. Some authors realize that, but they are a minority. Finally, legal scholars should realize that titles and abstracts in English are often the only traces of their work on databases such as Web of Science or Scopus.

All the above suggests that MT can only help in legal translation when used by a person with the necessary skills (PE competence) and even so only to a limited extent. To conclude, let me repeat the words of a language mediator expressing concerns about ELF:

> People think they speak the same language, but it is rarely the case. Interpretation will be important after miscommunication incidents in English. Right now, people think that if everyone speaks English everything will be fine (Gentile & Albl-Mikasa, 2017: 59).

REFERENCES:

Albl-Mikasa, M. (2017). ELF in translation/interpreting. In J. Jenkins, W. Baker, & M. Dewey (Eds.), *The Routledge Handbook of English as a Lingua Franca* (pp. 369–383). Routledge.

Albl-Mikasa, M., & Ehrensberger-Dow, M. (2019). ITELF: (E)merging Interests in Interpreting and Translation Studies. In E. Dal Fovo & P. Gentile (Eds.), *Translation and Interpreting. Convergence, Contrast and Interaction* (pp. 45–62). Peter Lang.

Albl-Mikasa, M., Ehrensberger-Dow, M., Hunziker Heeb, A., Lehr, C., Boos, M., Kobi, M., Jäncke, L., & Elmer, S. (2020). Cognitive load in relation to non-standard language input: Insights from interpreting, translationand neuropsychology. *Translation, Cognition & Behavior*, *3*(2), 263–286. https://doi.org/https://doi.org/10.1075/tcb.00044.alb

Albl-Mikasa, M., Fontana, G., Fuchs, L. M., Stüdeli, L. M., & Zaugg, A. (2017). Professional translations of non-native English: 'before and after' texts from the European Parliament's Editing Unit. *The Translator*, *23*(4), 371–387. https://doi.org/10.1080/13556509.2017.1385940

Alcaraz Varó, E., & Hughes, B. (2002). *Legal Translation Explained*. St.Jerome Publishing.

Bennett, K., & Queiroz de Barros, R. (2017). International English its current status and implications for translation. *The Translator*, *23*(4), 363–370.

Biel, Ł. (2021). Postedycja tłumaczeń maszynowych. *Lingua Legis*, *29*(1), 11–34.

Bowker, L., & Buitrago Ciro, J. (2019). *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Emerald Publishing. https://doi.org/doi:10.1108/978-1-78756-721-420191006

Cao, D. (1962-). (2007). *Translating law*. Multilingual Matters.

Daems, J., Vandepitte, S., Hartsuiker, R. J., & Macken, L. (2017). Identifying the machine translation error types with the greatest impact on post-editing effort. *Frontiers in Psychology*, *8*(AUG), 1–15. https://doi.org/10.3389/fpsyg.2017.01282

do Carmo, F., & Moorkens, J. (2020). Differentiating Editing, Post-Editing and Revision. In M. Koponen, B. Mossop, I. S. Robert, & G. Scocchera (Eds.), *Translation Revision and Post-Editing: Industry Practices and Cognitive Processes* (pp. 35–49). Routledge.

Engberg, J. (2013). Comparative law for translation: The key to successful mediation between legal systems. In A. Borja Albi & F. Prieto Ramos (Eds.), *Legal Translation in Context. Professional Issues and Prospects* (pp. 9–25). Peter Lang.

Gentile, P., & Albl-Mikasa, M. (2017). "Everybody Speaks English Nowadays". Conference Interpreters' Perception of the Impact of English as a Lingua Franca on a Changing Profession. *Cultus, the Journal of*

Intercultural Mediation and Communication, *10*, 53–66.

Gościński, J. (2019). *Egzamin na tłumacza przysięgłego. Angielskie orzeczenia w sprawach karnych*. Wydawnictwo C.H.Beck.

Jopek-Bosiacka, A. (2010). *Przekład prawny i sądowy*. Wydawnictwo Naukowe PWN.

Kecskes, I. (2008). Formulaic language in English Lingua Franca. In I. Kecskes & L. R. Horn (Eds.), *Explorations in Pragmatics: Linguistic, Cognitive and Intercultural Aspects* (pp. 191–218). De Gruyter Mouton.

Kecskes, I., & Kirner-Ludwig, M. (2019). "Odd structures" in English as a lingua franca discourse. *Journal of Pragmatics*, *151*, 76–90.

Kjær, A. L. (2007). Phrasemes in legal texts. In H. Burger, D. Dobrovol'skij, P. Kühn, & N. R. Norrick (Eds.), *Ein internationales Handbuch der zeitgenössischen Forschung / An International Handbook of Contemporary Research* (pp. 506–516). Walter de Gruyter.

Koponen, M. (2016). *Machine Translation Post-editing and Effort Empirical Studies on the Post-editing Process*. University of Helsinki.

Kurek, H. (2015). Język prawny i prawniczy na przełomie wieków (perspektywy badawcze i zagrożenia). *Zagadnienia Naukoznawstwa*, *3 (205)*, 303–310.

Łętowska, E. (2005). Dlaczego i po co pisze się glosy, czyli dwanaście uwag dla debiutujących autorów komentarzy do orzeczeń sądowych - forum. *Monitor Prawniczy*, *11*.

Martínez, E., Mollica, F., & Gibson, E. (2022). Poor writing , not specialized concepts , drives processing difficulty in legal language. *Cognition*, *224*(September 2021), 105070. https://doi.org/10.1016/j.cognition.2022.105070

Massardo, I., van der Meer, J., O'Brien, S., Hollowood, F., Aranberri, N., & Drescher, K. (2016). *MT Post-Editing Guidelines* (p. 41). TAUS Signature Editions. https://github.com/ygraham/direct-assessment/blob/master/post-editing-guidelines

Mossop, B. (2020). *Revising and Editing for Translators: Vol. Fourth ed.* Routledge.

Nitzke, J., & Hansen-Schirra, S. (2021). *A short guide to post-editing*. Language Science Press. http://langsci-press.org/catalog/book/319

O'Brien, S. (2002). Teaching Post-editing : A Proposal for Course Content. In *Proceedings of the 6th EAMT Workshop, Teaching Machine Translation* (pp. 99–106). %5Cnwww.mt-archive.info/EAMT-2002-OBrien.pdf

O'Brien, S. (2006). *Machine-translatability and post-editing effort: an empirical study using Translog and Choice Network Analysis* [Dublin City University]. https://doras.dcu.ie/18118/2/Sharon_O%27BrienV2.pdf

O'Brien, S. (2022). How to deal with errors in machine translation: Post-editing. In D. Kenny (Ed.), *Machine translation for everyone: Empowering users in the age of artificial intelligence* (pp. 105–120). Language Science Press.

Orlando, D. (2015). *The trials of legal translation competence: Triangulating processes and products of translators vs. lawyers* [UNIVERSITÀ DEGLI STUDI DI TRIESTE]. https://arts.units.it/retrieve/handle/11368/2908045/187217/ORLANDO_PhD_Thesis.pdf

Pokorn, N. K. (2005). *Challenging the Traditional Axioms: Translation Into a Non-mother Tongue*. John Benjamins Publishing Co.

Popiołek, M. (2020). ISO 20771 : 2020 overview and legal translator competence requirements in the context of the European Qualifications Framework , ISO 17100 : 2015 and relevant research. *Lingua Legis*, *28*, 7–40. https://lingualegis.ils.uw.edu.pl/index.php/lingualegis/article/view/60/49

Prieto Ramos, F. (2014). Legal Translation Studies as Interdiscipline: Scope and Evolution. *Meta*, *59*(2), 260–277. https://doi.org/10.7202/1027475ar

Šarčević, S. (1997). *New Approach to Legal Translation*. Kluwer Law International.

Scott, J. (2017). Legal translation – A multidimensional endeavour. *Comparative Legilinguistics*, *32*, 37–66.

Setkowicz-Ryszka, A. (2022). *Academic legal writing in Polish-English translation. Corpus analysis of lexical items that cause difficulties titles and abstracts of legal papers*.

Stefaniak, K. (2020). Evaluating the usefulness of neural machine translation for the Polish translators in the European Commission. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 263–269. https://aclanthology.org/2020.eamt-1.28

Translation Centre for Bodies of the European Union. (2019). *Writing for translation* (p. 21). https://doi.org/0.2817/95648

Translation Centre for Bodies of the European Union. (2021). *Writing for machine translation* (p. 14). https://doi.org/10.2817/191981

Voita, E., Sennrich, R., & Titov, I. (2019). When a Good Translation is Wrong in Context : Context-Aware Machine Translation Improves on Deixis , Ellipsis , and Lexical Cohesion. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1198–1212. https://doi.org/10.18653/v1/P19-1116

Whyatt, B. (2019). In search of directionality effects in the translation process and in the end product. *Translation, Cognition & Behavior*, *2*(1), 79–100. https://doi.org/10.1075/tcb.00020.why

Evaluation of adaptive machine translation from a gender-neutral language perspective

Aida Kostikova
Ghent University
aida.kostikova@ugent.be

Todor Lazarov
Ghent University
New Bulgarian University
tdlazarov@gmail.com

Joke Daems
joke.daems@ugent.be

Abstract

In this paper, we attempt to analyse the problem of conveying gender-neutral language when working with notional and grammatical languages (English and German) from the point of view of adaptive machine translation (MT). More specifically, we assess the efficiency of adaptive MT when it comes to gender-neutral language use, the purpose of which is to "reduce gender stereotyping, promote social change and contribute to achieving gender equality". We conclude that the initial output largely reflects cases of misgendering and generic masculine – problems that are well documented in the MT field, but which still remain unresolved. Moreover, our experiment revealed that ModernMT faces systematic difficulties in adapting to gender-neutral language when working with the English-German translation direction.

Introduction and Related Work

As the adoption of gender-neutral language (GNL) becomes more widespread, it is increasingly important to consider how these trends can be reflected in natural language processing (NLP) applications, especially given the fact that the purpose of GNL is to "reduce gender stereotyping, promote social change and contribute to achieving gender equality" (Papadimoulis, 2018: 3). To date the task of reflecting such linguistic trends as GNL has been addressed within the field of uncustomised, generic machine translation (MT) (Dev et al., 2021; Prates et al., 2019). At the same time, there are other promising and efficient solutions with the capacity of being more flexible in terms of use of gender-fair language. For example, adaptive MT is a technology which is characterised by its ability to learn from its users, make suggestions and improve accuracy over time. Adaptive MT builds on the concept of human-in-the-loop learning, which is the process by which a machine learning model receives and utilizes human intervention or feedback (Finkelstein, 2020).

Moreover, while notional gender languages, such as English, are more or less consistent in GNL strategies, more morphologically rich languages present a challenge in terms of adapting a universal gender-fair approach (Stahlberg et al., 2007). Existing strategies in German, for example, include declension rules modifications, various

gender-neutral wordings, neopronouns (Hornscheidt and Sammla 2021), and in most cases represent an individual, rather than systematic linguistic choice. This fits the purpose of adaptive MT, which adjusts to personal linguistic preferences, which can also include GNL use.

In this paper, we will assess the efficiency of adaptive MT when it comes to GNL use, focusing on non-binary oriented language use (that is, language that avoids bias toward not only females, but also individuals who identify outside the gender binary) (del Rio-Gonzalez, 2021). In particular, we will be putting the ModernMT[1] engine to the test and analyse whether and to which degree it can be retrained "on-the-fly" in attempting to ensure gender-neutrality in translation. English-German was chosen as a main working language pair in order to analyse how adaptive model of the engine adjusts the output to complex GNL modifications specific for grammatical gender languages as German (Stahlberg et al., 2007).

Methodology

In order to achieve the objectives of the study, we translated a text with the help of adaptive MT, identified bias which might be reflected in the initial output, concentrating exclusively on bias leading to under-representation of certain groups (Savoldi et al., 2021) and evaluated the adaptive model of the engine by post-editing the MT output and registering the process with the help of CharacTER (translation edit rate on character level) (Wang et al. 2016) and KSR (keystroke ratio), which was registered with the help of Inputlog, a keystroke logging program. Only gender-related items were edited. ModernMT, an adaptive MT system (integrated in MateCat, an online computer assisted translation tool), was chosen as the basis for the study. Its distinctive feature is that no changes are reflected in its base engine, and all modifications are introduced with the help of an "instance-based adaptive NMT" technology, which means that a system's generic model incrementally updates with the help of the dynamic configuration of the learning algorithm's hyperparameters (Farajian et al., 2017).

Texts developed by the International Quidditch Association[2] were used as the material for the study, as their texts are written in GNL and are available in different languages. The text size was 1138 words (divided into 45 segments) and it included 29 examples of gender-ambiguous nouns and a gender-neutral pronoun *they* in its different inflected forms, and we also made sure every word occurred at least three times in the text to increase the likelihood of the system being able to adapt after two repetitions. As a first step, an initial output generated by a baseline system was evaluated against a group of linguistic criteria derived from the European Parliament's guide on GNL. Then, the output was edited using the adaptive function of the ModernMT engine, with the increased emphasis on GNL forms, not on the overall quality of translation. As existing strategies in German are very complex due to the morphologically rich grammatical gender system (Hornscheidt and Sammla 2021), and represent an

[1] https://modernmt.com

[2] https://iqasport.org

individual, rather than systematic linguistic choice, two approaches were chosen to test the performance of an engine when working with potentially challenging elements: De-E-System, which introduces a whole new system of declension rules and neopronouns: for example, in order to eliminate a masculine gender marker in the plural noun *Spieler* (pl. *players*), which is used to refer to a group of people whose gender is unknown or irrelevant, it was changed to *Spielerne (Spielerne können in ihrem eigenen Namen mit den Offiziellen sprechen – players may speak to officials on their own behalf*); and the gender star — a nonstandard typographic style, where an asterisk (*) is used to separate gendered inflections in the German language to include individuals who identify themselves outside of the gender binary, like in the word *Spieler*innen*: *Jedes Team besteht aus zwischen 7 und 21 Spieler*innen* (*Each team is made up of between 7 and 21 players*).

Results and Discussion

The first objective of our study was the manual evaluation of gender bias, which may be present in the initial output of the MT system. Two different trends were identified during the analysis: generic masculine and misgendering.

Baseline Model

29 of 29 nouns were always translated in the masculine form, and none of the sentences were translated with at least double gender names. For example, *speaking captain* was always translated as *der sprechende Kapitän, player – ein/der Spieler, coach – ein/der Trainer, the Chair – der Vorsitzende, the IQA CEO – der IQA CEO*. Nouns, articles and pronouns in the plural have also been translated on the basis of the masculine form: *Teammitarbeiter (team staff), Kapitäne (captains), Spieler (players)*, although the German language has means for avoiding using generic masculine in the plural, which, however, are limited to the binary gender system: for example, using feminine-masculine word pairs, using feminine-masculine word pairs (e.g., *Ingenieurinnen und Ingenieure – engineers*).

The reason for that could be that its baseline model is trained in the same way as generic MT systems (Farajian et al., 2017), which are prone to pre-existing bias – any asymmetries which are rooted in society at large or languages' structure and use (Silveira, 1980; Hamilton, 1991). If present in the training data, asymmetries in the semantics of language use and gender distribution are respectively inherited by the output of the MT (Caliskan et al. 2017).

Misgendering

Another problem identified in the first part of the experiment was misgendering, which describes cases where a person is addressed by a gendered term that does not match their gender identity. For example, *they* (with one instance of *them* and five instances of *their*), which was present in 16 segments, was not translated with a gender-neutral

term in any of the cases. As noted by Dev et al. (2021), language models are prone to misgendering when there is insufficient information to disambiguate the gender of an individual, and so they default to binary pronouns and binary-gendered terms, as we observe in the case of the baseline of ModernMT.

In most cases, the pronouns *they/them/their* were treated as plural pronouns in the third person even if there is a direct reference to a single person: *Die IQA soll den Beschwerdeführer darüber informieren, wann sie zusätzliche Mitteilungen erwarten können* (*The IQA should inform the complainant as to when they can expect additional communication*). Moreover, in some cases, *they* was translated as a masculine singular pronoun: *Wenn der Kapitän in das Spielfeld zurückkehrt, nimmt er die Rolle des Kapitäns wieder auf* (*If the captain returns to the pitch, they shall resume the role of the captain*). This is also in line with the observation made by Dev et al. (2021), who noted that language models can also misgender individuals even when their pronouns are provided.

These findings indicate that the text translated by the baseline system would require a considerable amount of post-editing. In the next section, we verify whether using the adaptive function of the engine reduces that post-editing effort.

4.2. Adaptive model

CharacTER, which is common in post-editing efforts studies (Bentivogli et al., 2016), was calculated for each segment, and its change during the translation process indicated the rate at which the system adapts to the edits: for example, 0 would mean that the segment did not need any post-editing, and increase in the number of such sentences by the second half of the text (starting from the segment 22) would indicate that the system started picking up the gender-neutral forms. Possible edits included the insertion, deletion, and correcting punctuation errors; shifts of word sequences were avoided where possible (Snover et al., 2006).

As TER-derived metrics heavily depend on the length of the sentence, the text was pre-processed to ensure that every segment in the text is of an average length (around 65 characters) and has a comparable number of potentially problematic items (for instance, nouns, pronouns, articles). CharacTER scores, which reflect the final results of translation process, were complemented by KSR, to measure the number of keystrokes (and, therefore, the actual editing process) that are needed to edit the MT output.

De-E-System

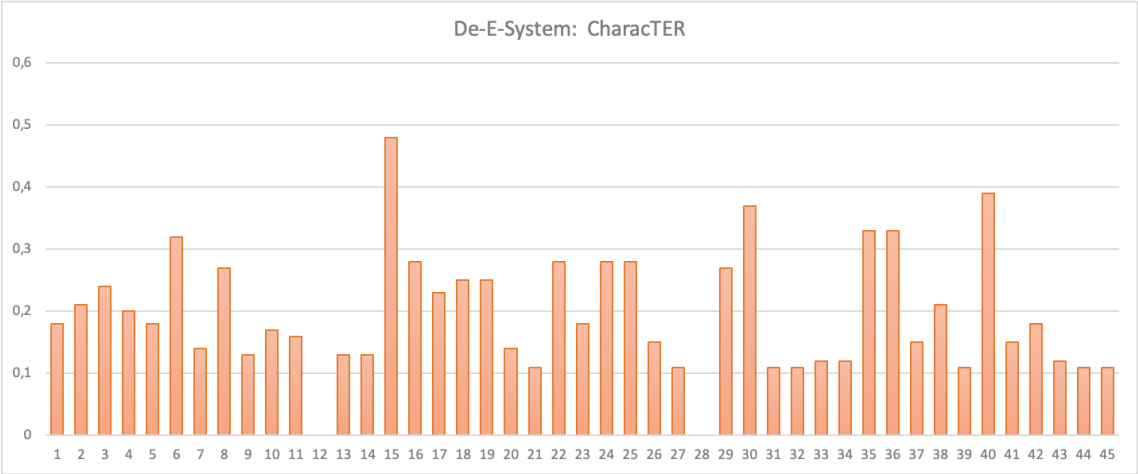


Figure 1: CharacTER scores for each segment edited according to the De-E-System

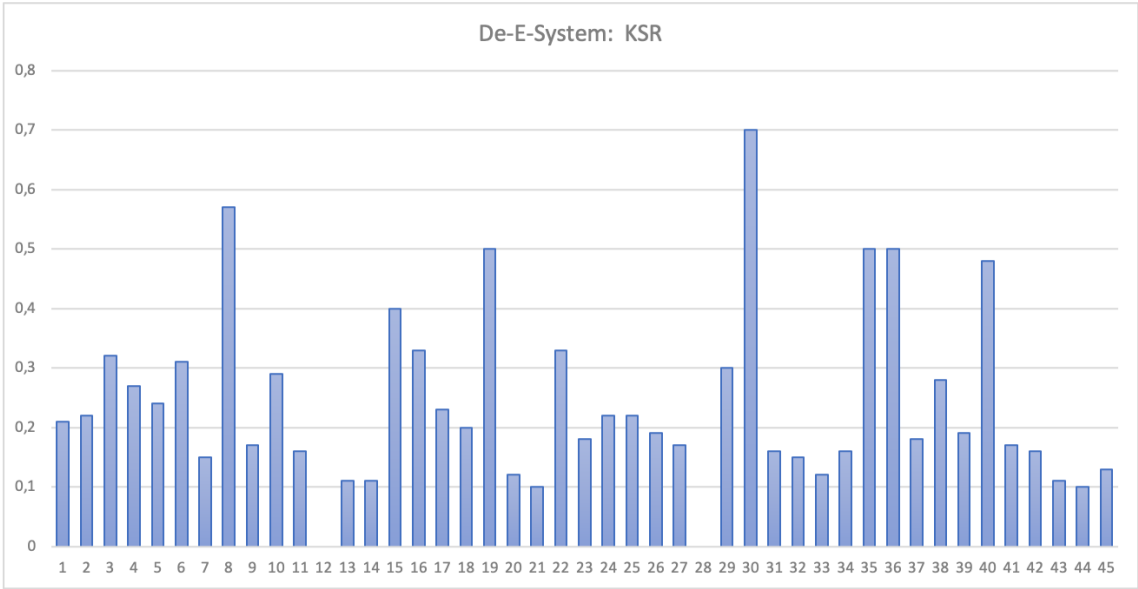


Figure 2: KSR for each segment edited according to the De-E-System

As can be seen from figures 1 and 2, only segments 12 and 28 have reached a zero value, which are in fact exact translation memory matches. The system did not show any improvements in adapting to the edits introduced by a translator; in fact, gender-ambiguous words invariably took a masculine form: for example, the word *a player* was translated as *ein* (or *der*) *Spieler* throughout the text after each segment was edited in line with the GNL strategies.

Gender Asterisk

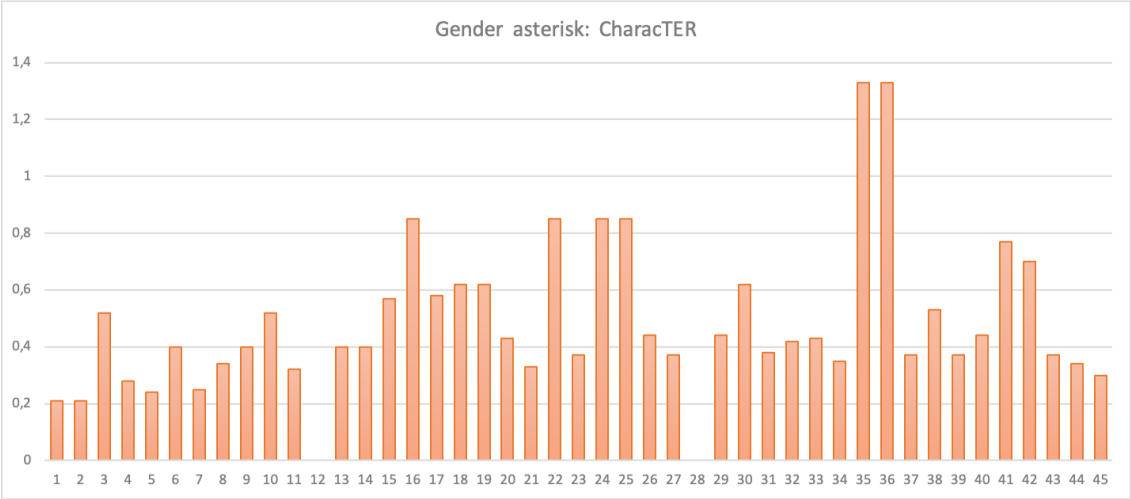


Figure 3: CharacTER scores for each segment edited according to the "gender asterisk" system

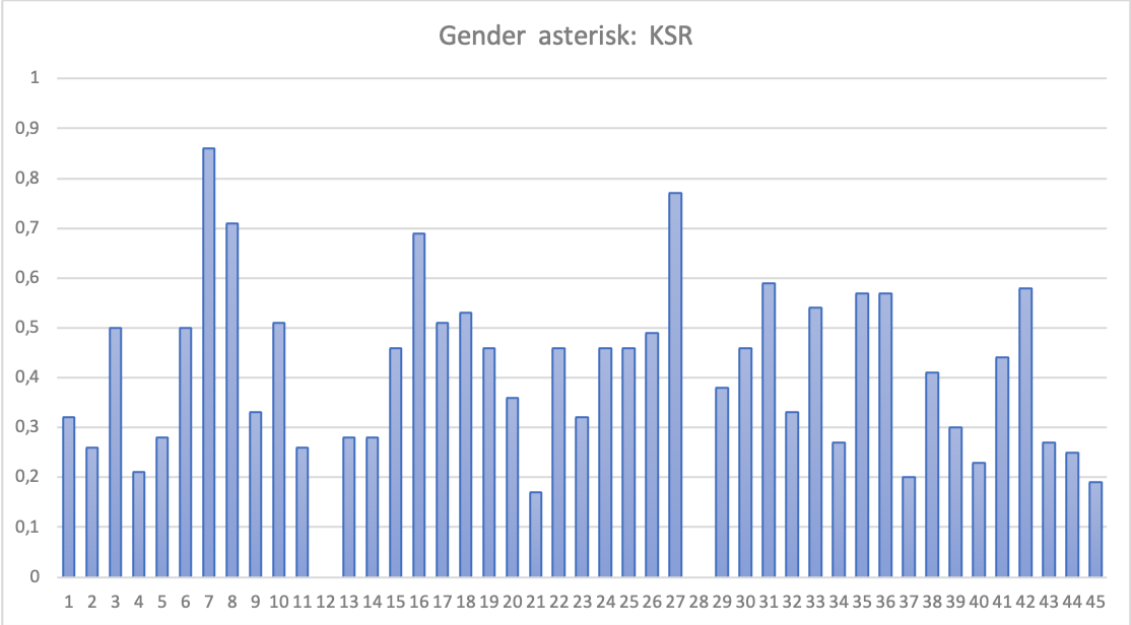


Figure 4: KSR for each segment edited according to the "gender asterisk" system

Similar results are observed with the gender asterisk: overall, this system requires more post-editing effort than the De-E-System, due to larger number of characters required to align the text with the gender-neutral strategies. Nevertheless, the system failed to adopt any changes made during the translation process, as no improvements are seen in CharacTER or KSR. It should also be noted that the active use of typographic characters did not have any effect on the rest of the text and no distortions were detected. On the other hand, for each edited segment the system reported a symbol mismatch, which

occurs when that the source and target segments do not contain the same elements and symbols. As in the case with the De-E-System, zero values are seen for segments 12 and 28, which were exact translation memory matches.

Conclusion and Future Work

In this paper, we analysed the problem of conveying GNL when working with the English-German translation direction from the point of view of adaptive MT. More specifically, we assessed the efficiency of adaptive MT by putting its baseline and adaptive functionality to the test. We conclude that the initial output largely reflects cases of misgendering and generic masculine – problems that are well documented in the MT field, but which still remain unresolved.

Some issues were also detected when working with the adaptive part of ModernMT: no progress in adaptation speed was registered when working with GNL, except for the cases of TM auto-propagation. For future work, we will additionally train the ModernMT engine by feeding it a translation memory containing GNL, and we will compare the adaptivity of another MT system, Lilt. A preliminary experiment with Lilt showed that this engine is capable of adapting to gender-neutral forms: for example, it suggested the gender-neutral noun Kapitän*in in the tenth segment.

References

Bentivogli, Luisa, Nicola Bertoldi, Mauro Cettolo, Marcello Federico, Matteo Negri, and Marco Turchi. 2016. On the evaluation of adaptive machine translation for human post-editing. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(2), pages 388-399.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. In *Science, 356(6334)*, pages 183-186.

Farajian, M. Amin, Marco Turchi, Matteo Negri, and Marcello Federico. 2017. Multi-domain neural machine translation through unsupervised adaptation. In *Proceedings of the Second Conference on Machine Translation*, pages 127-137.

Finkelstein, Paige. 2020. Human-assisted Neural Machine Translation: Harnessing Human Feedback for Machine Translation. University of Washington.

Hamilton, Mykol C. 1991. Masculine bias in the attribution of personhood: People= male, male= people. In: *Psychology of Women Quarterly*, 15(3), pages 393-402.

Hornscheidt, Lann and Sammla, Ja'n. 2021. *Wie schreibe ich divers? Wie spreche ich gendergerecht? Ein Praxis-Handbuch zu Gender und Sprache*. Insel Hiddensee: w_orten & meer.

María del Río-González, Ana. 2021. To Latinx or not to Latinx: a question of gender inclusivity versus gender neutrality. In *American Journal of Public Health, 111(6)*, pages 1018-1021.

Papadimoulis, Dimitros. 2018. *Gender-neutral language in the European Parliament*. Brussels: European Parliament.

Prates, Marcelo OR, Pedro H. Avelar, and Luís C. Lamb. 2019. Assessing gender bias in machine translation: a case study with Google Translate. In *Neural Comput & Applic* 32, pages 6363–6381.

Savoldi, Beatrice, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender Bias in Machine Translation. In *Transactions of the Association for Computational Linguistics; 9*, pages 845–874.

Silveira, Jeanette. 1980. Generic masculine words and thinking. In *Women's Studies International Quarterly*, 3(2-3), pp. 165-178.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223-231.

Stahlberg, Dagmar, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of the sexes in language. In *Social communication*, pages 163-187.

Wang, Weiyue, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. Character: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505-510.

Introducing Fairslator: a machine translation bias removal tool

Michal Měchura
Masaryk University and Dublin City University
michmech@lexiconista.com

Abstract

This paper introduces Fairslator (www.fairslator.com), an experimental tool for removing bias from machine translation. Fairslator is a plug-in for existing machine translation systems which examines their output, detects the presence of bias-causing ambiguities in gender and in forms of address, and offers the human user options for re-inflecting the translation into alternative genders and forms of address. The paper explains exactly which kinds of bias Fairslator handles, how the Fairslator method differs from other known attempts to solve the bias problem in machine translation, and discusses limitations of the method.

Introduction: bias and ambiguity in machine translation

In machine translation (MT), bias is well-known as an unintended side effect of machine learning. Language models end up being biased if and when they have been built from biased training data, for example from data where gender-neutral words in one language, like *doctor* and *nurse* in English, have been translated into other languages using certain gender-specific words more often that others: for example, *doctor* as *male doctor* and *nurse* as *female nurse*. Machine learning algorithms pick up on these tendencies, generalise from them, and eventually end up replicating them in their own translations – even to a larger extent than the training data, due to most machine learning algorithms' tendency to overgeneralise, to over-favour typicality.

Bias in MT is undesirable for two reasons. First, it subconsciously perpetuates stereotypes among speakers of the target language by cheaply injecting plentiful texts into the world in which doctors are disproportionately often male and nurses female (for example). Secondly, it produces translations which are sometimes factually incorrect: in a sentence such as "I am a doctor" the intended reading of *doctor* (male of female) depends on the human user's intention, which is unknown to the machine, and so the machine makes an unjustified, biased assumption instead to translate *doctor* as *male doctor*. A more satisfactory user experience would be if the machine asked the user at that point which reading of *doctor* they prefer, but present-day machine translators generally do not have the ability to ask such questions.

Ultimately, bias is caused by ambiguity. When the source text is ambiguous (= when it allows two or more readings, with two or more translations into the target language) and when the machine needs to produce a translation and cannot ask follow-up questions, then all the machine has to go on (= to decide which reading to choose) is its own biased understanding (metaphorically speaking) of how things usually are in the world. Hypothetically, if the machine's choices were completely random instead of

biased, then that would mitigate the social impact of bias: there would be no more perpetuation of stereotypes, nurses would now be translated as male half the time and as female the other half. But, importantly, it would still result in factually incorrect translations some of the time. In other words, the problem for MT is not so much the bias as the ambiguity: the fact that different languages encode the same messages with different amounts of collateral detail. To paraphrase Roman Jakobson, languages differ not in what *can* be said in them but in what *must* be said in them (Jakobson 1959).

Resolvable and unresolvable ambiguities

There are two kinds of ambiguity in MT which routinely result in biased and/or factually incorrect translations: resolvable ambiguity and unresolvable ambiguity.

If the source text contains an ambiguous expression but also, somewhere else in the text, contains a clue with the help of which the ambiguity can be disambiguated, then the ambiguity is resolvable. Example: "she is a doctor" were the pronoun *she* (and the fact that it co-refers with *doctor*) disambiguates the reading of *doctor* as female. Humans are proficient at picking up on such clues while machines are getting better at it all the time. It is theoretically possible that this problem can be solved by incrementally improving existing technology and that, one day, machines will achieve human parity at this task.

On the other hand, if the text contains an ambiguous expression but no clues which could help with its disambiguation, then the ambiguity is unresolvable. Example: "I am a doctor" where there are no clues anywhere in the text to help with deciding whether the intended reading of *doctor* is male or female. No artificial intelligence, however smart, will ever be able to resolve such ambiguities by inspecting only the text: this problem cannot be solved by improving existing technology. Human translators resolve such problems by inspecting the extralinguistic reality, typically by asking follow-up questions or simply looking who is talking. The challenge for MT is to acquire the ability to do that too.

While major MT providers such as Google and DeepL are busy solving the problem of resolvable ambiguities, Fairslator's contribution is that it offers a solution for ambiguities of the unresolvable kind.

Linguistic categories often affected by unresolvable ambiguities

Some linguistic categories are subject to unresolvable ambiguities (and therefore to bias) more often than others. Most prominently, unresolvable ambiguity affects gender (example: translating *doctor* as *male doctor* or *female doctor*). Gender bias in MT is well known and has been well studied (Savoldi *et al.* 2021).

In the author's experience, another category often affected by unresolvable ambiguities is grammatical number on second-person pronouns and verbs (translating *you* as either singular or plural) and, in combination with this, the register on forms of address (translating *you* as either formal or casual). Where English has *you* other languages often have a choice of two or more pronouns. Which one a machine translator chooses can be a biased choice in contradiction to the human user's intended meaning: how is a machine supposed to know whether, when I say "where are you?", I am talking

to one person or many, and in which register I wish to address them? This kind of MT bias has not been studied very well yet, the only contribution the author is aware of is Moryossef *et al.* (2019).

Fairslator is a tool for dealing with exactly these two kinds of unresolvable ambiguities: (1) gender and (2) second-person forms of address. In principle, unresolvable ambiguities can and do occur on any other aspect of meaning too, but that is outside of Fairslator's scope (for now).

How Fairslator works

It is important at this point to emphasize that Fairslator is not an MT system. It is a plug-in for other MT systems which filters and processes their output before it is shown to the human user. This section outlines how Fairslator works inside, and discusses its strengths and limitations.

The workflow: translate, detect, re-inflect

The process of getting something translated through Fairslator begins with a human user typing some text in the source language and selecting the MT provider they wish to use (currently: Google, DeepL, Microsoft). Fairslator sends the text to that provider's API and obtains a translation.

The two texts (source and translation) are then passed through Fairslator's bias detection algorithm. The algorithm looks for items in the source text which could be translated in more than one way (in terms of gender and forms of address) but where only one of these two ways actually occurs in the translation. When such a situation is detected by the algorithm, Fairslator shows the translation to the user along with disambiguatiin options as in Figure 1.

The user is now able to select alternative readings for the gender and forms of address of humans mentioned in the text. The user's choices are sent to Fairslator's re-inflection algorithm which updates the translation accordingly. The re-inflection process can range from trivial (substituting one word for another) to complex (substituting a word, then re-inflecting its modifiers such as adjectives so as not break grammatical agreement, then perhaps changing co-referring pronouns, and so on).

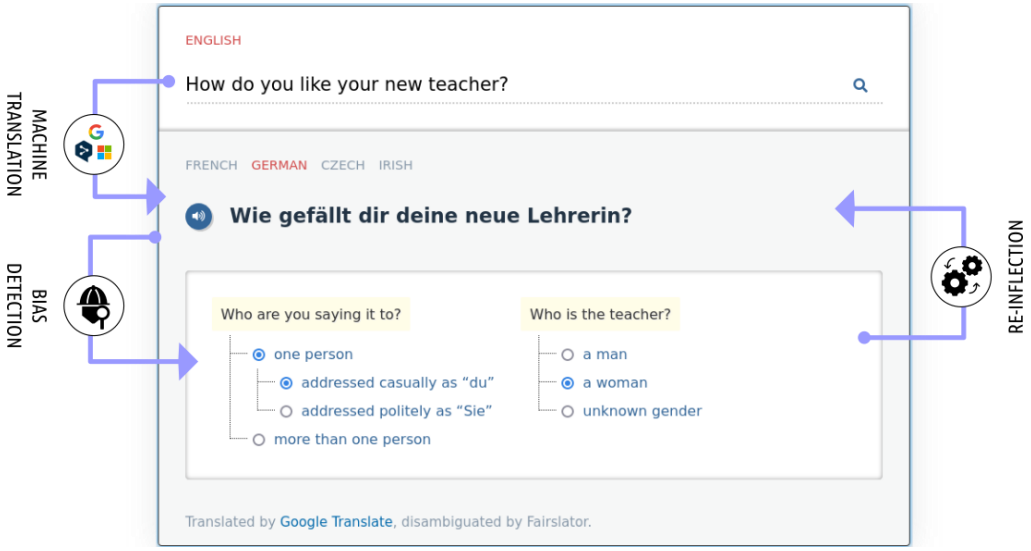


Figure 1. The Fairslator workflow

Classifying and describing the ambiguities

An important part of Fairslator is the ability to describe the ambiguities of a text, and to do it in a way a human user can understand, even if they do not speak the target language: with questions such as "who is the doctor, a man or a woman?" and "who are we saying this to, one person or many?". Internally, Fairslator represents the ambiguity of a text using a taxonomy developed for this project, for details see Měchura (2022b).

The technological stack

Internally, Fairslator makes use of UDPipe (Straka and Straková 2017), an industry-standard dependency parser, to parse texts in the source and target languages into syntax trees. On top of this Fairslator adds its own set of algorithms for detecting ambiguities and re-inflecting translations.

The algorithms are rule-based (ie. not the result of machine learning), have been hand-coded by the author, and make use of large-coverage lexicons. It takes between three and five days to prepare everything needed for a new language pair in Fairslator. This means that this method is relatively cheap compared to how industry giants such as Google usually approach computational problems: no Big Data is required for training and no expensive infrastructure is needed for machine-learning algorithms. The downside is that the person doing the hand-coding needs to be a skilled computational linguist with thorough knowledge of the grammar of the two languages in each language pair.

The fact that Fairslator is not the result of machine learning is also the reason why fixing errors and bugs is relatively easy. Unlike most machine learning artefacts, which are black boxes and even their author do not fully understand how they work, Fairslator's source code is readable to a programmer. When bugs are brought to the author's attention it is usually possible to fix them.

Limitations of the method

When Fairslator makes errors, for example when it fails to detect an ambiguity or when it re-inflects something incorrectly, it is usually caused by one of the following factors.

The underlying parser, UDPipe, has parsed something incorrectly. In such cases it is often – if not always – possible to compensate for it by building in a special case and re-writing the syntax tree before it goes further for analysis.

There is a gap in Fairslator's internal lexicon of gender-specific vocabulary. This is by far the commonest cause of errors, and also the easiest to fix by adding items to the lexicon.

Fairslator's ambiguity detection algorithm has failed to detect an ambiguity (a false negative) or claims an ambiguity where there is none (a false positive). These errors are usually quite intricate, fixing them requires changing the code, which requires good knowledge of the grammar of the language involved.

The taxonomy which Fairslator uses internally for describing ambiguities is not expressive enough to represent a given kind of ambiguity. Some discussion of these errors, which are rare, can be found in section 4 of Měchura (2022b).

Similar work elsewhere

Fairslator is of course not the only attempt to solve MT bias. Some big players on the MT market have been making tentative moves in this area, notably:

Google has, in its public web interface, implemented gender-related features for some language pairs which are somewhat similar to Fairslator, although the technology behind it is very different (and not as "cheap" as Fairslator's).

DeepL has a feature in its web interface and in its API for controlling forms of address (formal versus casual *you*) in translations. The user experience is somewhat similar to Fairslator.

Amazon has, as of 2022, added a similar feature for controlling forms of address to their API.[1]

A more detailed discussion of Google's and DeepL's approach to bias and ambiguity, compared to Fairslator's, can be found in section 3 of Měchura (2022a).

In academic research into MT bias, one contribution which is rather similar to Fairslator in its methodology if not in its user experience – it is, like Fairslator, a rule-based re-inflector – is by researchers at New York University Abu Dhabi for the English-to-Arabic language pair, see Alhafni *et al.* (2022).

Summary and conclusion

Most instances of bias in MT are caused by ambiguities which are unresolvable. As MT technology nudges closer and closer to human parity, the problem of unresolvable

[1] https://docs.aws.amazon.com/translate/latest/dg/customizing-translations-formality.html

ambiguities becomes more apparent as the only problem that cannot be solved simply by improving existing machine-learning technology. To properly solve this problem, machines need to acquire the ability to recognize unresolvable ambiguities when they occur and to know when to solicit disambiguation from humans. This paper has described an experimental application called Fairslator which does exactly this.

References

Alhafni, Bashar, Ossama Obeid, and Nizar Habash. 2022. "The User-Aware Arabic Gender Rewriter." arXiv. https://doi.org/10.48550/ARXIV.2210.07538.

Jakobson, Roman. 1959. "On Linguistic Aspects of Translation." In On Translation. https://web.stanford.edu/~eckert/PDF/jakobson.pdf.

Měchura, Michal. 2022a. "We Need to Talk about Bias in Machine Translation: The Fairslator Whitepaper." https://www.fairslator.com/fairslator-whitepaper.pdf.

Měchura, Michal. 2022b. "A Taxonomy of Bias-Causing Ambiguities in Machine Translation." In Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), 168–73. Seattle, Washington: Association for Computational Linguistics. https://aclanthology.org/2022.gebnlp-1.18.

Moryossef, Amit, Roee Aharoni, and Yoav Goldberg. 2019. "Filling Gender & Number Gaps in Neural Machine Translation with Black-Box Context Injection." In Proceedings of the First Workshop on Gender Bias in Natural Language Processing, 49–54. Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3807.

Savoldi, Beatrice, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. "Gender Bias in Machine Translation." Transactions of the Association for Computational Linguistics 9 (August): 845–74. https://doi.org/10.1162/tacl_a_00401.

Straka, Milan, and Jana Straková. 2017. "Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe." In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, 88–99. Vancouver, Canada: Association for Computational Linguistics. http://www.aclweb.org/anthology/K/K17/K17-3009.pdf.

Section C: Multi-word expressions, terminology and corpora

Expert data: a French MWE Manually Annotated Corpus

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

Fiorella Albasini
Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG, 38000 Grenoble, France
fiorella.albasini@etu.univ-grenoble-alpes.fr

Yacine Haddad
UT2J, Université Toulouse – Jean Jaurès, Toulouse, France
yacinehaddad.ut2j@gmail.com

Abstract

Multiword Expressions (MWE) are idiosyncratic expressions made of recurrent word combinations where the general meaning cannot be understood from the literal meaning of each constituent. Their identification is a demanding task in natural language processing (NLP). To figure out if manually annotated MWE corpora could be used to train Neural Networks to automatically detect MWE, we set up an experiment to have a French corpus, paraSHS-Témoigner (1,838 sentences and 57,162 words using Tutin's typology), annotated for MWE by three annotators on the Online Collaborative Annotation Platform ACCOLÉ. A total of 3,356 MWEs were annotated. We used the SemEval'13 metric adapted to MWE annotation to demonstrate the worthiness of manually annotated corpora. The first results on two annotators showed a high agreement with an F1-score at 0.71 on Strict cases (MWE delimitation and labelling), rising to 0.86 on Partial cases (overlap on the MWE delimitation) and 0.86 for Type labelling, not considering the MWE delimitation. This encourages the creation of manually annotated MWE corpora to train Neural Networks for MWE automatic detection.

Related work and State-of-the-Art

Multiword Expressions (MWE) are idiosyncratic expressions made of recurrent word combinations in which the general meaning cannot be understood from the literal meaning of each of its constituents (Firth, 1957; Sag et al., 2002). MWE identification is known to be a demanding task in natural language processing (NLP) (Sag et al., 2002; Baldwin and Kim, 2010; Bouamor, 2014). In machine translation (MT), one of the main sources of error is the incapability of recognizing MWEs (Constant et al., 2017). Despite the improved output quality, Neural MT (NMT) still struggles with MWEs (Rikters and Bojar, 2017; Zaninello and Birch, 2020, Han et al., 2020): Colson (2020) reports that Google Translate made mistakes in about 40% of MWE translations.

To study MWE, a lot of work has been done on their automatic identification, whether this is done while parsing other structures such as Mind the Gap (Coavoux et al., 2019), or specifically to spot MWEs as with MWEToolKit (Ramisch, 2015), and

* Institute of Engineering Univ. Grenoble Alpes

the PARSEME project (Savary et al., 2015). The latter also provides lexicons and annotated corpora.

This second approach is also well established in MT and NLP communities if we consider all the available MWE annotated corpora. Laporte et al. provided two French corpora, one dedicated to annotated Multiword Nouns (2008a) and the other to annotated MWEs with Adverbial Function (2008b). Another French corpus named PolyCorp (Tutin, 2015; Tutin and Esperança-Rodier, 2017) can annotate eight types of MWE of any part of speech. Vincze (2012) shared the SzegedParalelIFX corpus of parallel English-Hungarian texts annotated for light verb constructions. The multilingual corpus AlphaMWE (Han et al., 2020) provides annotations of verbal MWEs. Furthermore, several treebanks exist, among others, the French Treebank (Abeillé et al., 2003), the National Corpus of Polish (Głowińska and Przepiórkowski, 2010) annotated for Named Entities, and Głowińska (2012).

Those corpora are used to study the syntactic and semantic behaviour of MWEs, or to train, fine-tune, and test NMT systems. Nevertheless, to our knowledge, only PolyCorp deals with several types of MWEs and parts-of-speech while all the other corpora focus on one type of MWE or one part-of-speech and on continuous MWE. We thus wanted to follow PolyCorp's example by providing a new MWE annotated corpora on a large amount of MWE types, on any part of speech and on continuous and discontinuous MWE, to enable the study of a wider number of linguistic phenomena. We have chosen to work on paraSHS-Témoigner (Kraif, 2018) because it is composed of French documents along with their aligned translation in English, which is of interest to us as we want to further check the translation quality of MWE. While PolyCorp contains French texts from several domains such as scientific writing, news, extract of a novel, and film subtitles, paraSHS-Témoigner offers academic documents from social and human sciences.

To figure out if the annotation made by human annotators was of sufficient quality, we set up an experiment annotating this French corpus. As we wanted to see if the quality of the human annotations were consistent among the different annotators, we decided to focus on inter-annotator agreement.

After this overview of the state-of-the-art, this article introduces the experiment itself, by presenting in the following sections the experimentation protocol, along with the inter-annotator methodology used. In the remaining parts of this article, we present the metrics used to determine the inter-annotator agreement as well as the results obtained. We conclude with a discussion of the results and further work to be done. Examples extracted from the study will be found throughout.

Experiment

On top of focusing our study on the quality of the data, we wanted to provide a good amount of data. We asked three annotators to annotate the same documents to obtain more annotated data and to study the consistency among the annotators. We chose to evaluate the consistency of annotations by calculating the inter-annotator agreement using Precision, Recall, and F-measure as defined in the International Workshop on

Semantic Evaluation (SemEval'13) (Segura-Bedmar et al., 2013) adapted to MWE annotation.

a. Protocol

The experiment consisted in annotating with MWE a French corpus, paraSHS-Témoigner (Kraif, 2018), composed of 1,838 sentences and 57,162 words. To proceed with the annotation, we used ACCOLÉ, an Online Collaborative Annotation Platform (Esperança-Rodier and Brunet-Manquat, 2019) that allows us to annotate monolingual, bilingual, or multilingual corpora using predefined typologies. ACCOLÉ also allows collaboration: During the annotation, several annotators can comment on a specific MWE to agree on the type or the boundaries.

We used the Tutin (2015) typology developed during the Polycorp project. This typology addresses any MWE, continuous or discontinuous, assigning a part of speech to the MWE. For this experiment, although we asked the annotators to annotate the MWE using the typology parts-of-speech, we focused on inter-annotator agreement regarding only the boundaries and types of the MWE.

Three annotators were recruited, including one co-author, who is a French, English, and Spanish translator, graduated with a master's degree in Language Sciences, Linguistic specialty. The second annotator graduated with a master's degree in NLP, and the last annotator has a PhD in Language Sciences on Multiword Chaining in Language and Discourse from Laboratoire Textes, Théories, Numérique (TTN), Paris 13 University. Both latter annotators were French native speakers with a linguistic academic background.

We established the following workflow: training task and main task. For the training task, the annotators, referring to the annotation guide, were asked to annotate on ACCOLÉ the first 60 sentences of the French novel Thérèse Raquin for which we already had MWE annotations made by experts, to which we refer as gold annotations (Tutin et al., 2015). The annotations made by the annotators have been compared to the gold ones. When a difference between an annotation and the gold annotation was found, whether on the span of the MWE, or the type, the typology was again explained to the annotators and clarified with new examples in the annotation guide to ensure uniformity.

Once this training task was completed, the annotators started annotating the paraSHS-Témoigner corpus. Using ACCOLÉ, for each sentence, the annotators had to delimit the boundaries of the MWE according to the annotation guide and assign the corresponding type, still referring to the annotation guide.

If they had doubts, they could comment on the platform. Comments were shared with all annotators who consequently had to agree on a decision about the annotation, whether on the boundary or the MWE type according to the hesitation type. One of the annotators, the French-English-Spanish translator, oversaw the final decision. Finally, the annotators modified the boundary or the MWE type according to the final decision.

A total of 3,356 MWEs were annotated over 700 sentences. The typology is detailed in the following sub-section.

b. MWE Typology

Tutin et al. (2015) distinguish eight types of MWE as shown in Table 1. As the annotators only used five of the eight types while annotating in our experiment, we will only describe the five MWE types used for annotating.

MWE		**Example**
Idioms	Frozen multiword expressions	*Cul de sac* (fr) 'dead end'; *prendre en compte* (fr) 'take into account'
Collocations	Preferred binary association, including light verb constructions	*Gros fumeur* (fr) 'heavy smoker'; *faire une promenade* (fr) 'to take a walk'
Functional Multiword Expression	Functional adverbs, prepositions, conjunctions, determiners, pronouns	*C'est pourquoi (fr)* 'that is why'; *d'autre part* (fr) 'on the other hand'; insofar as
Pragmatic MWEs (pragmatemes)	Multiword expressions related to specific speech situations	*De rien* (fr) 'You're welcome'; *à plus tard*(fr) 'see you later'
Proverbs		*Pierre qui roule n'amasse pas mousse* (fr) 'A rolling stone gathers no moss'
Complex Terms		Natural Language Processing
Multiword Named Entities		*Université Grenoble Alpes*; the European Union
Routine Formulae	Routines generally associated with rhetorical functions	*Force est de constater* (fr) 'it must be noted'

Table 7: MWE Typology (Tutin et al., 2015)

This description of MWE types is taken from two papers: the article by Tutin et al. (2015) and the annotation guide established during the experimentation. The examples are taken from the annotation set.

Idioms: those are expressions whose meaning of the words is non-compositional, that is to say, they cannot be deduced from the meaning of the parts. In the example below, the expression ***point de vue*** 'point of view' is annotated as "Idiom".

*(2) « Wolf vise à faire entendre le **point de vue** des vaincus dans un récit »*

'Wolf aims to convey the **point of view** of the vanquished in a narrative.'

Collocations: those are compositional expressions, mainly binary, tending to frequently co-occur, where one word keeps its usual meaning while the other is more unpredictable. In the example below, the expression ***a l'impression*** 'feel like' has been annotated as an MWE of type "collocation".

*(3) « on **a l'impression** que le bateau continue de prendre l'eau »*

'it **feels like** the boat is still taking on water'

Functional MWEs: those MWEs are characterized by a functional meaning. They include grammatical words such as conjunctions (even if), prepositions (before),

pronouns (something), adverbs, etc. In (2), the multiword expression ***du haut des*** 'from atop' received the type "function word"

*(4) « **du haut des** murailles de Troie, elle apostrophe ses compatriotes pour les appeler à manifester... »*

'**From atop** the high walls of Troy, she shouts to her compatriots to call them to express'

Multiword named entity: "*Étant donné un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus*." 'Given an application model and a corpus, a named entity is any linguistic expression that refers to a single entity of the model autonomously in the corpus.' is the definition given by Ehrmann (2008). We find, in general, proper names, dates, durations, events, organizations, etc. In (3), the phrase ***L'Iliade d'Homère*** 'Homer's *Iliad*' is annotated as a named entity.

*(5) « Cassandre, [...], apparaît brièvement **dans L'Iliade d'Homère**... »*

'Cassandra, […], briefly appears **in Homer's *Iliad***.'

Proverbs: namely proverbs as in the example (5) below:

*(6) « Pour le dire de façon un peu sommaire : **les ennemis de nos ennemis ne sont pas nécessairement nos amis** »*

'To say it somewhat briefly: the enemies of our enemies are not necessarily our friends.'

Routine Formulae: These are prefabricated verbal expressions, frequent in a specific genre, such as ***pour ainsi dire*** 'so to speak/as it were'. In the following example,

*(7) « L'accumulation de ces dizaines de milliers d'heures maintenant en phrase de numérisation, **pour ainsi dire**, la promesse d'une éternité digitale ne pallie pas plus la disparition des témoins »*

'But the accumulation of tens of thousands of hours, currently being filed away, and the promise of a digital eternity **as it were**, do not solve the question of the disappearing of the witnesses.'

a. Inter-annotator Agreement Issues

To use these annotations, we had to re-organize the CSV output of ACCOLÉ so that we could better focus on the overlapping of annotations among annotators.

Annotating MWE is a dual-task process, as the annotator must first identify the MWE (select the units in the text) and then assign a type to those units. Since this is not a simple classification task, calculating inter-annotator agreement requires metrics that consider both the unit selection and the type labelling. In other words, it is necessary to calculate to what extent the annotators agree not only on the assignment of types but also on the prior identification of their boundaries.

The measure used to evaluate the inter-annotator agreement was proposed during SemEval'13 on the evaluation of Named Entities (Segura-Bedmar et al., 2013). We adapted it so that it is operational on the assessment of MWE annotations. As we did not have a gold standard, the metric consists in using one of the annotators as the gold standard and comparing it with the annotations from the other annotators, two by two. We develop the above in further detail in the following section.

Inter-annotator Agreement

b. Methodology

Inspired by the SemEval'13 metric for Named Entities (Segura-Bedmar et al., 2013), the overall idea is to consider Annotator 1 as the gold standard and Annotator 2 as the system output. In what follows we will refer to Annotator 1 as the gold standard using the expression "Gold annotator" and to Annotator 2 as the system output using the expression "annotator output". The metric proposes four cases to measure the precision, recall, and F-measure between the annotators:

STRICT: unit boundaries and type strictly match

EXACT: exact match of the unit boundaries, regardless of the type

PARTIAL: partial match of the unit boundaries (overlap), regardless of the type

TYPE: exact match of the types, regardless of the unit boundaries.

These four cases of annotation can be assigned one of the values from the five scores proposed in MUC (Message Understanding Conference) (Chinchor and Sundheim, 1993):

Correct (COR): the annotator output and the gold annotator match

Incorrect (INC): the annotator output and the gold annotator do not match

Partial (PAR): the annotator output and the gold annotator are somewhat similar, but not identical

Missing (MIS): The gold annotator annotation is not captured by the annotator output

Spurious (SPU): the annotator output produces an annotation that is not in the ones achieved by the gold annotator

The four cases of annotation, concerning these five scores, are used to establish correspondences between the annotations of each annotator. All cases such as "**STRICT**", "**EXACT**", "**PARTIAL**" and "**TYPE**" annotations can be assigned to the "**Missing**" score (when some error is present in the gold annotator but not detected by the annotator output) and "**Spurious**" score (when the annotator output detects any error that is not in the gold annotator annotations). They can also be set to "**Correct**", "**Incorrect**" or "**Partial**" scores:

"**STRICT**" is scored "**Correct**" when there is a strict match in terms of boundaries and MWE type, as in the following example (7): the two annotators, selected

the same sequence of units ***manifester leur deuil*** ‘to express their grief’ and attributed the same MWE type “Collocation”. If this condition is not met, "**STRICT**" is scored “**Incorrect**”.

*(8) Cassandre, fille du roi Priam, apparaît brièvement dans L'Iliade d'Homère : du haut des murailles de Troie, elle apostrophe ses compatriotes pour les appeler à **manifester leur deuil** au retour du cadavre d'Hector.*

‘Cassandra, daughter of King Priam, briefly appears in Homer’s *Iliad*. From atop the high walls of Troy, she shouts to her compatriots to call them **to express their grief** after Hector returns dead.’

Gold annotator: (***manifester leur deuil***, Collocation) ‘to express their grief’

Annotator output: *(**manifester leur deuil***, Collocation*)* ‘to express their grief’

“**EXACT**” is scored “**Correct**” when there is an exact match of the unit boundaries, regardless of the assigned type of MWE, as in the example (8): both annotators have selected the same sequence of units ***au bas de*** ‘to the bottom of’ but the gold annotator assigned the MWE type “Function Word” and the annotator output the type “Collocation”. If this condition is not met, “**EXACT**” is scored “**Incorrect**”.

*(9) […] il a proféré le terrible constat – non de la disparition des témoins, lui qui allait le 11 avril 1987 se jeter du troisième étage **au bas de** l'escalier de son immeuble.*

[…] he acknowledged not the disappearing of the witnesses (on 11 April 1987 he who would throw himself from the third floor **to the bottom of** the stairs in his apartment block)

Gold annotator: (*au bas de*, Function Word) ‘to the bottom of’

Annotator Output: (*au bas de*, Collocation) ‘to the bottom of’

“**PARTIAL**” is scored “**Partial**” when there is a partial match of unit boundaries (at least one character in common), regardless of the MWE type, as in the example (9) below: the gold annotator selected ***les droits de l'homme*** ‘the human rights’, including the determiner, while the annotator output selected ***droits de l'homme*** ‘human rights’ excluding the determiner (overlapping). If this condition is not met, “**PARTIAL**” is scored “**Correct**”. For further explanation, where “**EXACT**” is scored “**Correct**”, so is “**PARTIAL**”; when “**EXACT**” is scored “**Incorrect**”, “**PARTIAL**” is scored “**Partial**”, which makes sense because if the boundaries of the two units are exactly aligned, then there is no question of overlap.

*(10) […] en considérant qu'un questionnement sur les fondements du monde que nous voulons, résolument ancré sur **les droits de l'homme**, doit passer par Auschwitz, tout autant que par la critique de modèles […]*

‘by stating that any investigation into the foundations of our “ideal” world, which we so resolutely want to establish on **human rights**, should reckon with Auschwitz and critically review the models’

Gold annotator: (*les droits de l'homme*, Full Phraseme) ‘the human rights’

Annotator output: (*droits de l'homme*, Collocation) ‘human rights’

“**TYPE**” is scored “**Correct**” when both annotators assign the same MWE type, regardless of the boundaries of the selected units, as in the example (10) hereafter: the gold annotator selected ***Agamemnon d’Eschyle*** ‘*Agamemnon* by Aeschylus’, without the determiner, while the annotator output selected ***l’Agamemnon d’Eschyle*** ‘The *Agamemnon* by Aeschylus’, with the determiner. They both assigned the same MWE type “Named Entity”. If this condition is not met, “**TYPE**” is scored “**Incorrect**”.

*(11) Prophétesse inspirée par Apollon (à partir de **l'Agamemnon d'Eschyle**) ou faisant bon usage de sa raison (dans nombre de versions modernes), elle devient une figure […]*

‘As a prophet inspired by Apollo (from **the *Agamemnon* by Aeschylus**) and through the good use of her reason (in many modern versions), she becomes a figure of inaudible knowledge’

Gold annotator: (*Agamemnon d’Eschyle*, Named Entity) ‘*Agamemnon* by Aeschylus’

Annotator output: (*l’Agamemnon d’Eschyle*, Named Entity) ‘The *Agamemnon* by Aeschylus’

Table 2 illustrates the cases we have just mentioned with the scores to which they have been set. As an example, we calculate the inter-annotator agreement of annotations in this table.

Gold Annotator		**Annotator**		**Evaluation Scheme**			
Phrase	**MWE Type**	**Phrase**	**MWE Type**	**TYPE**	**PARTIAL**	**EXACT**	**STRICT**
-	-	une nouvelle fois ‘one more time’	Collocation	SPU	SPU	SPU	SPU
Agamemnon d’Eschyle	Named Entity	l’Agamemnon d’Eschyle	Named Entity	COR	PAR	INC	INC
au bas de	Function Word	au bas de	Collocation	INC	COR	COR	INC
manifester leur deuil	Collocation	manifester leur deuil	Collocation	COR	COR	COR	COR
Les droits de l’homme	Full Phraseme	droits de l’homme	Collocation	INC	PAR	INC	INC

Table 2: Inter-Annotator Agreement Evaluation Scheme Matrix with examples

a. Metrics

From above table 2, we have to calculate two values: the "**Possible(POS)**" which corresponds to the sum of the annotations of the gold annotator (true positives–TP + false negatives–FN) for each of the 4 cases:

Possible(POS) = COR+INC+PAR+MIS=TP+FN

As well as the "**Actual(ACT**)" which is the sum of the effective annotations of the annotator output (true positives–TP + false positives–FP) for each of the 4 cases.

Actual(ACT) = COR+INC+PAR+SPU=TP+FP

The two "**Possible(POS)**" and "**Actual(ACT)**" sums will allow us to calculate the Precision, the Recall, and the F-measure, for each of the 4 cases "**STRICT**", "**EXACT**", "**PARTIAL**" and "**TYPE**". For the "**STRIC**" and "**EXACT**" cases, the Precision is calculated as the standard Precision ($Precision_{Std}$) metric is, by dividing the "**Correct**" of each case by the "**Actual(ACT)**", which corresponds, as illustrated in Equation 1 below, to the true positives divided by the sum of the true positives and the false positives. The recall is also calculated as the standard Recall ($Recall_{Std}$) metric is, by dividing the "**Correct**" of each case by the "**Possible(POS)**", which corresponds to the true positives divided by the sum of the true positives and the false negatives.

$$Precision_{Std} = \frac{COR}{ACT} = \frac{TP}{TP + FP}$$

$$Recall_{Std} = \frac{COR}{POS} = \frac{TP}{TP + FN}$$

Equation 8: Standard Precision and Recall formula

For the two other cases "**PARTIAL**" and "**TYPE**", the Precision ($Precision_{PC}$)is calculated by multiplying the "**Partial**" of each case by 0.5 (coefficient used to set the value of the "**TYPE**" and "**PARTIAL**" at 0.5 point compared to "**STICT**" and "**EXACT**"), plus the "**Correct**" of each case, divided by the "**Actual(ACT)**" sum. The Recall ($Recall_{PC}$) is calculated by dividing these same values by the "**Possible(POS)**" sum, as can be seen in Equation 2, below. In our case, we do not get any "**Partial**" for "**TYPE**" as there is no hierarchy in the MWE typology we use, and thus all MWE types are distinct.

$$Precision_{PC} = \frac{COR + 0.5 \times PAR}{ACT} = \frac{TP + 0.5 \times PAR}{TP + FP}$$

$$Recall_{PC} = \frac{COR + 0.5 \times PAR}{POS} = \frac{TP + 0.5 \times PAR}{TP + FN}$$

Equation 2: Partial Case Precision and Recall formula

From there, we can calculate the F-measure for each of the 4 cases, as shown in Table 3, using the standard F-measure formula as shown in Equation 3.

$$F - measure = 2\,\frac{Precision \times Recall}{Precision + Recall}$$

Equation 3: F-measure formula

For example, to calculate the Precision for the "**STRICT**" case in this example: we divide the number of "**Correct**" (=1) by the "**Actual(ACT)**" (=5), which gives 0.2 of Precision. Table 3 shows the agreement results for this example.

Measure	**TYPE**	**PARTIAL**	**EXACT**	**STRICT**
Correct	2	2	2	1
Incorrect	2	0	2	3
Partial	0	2	0	0
Missing	0	0	0	0
Spurious	1	1	1	1
Possible(POS)	4	4	4	4
Actual(ACT)	5	5	5	5
Precision	0.4	**0.6**	0.4	0.2
Recall	0.5	**0.75**	0.5	0.25
F-measure	0.44	**0.66**	0.44	0.22

Table 3: Precision, Recall, and F-measure Matrix from the Evaluation Scheme. Precision and Recall are calculated using $Precision_{Std}$ and $Recall_{Std}$ formula for STRICT and EXACT cases and $Precision_{PC}$ and $Recall_{PC}$ for PARTIAL and TYPE.

Result

Given that the three annotators' results are virtually the same, we provide the evaluation of the agreement between two annotators only to avoid redundancy. As explained in Section 2.1 above, annotators agreed on how to annotate boundaries and types on difficult occurrences of MWEs. This allowed us to ensure uniformity in annotations up to a certain point, and consequently show the results on the comparison of only the two first annotators.

Measure	**TYPE**	**PARTIAL**	**EXACT**	**STRICT**
Correct	694	599	599	575
Incorrect	71	0	166	190
Partial	0	166	0	0
Missing	41	41	41	41
Spurious	35	35	35	35
Possible(POS)	806	806	806	806
Actual(ACT)	800	800	800	800
Precision	**0.86**	0.84	0.74	0.71
Recall	**0.87**	0.85	0.75	0.72
F-measure	**0.86**	0.84	0.74	0.71

Table 4: Precision, Recall, and F-measure Matrix from the Experiment (whole data)

Table 4 shows the results of the evaluation of inter-annotator agreement. We observe that the agreement is good overall, despite the distinctions between the four cases of annotation: less agreement among the "**EXACT**" and "**STRICT**" than among the "**TYPE**" and "**PARTIAL**" cases. We also notice that the number of "**Missing**" and "**Spurious**" is not very high, which explains the high value of the recall. Those results on two annotators show a quite high agreement with an F-measure at 0.74 on "**EXACT**" cases and 0.71 on "**STRICT**" cases. F-measure rises to 0.84 on "**PARTIAL**" cases, when there is only overlap on the MWE boundaries, and reaches 0.86 for "**TYPE**" cases, not considering the delimitation of the MWE.

To conclude this section, the evaluation of the MWE inter-annotator agreement gave good results, which is most likely the result of several elements. First, an annotation guide was established and enriched as the annotation progressed. In addition, the annotators had clear instructions and collaborated on the ACCOLÉ platform solving a lot of issues while annotating. Finally, the MWE typology does not contain many types which helped the annotators to stay consistent.

Conclusion and Discussion

This study has demonstrated that high inter-annotator agreement (0.86 on the MWE labelling case) can be reached between three annotators, using the annotation platform ACCOLÉ with a complete MWE typology (Tutin et al., 2015) on an MWE annotation task. This means that one can create high-quality MWE-annotated corpora following our methodology, thus addressing the first issue we mentioned in Section 2: i.e. the need for high-quality data to address specific linguistic issues.

As regards the second issue mentioned in Section 2 which is the amount of data needed to train Neural Networks (NN), this is another matter. The threshold on the amount of data to train or fine-tune a NN System depends on the task to be achieved (MT, Word Sense Disambiguation, etc.). In the MWE annotation case, as far as we know, no work has yet been published on the required size of corpora to enable training or fine-tuning of NN Systems. We believe that our results show that human annotation is consistent enough to be used for this purpose. Nevertheless, without a defined threshold for the amount of data required, it is difficult to establish such a threshold. Consequently, looking at the already existing huge corpora such as PARSEME or the others mentioned in Section 1, we would rather take the evaluation approach. Hence, we will use those huge corpora to train and fine-tune the NN systems for the MWE annotation task and use our smaller but higher quality corpus, as we deal with much more MWE types, to test the NN Systems, and thus process to a quality assessment of those systems.

Furthermore, we can notice that delimitation issues in terms of MWE boundaries lower the annotator agreement, however, we could face this phenomenon as a natural aspect of the annotation process. NN Systems can benefit from the detection of MWEs even with non-standardized boundaries. The fact that several annotators select overlapping MWE boundaries indicates the possibility of having a potential MWE in the area delimited by those boundaries. We also found out that the inter-annotator agreement increased when annotators used the discussion feature of the platform while

annotating. This allowed them to agree on the boundary as well as on the type of the MWE after examining the MWE typology.

Future work will focus on the use of decision flowcharts while annotating on top of using the discussion feature. We will also set up experiments to find out the right amount of data necessary to train or fine-tune NN systems on MWE annotation tasks. On the inter-annotator agreement protocol, we would like to further investigate best practice when the Gold annotator disagrees with the other two annotators. We would also consider doing the same calculation by pairs, using each one of the annotators in turn as the Gold annotator. Finally, we will train and fine-tune NN systems on existing big MWE annotated corpora and use our high-quality level corpus to test the NN systems.

Acknowledgements

This work has been supported and financed by the Institut Cognition, and the LIG via Emergence funding.

References

Abeillé, A., Clément, L. and Toussenel, F. (2003) Building a treebank for French, in *Treebanks*, Kluwer, Dordrecht. (p.165-187)

Baldwin, T., & Kim, S. N. (2010). Multiword expressions. Handbook of natural language processing, 2, 267-292.

Bouamor, D. (2014). Constitution de ressources linguistiques multilingues à partir de corpus de textes parallèles et comparables. Université Paris Sud - Paris XI,

Chinchor, N and Sundheim, B. (1993) MUC-5 Evaluation Metrics. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*.

Coavoux, M., Crabbé, B., Cohen, S. H. (2019) Unlexicalized Transition-based Discontinuous Constituency Parsing. *Transactions of the Association for Computational Linguistics* 2019; 7 73–89 DOI: https://doi.org/10.1162/tacl_a_00255.

Colson, J.-P. (2020) Computational phraseology and translation studies: from theoretical hypotheses to practical tools. In: Gloria Corpas Pastor, Jean-Pierre Colson, *Computational Phraseology*, John Benjamins: Amsterdam / Philadelphia 2020, p. 65-81

Constant, M., Eryiğit, G., Monti, J., Van Der Plas, L., Ramisch, C., Rosner, M. and Todirascu, A. (2017) Multiword Expression Processing: A Survey. Computational Linguistics, 43(4), 837-892.

Ehrmann, M. (2008). Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation. Thèse de doctorat, Université Paris 7 - Centre de recherche Xerox, Grenoble (XRCE).

Esperança-Rodier, E., Brunet-Manquat, F., Eady, S. (2019). ACCOLÉ: A Collaborative Platform of Error Annotation for Aligned Corpora. Translating and the computer 41, Nov 2019, Londres, United Kingdom. ⟨hal-02363208⟩

Firth, J.R. (1957). A Synopsis of Linguistic Theory, 1930-1955 Studies in Linguistic Analysis Special Volume, Philological Society. 1-32.

Głowińska, K. and Przepiórkowski, A. (2010). The design of syntactic annotation levels in the National Corpus of Polish. In Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010, Valletta, Malta. European Language Resources Association (ELRA).

Głowińska, K. (2012). Anotacja składniowa. In Adam Przepiórkowski, Mirosław Ba´ nko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors, Narodowy Korpus J¸ ezyka Polskiego, pages 107–127. Wydawnictwo Naukowe PWN, Warsaw.

Han, L., Jones, G. J. F. and Smeaton, A. (2020). AlphaMWE: Construction of Multilingual Parallel Corpora with MWE Annotations in the Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons Proceedings of the Workshop (MWE-LEX 2020). 2020. <hal-03013612>.

Kraif, O. (2018). Constitution et traitement d'un corpus bilingue d'articles scientifiques : exemple de mise en œuvre automatique avec une architecture légère en Perl. In *Journées LTT 2018*, sept. 2018, Grenoble

Laporte, E., Nakamura, T., & Voyatzi, S. (2008a). A French corpus annotated for multiword nouns. In Language Resources and Evaluation Conference. Workshop Towards a Shared Task on Multiword Expressions. pp. 27-30.

Laporte, E., Nakamura, T., & Voyatzi, S. (2008b). A French corpus annotated for multiword expressions with adverbial function. In Language Resources and Evaluation Conference (LREC). Linguistic Annotation Workshop. pp. 48-51.

Ramisch, C. (2015) Multiword Expressions Acquisition: A Generic and Open Framework", Theory and Applications of Natural Language Processing series XIV, Springer, ISBN 978-3-319-09206-5, 230 p.

Rikters, M., and Bojar, O. (2017) Paying Attention to Multi-Word Expressions in Neural Machine Translation. Preprint.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002) Multiword Expressions: A Pain in the Neck for NLP. International Conference on Intelligent Text Processing and Computational Linguistics. 1-15. Springer, Berlin, Heidelberg.

Savary, A., Sailer, M., Parmentier, Y., Rosner, M., Rose´n, V., Przepio´rkowski, A., Krstev, C., Vincze, V., Wo´jtowicz, B., Losnegaard, G. S., Parra Escart´ın, C., Waszczuk, J., Constant, M., Osenova, P., and Sangati, F. (2015). PARSEME–PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*

Segura-Bedmar, I., Martínez, P., and Herrero-Zazo, M. (2013). SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

Tutin, A. & Esperança-Rodier, E. (2017). La difficile identification des expressions polylexicales dans les textes : critères de décision et annotation. In *"La phraséologie française : débats théoriques et dimensions appliquées (didactique, traduction et traitement informatique)"*, Sep 2017, Arras, France.

Tutin, A., Esperança-Rodier, E., Iborra, M., and Reverdy, J. (2015). Annotation of multiword expressions in French, in *Proceedings of the. European Society of Phraseology Conference (EUROPHRAS 2015), Jun 2015, Malaga, Spain. pp.60-67.*

Vincze, V. (2012). Light verb constructions in the SzegedParalellFX English–Hungarian parallel corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC'12), pages 2381–2388, Istanbul, Turkey, May. European Language Resources Association

Zaninello, A. and Birch, A. (2020) 'Multiword Expression Aware Neural Machine Translation', Proceedings of the 12th Language Resources and Evaluation Conference. 3816-3825. Marseille, France, 11-16 May.

gApp: a text preprocessing system to improve the neural machine translation of discontinuous multiword expressions

Carlos Manuel Hidalgo-Ternero
Universidad de Málaga
Avda. Cervantes, 2. 29071
Malaga, Spain
cmhidalgo@uma.es

Xiaoqing Zhou-Lian
Universidad Rey Juan Carlos
Paseo de los Artilleros s/n. 28032
Madrid, Spain
xiaoqing.zhou@urjc.es

Abstract

In this paper we present research results with *gApp*, a text-preprocessing system designed for automatically detecting and converting discontinuous multiword expressions (MWEs) into their continuous forms so as to improve the performance of current neural machine translation systems (NMT) (see Hidalgo-Ternero, 2021 & 2022, Hidalgo-Ternero & Corpas Pastor, 2020, 2022a & 2022b, Hidalgo-Ternero, Lista, and Corpas Pastor, 2022, and Hidalgo-Ternero and Zhou-Lian, 2022a & 2022b). To test its effectiveness, eight experiments with several NMT systems such as DeepL, Google Translate, ModernMT and VIP have been carried out in different language directionalities (ES/FR/IT > ES/EN/DE/FR/IT/PT/ZH) for the translation of somatisms, i.e., MWEs containing lexemes referring to human or animal body parts (Mellado Blanco, 2004). More specifically, we have analysed both flexible verb-noun idiomatic constructions (VNICs) and flexible verb + prepositional phrase (VPP) constructions. In this regard, the promising results obtained for these typologies of MWEs throughout experiments 1-8 will shed some light on new avenues for enhancing MWE-aware NMT systems.

Introduction

The recent emergence of neural networks in natural language processing has represented a real breakthrough in the field of machine translation, bringing forth Neural Machine Translation (NMT), which has resulted in a considerable qualitative leap compared to previous ruled-based and statistical models (Bentivogli et al., 2016; Junczys-Dowmunt et al., 2016; Shterionov et al., 2018). Despite these advances, NMT systems still have an important weak spot: multiword expressions (MWEs). As well as their quintessential problematic features such as syntactic anomaly, non-compositionality, diasystematic variation and ambiguity, among others, a further challenge arises for NMT: MWEs do not always consist of adjacent tokens (e.g., *He took all their remarks into consideration.*), which seriously hinders their automatic detection and translation (Corpas Pastor, 2013; Foufi et al., 2019; Monti et al., 2018; Ramisch & Villavicencio, 2018; Rohanian et al., 2019). To overcome the challenges that discontinuous MWEs still pose for even the most robust NMT systems (cf. Colson,

2019; Zaninello & Birch, 2020), we have designed *gApp*,[1] a text-preprocessing system for the automatic identification and conversion of discontinuous MWEs into their continuous form in order to improve NMT performance. In this regard, 8 experiments, summarised in the *Results* section, have been carried out to prove *gApp*'s effectiveness.

Against this background, the remainder of the paper is structured as follows. Section 2 illustrates the research methodology. In Section 3, *gApp*'s precision and recall from experiments 1-8 is tested, in order to then assess to what extent this system can enhance NMT performance under the challenge of MWE discontinuity. Finally, Section 4 provides concluding remarks on how to further enhance MWE-aware NMT systems through *gApp*.

Methodology

This section presents the research methodology employed in order to assess to what extent *gApp* can optimise the performance of the NMT systems of DeepL, Google Translate, ModernMT and VIP in different language directionalities (ES/FR/IT > ES/EN/DE/FR/IT/PT/ZH). Analogously to Hidalgo-Ternero (2020), the concordances containing the discontinuous somatisms under study have been retrieved from two giga-token web-crawled corpus families (TenTen and Timestamped JSI web corpus) and the subcorpora available for the different languages under study (esTenTen18 and Timestamped JSI web corpus 2014-2021 Spanish, for Spanish; enTenTen20 and Timestamped JSI web corpus 2014-2021 English, for English, etc.). All these corpora are accessible through the corpus management system Sketch Engine (Kilgarriff et al., 2004).

The MWEs analysed through experiments 1-8 belong to the category of idiomatic expressions, since they have a non-compositional meaning (which is why they are also defined as semantically non-decomposable idioms or SNDIs [Bargman & Sailer, 2018]). Concerning their fixedness, following Parra et al.'s (2018) taxonomy for MWEs in Spanish, they can be classified as flexible, since other elements can appear embedded within the constituents of the MWEs. With regards to their morphosyntactic structure, they belong to two main categories: verb-noun idiomatic constructions (VNICs) and verb + prepositional phrase (VPP) constructions. Finally, considering the nature of their constituents, they are somatisms, i.e., idioms containing terms that refer to human or animal body parts (Mellado Blanco, 2004). In this regard, we have decided to analyse specifically idiomatic expressions because their non-compositional meaning makes them become potentially easier to detect and translate by NMT systems when all the constituents are contiguous, as we proved in experiments 1-8 (see Table 1 in the *Results* section).

Despite the challenges that user-generated content's (UGC) ubiquitous source-text error, noise and out-of-vocabulary tokens still pose to even the most robust NMT systems (Belinkov & Bisk, 2018; Lohar et al., 2019), a heterogeneous sample in terms of language varieties, text sources and types (including UGC) was selected for the

[1] *gApp* is accessible through the following link: http://lexytrad.es/gapp/app.php . This application is registered in Safecreative: https://www.safecreative.org/work/2011165898461-gapp.

analysis so as to alleviate sampling bias, which could otherwise originate from uniquely examining NMT canonical training data for the somatisms under study. In this way, a total of 3360 cases was analysed, comprising 1680 discontinuous and 1680 continuous forms (i.e., after the conversion) of somatisms, split by different unigrams, bigrams or trigrams. Besides these relevant results, for each somatism 50 irrelevant results (i.e., concordances containing analogous patterns to the MWEs but unrelated to the idiomatic sequences) were compiled, in order to calculate, at a first stage, both the precision and recall of this system, considering all the constituents of the MWE.

Once both parameters were quantified, at a second stage, the results concerning the NMT performance for the different concordances were classified within three main categories: before *gApp*, after the automatic conversion with *gApp*, and after the manual conversion, which hence constituted our gold standard. The same study was conducted for all the language directionalities. The NMT outputs for these different scenarios were then manually assessed following an instance-based MT evaluation (Zaninello & Birch, 2020) with several possible target-text candidates for each of the somatisms in both their continuous and discontinuous forms. To this end, morphological, syntactic, and/or orthotypographic divergences or source-text/translation imprecisions affecting other elements in the sentences were not considered *per se* as errors if they were unrelated to the phenomenon of MWE discontinuity for the somatisms under study.

Results

Eight different experiments, summarised in Table 1, have been carried out to prove *gApp*'s effectiveness.

	Type of MWE	**Language directionalities**	**NMTs analysed**	**NMTs' accuracy before *gApp***	**NMTs' accuracy after *gApp***	**Improvement after *gApp***	**Manual conversion**
1.	VNC	ES>EN	DeepL	80.7%	90.7%	10%	+3.2%
			Google Translate	60.7%	75.4%	14.6%	+2.1%
2.	VNC	ES>EN	DeepL	49%	62.5%	13.5%	+0.5%
		ES>DE		43.5%	52.5%	9%	+0.5%
3.	VNC	FR>EN	DeepL	40%	58%	18%	=
		FR>ES		41.5%	58%	16.5%	=
4.	VPP	ES>EN	Modern MT	50%	60%	10%	=
		ES>DE		23.3%	33.3%	10%	=
		ES>FR		49.3%	60%	10.7%	=
		ES>IT		56.7%	60.7%	4%	=
		ES>PT		56%	58.7%	2.7%	=
		ES>EN	DeepL	70.7%	81.3%	10.7%	+0.7%
		ES>DE		59.3%	66.7%	7.3%	+0.7%
		ES>FR		69.3%	74%	4.7%	+0.7%
		ES>IT		76%	80%	4%	+0.7%

		ES>PT		68%	74%	6%	+0.7%
		ES>EN	Google Translate	66%	75.3%	9.3%	=
		ES>DE		35.3%	43.3%	8%	=
		ES>FR		65.3%	73.3%	8%	=
		ES>IT		78.7%	79.3%	0.7%	=
		ES>PT		72.7%	79.3%	6.7%	=
5.	VPP	ES>EN	VIP	45.5%	67%	21.5%	-0.5%
		ES>EN	DeepL	77%	85.5%	8.5%	=
		ES>EN	Google Translate	64%	77.5%	13.5%	-1%
6.	VPP	IT>EN	Modern MT	25.5%	42%	16.5%	+1%
		IT>DE		28%	37%	9%	=
		IT>EN	Google Translate	64.5%	82.5%	18%	+0.5%
		IT>DE		50.5%	61%	10.5%	-1%
		IT>EN	DeepL	75%	75%	0%	-0.5%
		IT>DE		62%	67.5%	5.5%	=
7.	VNC	ES>EN	Google Translate	21.5%	25%	3.5%	=
			DeepL	57%	54.5%	-2.5%	=
		ES>ZH	Google Translate	11%	14%	3%	=
			DeepL	42.5%	39.5%	-3%	=
8.	VPP	ES>EN	Google Translate	13.6%	66.0%	52.4%	+0.8%
			DeepL	66%	90%	24%	=
		ES>ZH	Google Translate	13.6%	64.8%	51.2%	+0.8%
			DeepL	54.8%	73.6%	18.8%	=
Total (experiments 1-8)				**52.1%**	**66.8%**	**14.6%**	**+0.5%**

Table 1. *gApp* results through experiments 1-8

In Table 1, it is possible to observe a considerable improvement in NMT performance from 52.1% before *gApp* up to 66.8% after *gApp* (i.e., a final enhancement by 14.6%). Global results have also shown how *gApp*'s automatic conversion managed to achieve an analogous performance to the manual conversion (with only a 0.5% difference between the two types of conversion). This is chiefly due to *gApp*'s refined detection system both in terms of final average precision (95.9%) and recall (97.3%), which means that only 4.1% of the irrelevant results could enter the system and exclusively 2.7% of the relevant results were not successfully detected. A summary of *gApp*'s precision and recall through experiments 1-8 can be observed in Table 2.

	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5	Exp. 6	Exp. 7	Exp. 8	Global
Precision	94.8%	95.1%	96.1%	94.9%	95.2%	98.3%	95.3%	97.3%	95.9%
Recall	96.8%	97.5%	98.5%	99.3%	96%	92%	99.5%	98.8%	97.3%
F_1	95.8%	96.3%	97.3%	97.1%	95.6%	95.2%	97.4%	98.1%	96.6%

Table 2. *gApp* precision and recall through experiments 1-8

Other interesting findings can be observed in target-text errors in different language directionalities due to NMT pivoting through English. In this regard, let us contrast some instances of DeepL's performance for the Spanish somatisms *bajar los brazos* and *arrimar el hombro* into English and German.

	KWIC extracts
ST [ES]	Las dificultades del primero para iniciar el juego colaboraron en alguno de los goles rivales; el segundo trato de dar coherencia al juego de un equipo horroroso en la transición defensiva, hasta que bajó los brazos definitivamente.
	DeepL's outcomes
TT [EN]	The difficulties of the first one to start the game collaborated in some of the rival goals; the second one tried to give coherence to the game of a horrendous team in the defensive transition, until it gave up the arms definitively.
TT [DE]	Die Schwierigkeiten des ersten, das Spiel zu beginnen, wirkten bei einigen der rivalisierenden Tore mit; der zweite versuchte, dem Spiel einer horrenden Mannschaft in der defensiven Übergangsphase Kohärenz zu verleihen, bis er die Waffen endgültig abgab.

Table 3. Instances of DeepL mistranslations in English and German for the somatism *bajar los brazos*

	KWIC extracts
ST [ES]	Por esta razón sólo cabía la posibilidad de arrimar el hombro un poco y realizar las aportaciones y modificaciones económicas necesarias, para conseguir una plaza de toros más viable.
	DeepL's outcomes
TT [EN]	For this reason, there was only the possibility of putting the shoulder to the wheel a little and making the necessary contributions and economic modifications, in order to achieve a more viable bullring.
TT [DE]	Aus diesem Grund gab es nur die Möglichkeit, ein wenig die Schulter ans Steuer zu legen und die notwendigen Beiträge und wirtschaftlichen Änderungen vorzunehmen, um eine lebensfähigere Stierkampfarena zu erreichen.

Table 4. Instance of DeepL mistranslation in German for the somatism *arrimar el hombro*

In Table 3 it is possible to observe that, in the ES>DE directionality, *bajar los brazos* has been translated as *die Waffen angeben* ('to give up the weapons'). The only way to understand what yielded this unpredictable outcome is to analyse the English target text for the source-text (ST) somatism in the ES>EN directionality: *to give up the arms*. Therefore, in the ES>DE scenario, the German version was mostly determined by the training data with English with a misinterpretation of *the arms* as *die Waffe* ('the weapons') instead of *die Arme* ('the arms' as body parts). Analogous mistranslations

can be observed when examining DeepL's outcomes in other language directionalities for this Spanish ST idiom: *abandoner les armes* ('to abandon the weapons') in French, and *cedere le armi* ('to give in the weapons') in Italian. In Table 4, a similar problem can be detected. In this case, while in ES>EN an appropriate equivalent for the ST somatism *arrimar el hombro* has been offered (*to put the shoulder to the wheel*), in the ES>DE scenario it is possible to detect the sequence *die Schulter ans Steuer legen* ('to put the shoulders on the [steering] wheel'), with no idiomatic meaning. Once again, similar mistranslations with no idiomatic readings are to be found in other language directionalities for this Spanish ST somatism: *mettre l'épaule à la roue* in French, *mettere la spalla alla ruota* in Italian, or *colocar o ombro na roda* in Portuguese. These mistranslations hence emphasise the necessity for more training data in language combinations different from English, in order to avoid English-centred NMT outcomes.

Conclusion

The findings of our study confirm our hypothesis: the system *gApp* can, on average, improve the quality of the neural machine translation of discontinuous MWEs by converting them into their continuous form. More specifically, *gApp* has proved to enhance NMT for the analysed MWEs with a final average amelioration by 14.6%, which is only a 0.5% lower than the gold standard (15.1%).

These promising results with VNC and VPP somatisms in different language directionalities invite to further increase *gApp*'s detection lexicon and conversion mechanism so as to evaluate to what extent it can also result in NMT enhancement for other discontinuous MWE categories. In addition, the present study can also constitute the basis for further research to assess the scalation of this model to other language-dependent text preprocessing systems for the automatic conversion of discontinuous MWEs in syntactically flexible languages, with the purpose of enhancing MWE-aware NMT systems.

Acknowledgements

This research has been carried out within the framework of several research projects (ref. PID2020-112818GB-I00, UMA18-FEDERJA-067, P20-00109, E3/04/21, UMA-CEIATECH-04 and 03/2021-Embassy of France in Spain) at Universidad de Málaga (Spain) and at Research Institute of Multilingual Language Technologies (IUITLM). It has also been funded by a post-doc grant entitled *Ayuda para la recualificación del Sistema Universitario Español 2021-2023 (Modalidad «Margarita Salas»)* at Universidad de Málaga and at Université catholique de Louvain (Belgium).

References

Bargmann, Sascha and Manfred Sailer. 2018. The syntactic flexibility of semantically non-decomposable idioms. In Manfred Sailer and Sascha Markantonatou (Eds.), *Multiword expressions: Insights from a multi-lingual perspective*, pages 1–29. Language Science Press.

Belinkov, Yonatan, and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. *ArXiv*. https://arxiv.org/abs/1711.02173

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. *arXiv*. preprint arXiv:1608.04631.

Colson, Jean-Pierre. 2019. Multi-word Units in Machine Translation: why the Tip of the Iceberg Remains Problematic – and a Tentative Corpus-driven Solution. *MUMTT2019*.

Corpas Pastor, Gloria 2013. Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas. In Inés Olza and Elvira Manero (Eds.), *Fraseopragmática* (pages 335-373). Frank & Timme.

Foufi, Vasiliki, Luca Nerima and Eric Wehrli. 2019. Multilingual parsing and MWE detection. In Yannick Parmentier and Jakub Waszczuk (Eds.), *Representation and parsing of multiword expressions: Current trends* (pages 217–237). Language Science Press.

Hidalgo-Ternero, Carlos Manuel. 2020. Google Translate vs. DeepL: analysing neural machine translation performance under the challenge of phraseological variation. *MonTI. Monografías de Traducción e Interpretación, Special Issue 6*, 154-177. https://doi.org/10.6035/MonTI.2020.ne6.5

Hidalgo-Ternero, Carlos Manuel. 2021. El algoritmo ReGap para la mejora de la traducción automática neuronal de expresiones pluriverbales discontinuas (FR>EN/ES). In Gloria Corpas Pastor, María Rosario Bautista Zambrana and Carlos Manuel Hidalgo-Ternero (Eds.), *Sistemas fraseológicos en contraste: enfoques computacionales y de corpus*. Comares (pages 253-270)

Hidalgo-Ternero, Carlos Manuel. 2022/forthcoming. A la cabeza de la traducción automática neuronal asistida por gApp: somatismos en VIP, DeepL y Google Translate. In Gloria Corpas Pastor and Míriam Seghiri (Eds.). Comares.

Hidalgo-Ternero, Carlos Manuel, and Gloria Corpas Pastor. 2020. Bridging the 'gApp': improving neural machine translation systems for multiword expression detection. *Yearbook of Phraseology*, *11*(1), 61–80. https://doi.org/10.1515/phras-2020-0005

Hidalgo-Ternero, Carlos Manuel, and Gloria Corpas Pastor. 2022a/forthcoming. ReGap: a text preprocessing algorithm to enhance MWE-aware neural machine translation systems. In Johanna Monti, Gloria Corpas Pastor and Ruslan Mitkov (Eds.), *Recent Advances in MWU in Machine Translation and Translation technology*. John Benjamins Publishing Company.

Hidalgo-Ternero, Carlos Manuel and Gloria Corpas Pastor. 2022b/forthcoming. Qué se traerá gApp entre manos… O cómo mejorar la traducción automática neuronal de variantes somáticas (ES>EN/DE/FR/IT/PT). In Míriam Seghiri and Míriam Pérez Carrasco (Eds.). *Aproximación a la traducción especializada*. Peter Lang.

Hidalgo-Ternero, Carlos Manuel, Francesco Lista, and Gloria Corpas Pastor. 2022/under review. gApp-assisted NMT: how to improve the neural machine translation of discontinuous multiword expressions (IT>EN/DE). *Language Resources and Evaluation*.

Hidalgo-Ternero, Carlos Manuel, and Xiaoqing Zhou-Lian. 2022a. Reassessing gApp: Does MWE Discontinuity Always Pose a Challenge to Neural Machine Translation?. In Gloria Corpas Pastor and Ruslan Mitkov (eds) *Computational and Corpus-Based Phraseology. EUROPHRAS 2022. Lecture Notes in Computer Science*, vol 13528. Springer, Cham. https://doi.org/10.1007/978-3-031-15925-1_9

Hidalgo-Ternero, Carlos Manuel, and Xiaoqing Zhou-Lian. 2022b/under review. Minding the gApp in the ES>EN/ZH neural machine translation of discontinuous multiword expressions. *Natural Language Engineering*

Junczys-Dowmunt, Marcin, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? A case study on 30 translation directions. *Arxiv*. https://arxiv.org/pdf/1610.01108.pdf

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The Sketch Engine. *Proceedings of the 11th EURALEX International Congress*, pages 105-116.

Lohar, Pintu, Maja Popović, Haithem Alfi, and Andy Way. 2019. A systematic comparison between SMT and NMT on translating user-generated content. *20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2019)*.

Mellado Blanco, Carmen. 2004. *Fraseologismos somáticos del alemán*. Peter Lang, Frankfurt.

Monti, Johanna, Violeta Seretan, Gloria Corpas Pastor and Ruslan Mitkov. 2018. Multiword units in machine translation and technology. In R. Mitkov, J. Monti, G. Corpas Pastor & V. Seretan (Eds.), *Multiword Units in Translation and Translation Technology*, pages 1-37. John Benjamins.

Parra Escartín, Carla, Almudena Nevado Llopis, and Eoghan Sánchez Martínez. 2018. Spanish multiword expressions: Looking for a taxonomy. In Manfred Sailer and Stella Markantonatou (eds.), *Multiword expressions: Insights from a multi-lingual perspective*, 271–323. Language Science Press.

Ramisch, Carlos and Aline Villavicencio. 2018. Computational treatment of multiword expressions. In Ruslan Mitkov (Ed.), Oxford Handbook on Computational Linguistics (2ª ed). https://doi.org/10.1093/oxfordhb/9780199573691.013.56

Rohanian, Omid, Shiva Taslimipoor, Samaneh Kouchaki, Le An Ha, and Ruslan Mitkov. 2019. Bridging the Gap: Attending to Discontinuity in Identification of Multiword Expressions. In Jill Burstein, Christy Doran, and Thamar Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1*. (pages 2692–2698). Association for Computational Linguistics.

Shterionov, Dimitar, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'Dowd, and Andy Way. 2018. Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation*, 32, 217–235. https://doi.org/10.1007/s10590-018-9220-z

Zaninello, Andrea and Alexandra Birch. 2020. Multiword expression aware neural machine translation. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 3816–3825.

The use of CAT tools and corpus analysis in comparative literary translation research: an English-Arabic case study

Amal Haddad Haddad
Universidad de Granada
amalhaddad@ugr.es

Abstract

Omission is one of the common techniques used in translation. It is considered as a solution in cases of non-equivalence or implicit conveyance of meaning, and a parameter to detect manipulation and censorship in other cases. In the field of translation studies, many researchers compare original work with its translation by reading and manually annotating the original work and its translation and/or translations. This task is time consuming and in occasions may lead to the loss of relevant information due to inaccuracy, above all in long and extensive texts, such as some literary texts. In this paper, we propose an automated mechanism to detect omissions in literary translated works with the help of the Computer-Assisted Translation Tool (CAT), Trados Studio 2021, combined with parallel corpus analysis to compare the translation techniques used. As results of this study, we recommend the use of Trados Studio 2021 in conducting research related to comparative translation as it saves time and effort. In our opinion, the automatic detection of omissions is considered more precise than manual annotation and analysis when the texts, subjects of study are extensively long.

Keywords: computer-assisted literary translation; corpus analysis; close reading; distant reading; *Daddy-Long-Legs*.

Introduction

Omission and manipulation are considered two common translation techniques (Rodica, 2004: 163; Baker 2011). Omission is considered as a solution in cases of non-equivalence or implicit conveyance of meaning (Baker, 2011) and a parameter to detect manipulation and censorship in other cases (Klimovich, 2016). Baker and Saldanha (2009: 4) define omission as "the elimination or implicitation of part of the text" while other researchers define it as "translation loss" (Dickins *et al.*, 2017). On the other hand, manipulation in translation may involve "changes of individual lexical items or larger scale alterations such as the restructuring of the text and removal of significant sections which […] often have an ideological motivation and seek to affect the image of the text in the receiving culture" (Sherry, 2010: 3).

One of the most common approaches in translation studies research is to compare the original texts with their translation and/or translations to comprehend which techniques have been used and the reason behind choosing them during the translation process. This approach is called close reading (Youdale, 2019). Some of the techniques which investigators try to detect are omission and manipulation. When conducting comparative literary translation research, detecting omission and manipulation is very important, as they reveal the different preferences of different translators towards the target texts and in some cases, give information about their sociocultural and ideological tendencies.

Omission and manipulation techniques are also relevant in the cases of research related to translation and censorship. Some researchers consider intentional omission as a direct strategy of censorship (Klimovich, 2016; Cámara, 2016; Alimen and Kalaycio, 2021), while others consider censorship as "a manipulative rewriting of discourses" (Sherry, 2010: 1). Baker and Saldanha (2009: 289) define censorship as "a coercive and forceful act that blocks, manipulates and controls cross-cultural interaction in various ways". Other authors such as Izwaini (2017: 47) define censorship as: "a legal, administrative and socio-economic practice based on laws, rules, directives, guidelines, instructions, criteria and attitudes that has a direct impact on translation as an activity as well as on translators".

In the case of translation of Children and Youth Literature (CYL), the signs of manipulation are usually more noticeable as "both the target culture and society may decide what is wrong and what is acceptable for their children" (Leonardi, 2020: 26-27). That is why according to Giugliano and Hernández (2019: 314), CYL is considered an "ideal field for research on issues related to censorship" as they belong to both literary and educational field (Shavit, 1994: 11).

In the majority of research carried out until now to compare original texts with their translations, researchers read and manually annotate the original work and its translation(s). This task is time consuming, labour intensive, and on many occasions could lead to the loss of relevant information due to inaccuracy, mainly in long and voluminous texts, such as literary texts.

For this reason, in this paper, we propose the use of a combination of the close and distant reading approaches (Youdale, 2019), which implies the use of new technologies such as CAT tools and corpus analysis tools to analyse texts to acquire new insights into more comprehensive results in translation studies. With this objective, we suggest an automatic approach to detecting omissions in literary translated works with the help of the CAT tool Trados Studio 2021, combined with the use of corpus analysis to detect manipulation.

Trados Studio 2021, is one of the world's most widely-used CAT tools, both by autonomous translators and translation agencies (Zhang and Cai, 2015: 430), nevertheless, it still has an untapped potential. This software was designed to help translators during the translation process, by generating reusable Translation Memories (TM) that make translators benefit as much as possible from previous translations

(Mitkov, 2022: 367); though it can also be used in research tasks as well. To show the feasibility of this last facet of Trados Studio 2021, as part of a case study, we propose a methodology to detect omissions in the translations of the epistolary novel *Daddy-Long-Legs* written originally by Jean Webster in 1912 and compared to two of its translations into Arabic, published in 2009 and 2018.

Computer assisted tools

- **Trados Studio 2021**

Trados Studio 2021[1] is a CAT tool and an Artificial Intelligence (AI) application which offers a complete centralised translation environment for editing, reviewing and managing translation projects and terminology. It offers many features designed to help increasing the efficiency of translation processes and workflow and it is considered as an excellent solution to improve consistency. Its use is also helpful as it has the advantage of retaining the original text format when translating (Huynh: 183). It is built on an open platform and uses the technology of bilingual file, translation memory and termbase formats (Trados Studio). Authors like Ike and D'Angelo (2020: 190) praise the high grade of accuracy achieved owing to the provided help at phrase level and with the help of TMs. Trados Studio currently has a new version called Trados Studio 2022, which offers major synchronisation options and supports more file types; however, for the elaboration of this study, we used the previous version Trados Studio 2021.

In spite of the fact that Trados Studio 2021 is originally designed to help translation agencies and professional translators in managing the workflow and in producing more efficient translations, in this case-study, it has been observed that the alignment options offered by Trados Studio 2021 may be helpful for other tasks as well, and can be useful for researchers in the field of translation to conduct comparative research. This new method will be discussed. Based on the provided results we propose using this method as a base to develop a new *ad hoc* tool to detect omission.

- **Sketch Engine**

Sketch Engine (SE) is a corpus analysis tool (Vojtěch, 2016). It is one of the widely used AI applications for terminology management tasks. It allows the creation of comparable as well as parallel corpora and offers different options to analyse concordances and frequencies of words. For this reason, this tool is also used as didactic tool for the formation of professional translators and linguists (Matvieieva, 2022; Faya-Ornia, 2022, Gorbunov, 2020) and in comparative translation research (Perak and Kirigin, 2021). In this research, in line with the preliminary results of Le Poder (2021), we used SE to provide more insightful results related to comparative literary translation research. We believe that also SE is an excellent tool for distant reading, and more research is still required to unveil all its potential for comparative literary research.

[1] Available from: https://www.trados.com/

Materials

As a case-study, we carried out comparative research on the epistolary novel, *Daddy-Long-Legs* and two of its translations into Arabic. *Daddy-Long-Legs* was written originally by the American writer Jean Webster in 1912 and is considered one of the symbols of American national identity (Phillips, 1999: 79). This novel remains an international success, as it is regularly reedited and retranslated into different languages, and it has been adapted into stage and into screen.

Daddy-Long-Legs is classified as a youth literature novel (Guadamillas, 2019: 204; Hermida *et al.*, 2020: 10). It narrates the story of Jerusha Abbott (Judy), a girl who was brought up at the orphanage of John Grier Home until she was 18. One of the trustees heard about her talent in writing and promised to finance her studies at college to become a writer, with the condition of receiving a monthly letter from her, describing her advances in her career and education. The trustee did not want to reveal his identity and said that he will never reply to her letters. On the day the trustee left the orphanage, Judy noticed only a glimpse of his shadow in the dark projected on a wall, and she starts calling him mockingly Daddy-Long-Legs, hence the title of the novel. When Judy started her life at college, she not only started writing one letter a month, but she used to send letters on weekly or daily basis describing all the details of her daily life. Through those letters, the educational, cultural, emotional, social and ideological growth of Judy are made tangible and visible.

Previous research has been carried out to compare *Daddy-Long-Legs* with its corresponding translations into different languages. For example, Sharifi and Karimnia (2014) analysed the translation of the translated book in comparison to the film dubbing in Persian language, by using the critical discourse analysis approach. Rahbar *et al.* (2013) identified the ideological content of the novel and study the dimension of censorship in the translations of the novel published in Iran, before and after the Islamic Revolution. Other authors such as Alimen and Kalaycioğlu (2021) compare two translations of the novel into Turkish adapted to children. However, to my knowledge, no studies have been carried out on Arabic translations of the novel and none of the studies implemented new technologies to compare results.

For this reason, in this case-study, we compare two translations of *Daddy-Long-Legs* into Arabic: the first translation entitled *"أبي طويل الساقين" "aby ṭawyl assāqayn*[2]*"* [Daddy Long Legs[3]] was carried out by Samir Mahfouz Bashir, published by the National Center for Translation in Egypt in 2009. The second translation, entitled *"صاحب الظل الطويل"* "*ṣāḥib aẓẓill aṭṭawyl*" was by Buthaina Al-Ibrahim, published by Takween in Kuwait in 2018.

[2] Transliteration of Arabic text is provided between quotations.

[3] Literal translation of the Arabic text is provided between square brackets.

Methodology

For the implementation of this case study, first of all, the three versions of the novel were converted into an editable format, i.e. the original English and the two translations into Arabic. Afterwards, the texts were inserted in the CAT tool program, Trados Studio 2021 with the objective of aligning the translations with the original text in English to create a parallel corpus. The automatic alignment offered by Trados Studio 2021 was revised by adjusting the segments and realigning them when needed. The alignment was applied at phrase and paragraph level. The split segment option was used when part of the sentence was omitted so that the unaligned segments would contain only the parts that were totally omitted. The untranslated segments were left unaligned.

When the two translations were adequately aligned, we used the option of identifying all unaligned segments available in Trados Studio 2021. The functionality in Trados Studio 2021 appears in the alignment window as shown in Figure 1.

Figure 1. Alignment window in Trados Studio 2021

From that window, the functionality of *Select the alignment status, quality or connection type to go to* was selected as can be seen in Figure 2. Afterwards, the unconfirmed segments option was selected, as can be observed in Figure 3.

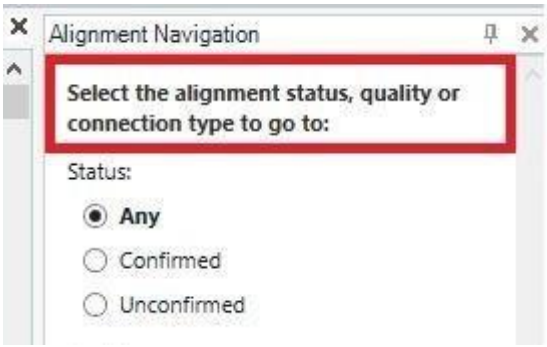


Figure 2. Select the alignment status, quality or connection type to go to

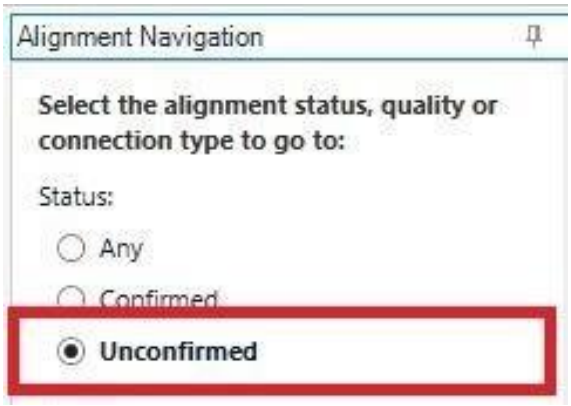


Figure 3. Select the unconfirmed segments option

This way all the segments that were not translated in the two translations into Arabic were identified. This was the technique to identify omissions of segments. The aligned texts were inserted then in the Sketch Engine[4] corpus analysis tool with the aim of comparing the original text and the two translations in Arabic. In other words, a bidirectional parallel corpus analysis was possible, using both a top down and bottom up strategy. Top down analysis refers to the close reading and examining the pre-annotated examples within its context, while bottom up analysis refers to corpus analysis, by examining the corpus with the help of the corpus analysis functions available in SE, such as wordlists, frequency, etc (Biber *et al*., 2007: 12-16).

Finally, the concordances functionality in Sketch Engine and frequency list were used, to compare the techniques of translation, above all, comparing segments in the two target texts, where it was estimated that there were signs of manipulation or censorship. In this phase, the study of Rahbar *et al.* (2013) in which the authors identified the ideological content of the novel was used to facilitate the process.

Analysis of results

At first, the paragraphs and sentences which were omitted were identified by highlighting the unaligned segments with the help of Trados Studio 2021. The result of this process indicated that the translation of Samir Mahfouz Bashir, to which we will refer in this study as T1, had 62 instances of omissions, some of which were sentences and others whole paragraphs. On the other hand, the translation by Buthaina Al-Ibrahim, to which we will refer to as T2 in this study, had no omissions at the sentence and paragraph level.

Secondly, we classified the reasons for omissions in T1 into five *ad hoc* categories, based on our background knowledge of sociocultural aspects, common in the Arab world in general: a) omissions related to unacceptable social behaviour, above all, related to relations between men and women; b) omissions related to religious

[4] Available from: http://www.sketchengine.eu

information; c) omissions related to ideological references; d) omissions related to unacceptable moral conduct; e) omissions due to linguistic reasons. Table 1 shows the frequency of omissions associated to each category.

Motive of omission	Frequency
unacceptable social behaviour	23
religious information	5
ideological references	4
unaccepted moral conduct	22
linguistic reasons	8
Total	62

Table 1. Frequency of omission and motive of omission in T1

After carrying out the analysis, it has been observed that social behaviour and moral conduct are the most frequent reasons of omissions in the translation, with a frequency of 23 and 22 occurrences consecutively. In the following, examples on each motive will be given.

1. Jimmie McBride is going to teach me how to ride horseback and paddle a canoe, and how to shoot and--oh, lots of things I ought to know. It's the kind of nice, jolly, carefree time that I've never had; and I think every girl deserves it once in her life (p. 51).

2. I didn't know that people used to be monkeys and that the Garden of Eden was a beautiful myth (p. 14).

3. Seems a little early to commence entertaining, doesn't it? A friend of Pepys devised a very cunning manner whereby the king might pay his debts out of the sale to poor people of old decayed provisions. What do you, a reformer, think of that? I don't believe we're so bad today as the newspapers make out (p. 82).

4. Oh, you see, I know! You're a snappy old thing with a temper (p. 14). 5. I'd hate to retouner chez John Grier (p. 41)

Example (1) shows a case of omission of a whole paragraph, as it describes a situation in which a man and a woman would have a close relationship doing certain activities together, which are not accepted in certain cultures. In this case, Jimmie McBride is one of Judy's friends and the brother of her best friend Sally, and he invited Judy to spend the summer with him to teach her certain activities, like horse riding and paddle a canoe. These types of omissions were frequent also in scenes where Judy and Jervis Pendleton were together. Example (2) shows a case of omission in order to hide information that is considered contradictory to the teachings of religion; in this particular example, the omission was implemented to avoid telling information relevant to stating that the

origin of people is monkeys, and questioning the veracity of existence of the Garden of Eden. Nevertheless, it has also been observed that there were omissions were Christian teachings were praised, while all the sentences which criticise it were preserved. In the case of example (3), it shows omission due to ideological contents. In this example, the word "reformer" is the clue. The translator estimated that he should not include this type of information related to a particular political trend. With respect to example (4), it shows a disrespectful behaviour on the part of the main character towards the trustee, who she supposes is an old man. For this reason, the way she picked her words is considered disrespectful. Finally, in cases like example (5) the linguistic difficulty which led to the use of omission was due to mixing English with French which the translator preferred to omit.

After detecting and analysing all the cases of omissions in T1, we compared the T1 and T2 by using the *parallel concordances* option available in Sketch Engine, to analyse which techniques the translator T2 used in the parts of the novel that were omitted in T1. We could also analyse the techniques that the two translators used, where Rahbar *et al.* (2013) identified ideological content. Table 2 shows some of the words or phrases that went through alteration of meaning techniques, comparing the original in English with T1 and T2.

Original text	T1	T2
anarchists	المحافظين *almuḥāfiẓyn* *[The conservatives]*	الفوضويين *alfawḍawyīn* *[The anarchistss]*
plutocrat	رأسمالية *Ra'smālyīa* *[Capitalism]*	بلوتوقراطية *blūtwqraṭya* *[Plutocracy]*
Yours ever	المطيعة دائما *almuṭy'a da'iman* *[The always obedient]*	المخلصة لك أبدا *almukhliṣa laka abadan* *[the always failthful]*

Table 2. Manipulated words or phrases in T1 and T2 in comparison with the original text

Table 2 shows some of the examples where the techniques of meaning alteration were used either in T1 or in T2. With respect to the word "anarchist" which appeared in the following context: "You know, I think I'll be a Socialist, too. You wouldn't mind, would you, Daddy? They're quite different from Anarchists; they don't believe in blowing people up" (p. 67), it is observed that the translator in T1 decided to change the meaning of the word into *"المحافظين" "almuḥāfiẓyn"* which means "the conservatives" while the translator in T2 used the word *"الفوضويين" "alfawḍawyīn"* which is a literal translation. In the same way, the translator in T1, substituted the word "plutocrat" by the word *"رأسمالية" "Ra'smālyīa"* which means "capitalism", while in T2, the translator used the literal translation *"بلوتوقراطية" "blūtwqraṭya"*.

Those words appeared in the following context: “I'm a Socialist, please remember; do you wish to turn me into a Plutocrat?” (128). Finally, in T1, the translator changed the way in which the main character, Judy, finishes her letters, by substituting the phrase “yours ever” with *“المطيعة دائما”* *“almuṭy‘a da’iman”* which means “the always obedient”, while the translator in T2 uses a more literal translation.

As shown in the previous examples, both translations of *Daddy-Long-Legs* went through omission and/or manipulation techniques. However, T1 contains obvious examples on omission and alteration of meaning for ideological reasons. In this case, the use of the automatic detection of omission facilitated identified more easily that the translation may have evidence of censorship and manipulation.

Conclusions

In this case study, we aim to contribute to the new field of computer-assisted literary translation (Youdale and Rothwell, 2022: 384). With this objective we propose the use of new technologies in literary translation research, such as the CAT tool Trados Studio 2021, as it allows for the automatic detection of omission after the proper alignment of the original text with its corresponding translation and/or translations. The use of corpus analysis methodology to analyse further translation techniques, such as alteration of meaning is also proposed. Those translation techniques are considered especially relevant to censorship studies. For this reason, we recommend using corpus analysis tools such as Sketch Engine, as it allows comparing word lists and its frequency as well as using the parallel concordance function to see how certain words or sentences were used in parallel bilingual corpora. The use of those methods in combination with the close reading approach would give more insights and a wider perspective towards original and target texts.

Finally, and notwithstanding that Trados Studio 2021 helped in the automatic detection of omissions in this study and saved time, it is still not optimum for this task. For this reason, we believe that there is a need to design more professional software to help translators and researchers in detect omissions in a more straightforward way. We also highlight the need to carry out more research orientated towards the understanding of the real needs of researchers, and accordingly design more AI tools which may help them in the tasks related to comparative literary translation.

Acknowledgements

I would like to express my gratitude to Prof. Ruslan Mitkov and RGCL Research Group.

References

Alimen, Nilüfer and Kalaycioğlu, Esra. 2021. An Analysis of the Turkish translations of Jean Webster's DaddyLong-Legs from the perspective of systemic affiliation. *TransLogos,* 4(2), 70 – 97. https://doi.org/10.29228/transLogos.38

Baker, Mona (2011). *In other words: A Coursebook on translation*. Routledge.

Baker, Mona and Saldanha, Gabriela. 2009. (2nd ed.). *Routledge Encyclopedia of translation studies*. Routledge.

Biber, Douglas, conner, Ulla, Upton, Thomas A. (2007) *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure.* John Benjamis.

Cámara Aguilera, Elvira. 2016. Traducción y asimetriá: "Manolito Gafotas" y su traducción al inglés como ejemplo de intervencionismo. *Anuario de investigación en literatura infantil y juvenil: ANILIJ*, 14, 23-42.http://revistas.webs.uvigo.es/index.php/AILIJ/article/view/971

Dickins, James, Hervey, Sándor and Higgins, Ian. 2017 *Thinking Arabic translation* (2nd ed.). Routledge.

Faya-Ornia, Goretti, Barranco-Izquierdo, Natalia, Calderón-Quindós, Teresa, & Quijada-Diez, Carmen. 2022. Subtitle creation and subtitle alignment as didactic resources for foreign language teaching. *Culture and Education,* 34(1), 180-187.

Giugliano, Marcello and Hernández Socas, Elia. 2019. Ambivalence, Gender, and Censorship in two Spanish Translations of Little Women. Meta, 64(2), 312–333. https://doi.org/10.7202/1068197ar

Gorbunov, Yurij,, Gorbunov, Evgenij, and Mkrtychev, Sergei. (2020). The Informational and Semiotic Nature of French Socio-Political Terminology and its Thesaurus Modeling. In CEUR Workshop Proceedings (pp. 211221).

Hermida, Carola, Couso, Lucía Belén, Bayerque and María Ayelén. 2020. Jóvenes y literatura: cruces entre el campo editorial y escolar. In proceedings of *2nd Congreso Latinoamericano de Comunicación de la UNVM. Instituto Académico Pedagógico de Ciencias Sociales*. Universidad Nacional Villa María, Villa María, Córdoba.

Huynh Tan Hoi. (2020). Efficiency of Japanese-Vietnamese Translation Job Thanks to the Use of Technology in the Fourth Industrial Revolution. In Proceedings of the 3rd International Conference on Digital Technology in Education (ICDTE '19). Association for Computing Machinery, New York, NY, USA, 181–184. https://doi.org/10.1145/3369199.3369228

Ike, Saya, and James D'Angelo. (2020). "English in Japan: The Applicability of the EIF Model." Chapter. In *Modelling World Englishes: A Joint Approach to Postcolonial and Non-Postcolonial Varieties,* edited by Sarah Buschfeld and Alexander Kautzsch, 179–201. Edinburgh University Press.

Izwaini, Sattar. 2017. Censorship and manipulation of subtitling in the Arab world. In Jorge Díaz Cintas y Kristijan Nikolić, editors-in-chief, Fast-Forwarding with Audiovisual Translation. Multilingual Matters, 47-57. DOI: 10.21832/9781783099375-006

Klimovich, Natalya. 2016. Manipulative Strategies in the Translations of Literary Texts Carried Out in the Soviet Union. Journal of Siberian Federal University. *Humanities & Social Sciences,* 3 (9) 543-550.

Leonardi, Vanessa. 2020. *Ideological manipulation of children's literature through translation and rewriting: travelling across times and places.* Palgrave Macmillan. https://doi.org/10.1007/978-3-030-47749-3

Le Poder, Marie-Évelyne. 2021. Traducción, censura y género: un análisis de cómics para adolescentes traducidos del francés al español. In proceedings of the CIUTI 2021 Conference. Granada.

Matvieieva, Svitlana. A., Lemish, Nataliya Y., Zernetska, Alla, A., Babych, Volodymyr. O., & Torgovets, Maryna. A. 2022. English-Ukrainian Parallel Corpus: Prerequisites for Building and Practical Use in Translation Studies. Studies about Languages, (40), 61-74.

Mitkov, Ruslan.2022. Translation Memory. In Sharon Deane-Cox and Anneleen Spiessens, editors-in-chief, *The Routledge Handbook of Translation and Memory.* Basingstoke: Routledge.

Perak, Benedikt, and Tajana Ban Kirigin. 2021. Dependency-based labeling of associative lexical communities. Central European Conference on Information and Intelligent Systems: 35-42.

Rahbar, Muhamad, Bateni, Bijan and Abad-Najaf, Ranjbar. 2013. Ideological manipulation in translation: A case study of Jean Webster's "Daddy Long Legs". *International Journal of Language Learning and Applied Linguistics World (IJLLALW)* 4(4), 373-381.

Rodica Dimitriu. 2004. Omission in translation, Perspectives: Studies in Translatology. *Studies in Translation Theory and Practice* 12 (3), 163-175, DOI: 10.1080/0907676X.2004.9961499

Sharifi, Leila and Karimnia, Amin. (2014). Differences between Webster's Daddy-Long-Legs Translation for Publication and Animation: Insight From Van Dijk's Ideology Framework of Critical Discourse Analysis. *Modern Journal of Language Teaching Methods (MJLTM)* 4(4), 163-169 http://mjltm.org/browse.php?mag_id=16&&slc_lang=en&sid=1

Shavit, Zohar. 1994. Beyond the Restrictive Frameworks of the Past: Semiotics of Children's Literature — A New Perspective for the Study of the Field. In Hans-Heino Ewers, Gertrud Lehnert, Emer O'Sullivan, editors-in-chief, *Kinderliteratur im interkulturellen Prozeß*, 1-6. DOI: 10.1007/978-3-476-03522-6_1 Webster, Jean. 2003. *Daddy-Long-Leg.* Project Gutenberg: Rainfield.

Sherry, Samantha. 2010. Censorship in Translation in the Soviet Union: The Manipulative Rewriting of Howard Fast's Novel The Passion of Sacco and Vanzetti, *Slavonica,* 16:1, 1-14, DOI: 10.1179/136174210X12639903087016

Trados Studio. Trados Studio 2022 FAQs. https://www.trados.com/products/trados-studio/faq.html

Vojtěch Kovář, Vít Baisa, Miloš Jakubíček. 2016. Sketch Engine for Bilingual Lexicography, International Journal of Lexicography, Volume 29, Issue 3, 339–352, https://doi.org/10.1093/ijl/ecw029

Youdale, Roy. 2019. Using Computers in the Translation of Literary Style: Challenges and Opportunities (1st ed.). Routledge. https://doi.org/10.4324/9780429030345

Youdale, Roy and Rothwell, Andrew. 2022. Computer-assisted translation (CAT) tools, translation memory, and literary translation. In Sharon Deane-Cox and Anneleen Spiessens, editors in chief, *The Routledge Handbook of Translation and Memory*. Routledge, London.

Zhang, Chengzhi and Cai Hui. 2015. On Technological Turn of Translation Studies: Evidences and Influences. *Journal of Language Teaching and Research,* 6 (2), 429-434. DOI: http://dx.do.i.org/10.17507/jltr.0602.25

HypoLexicon:

A Terminological Resource for Describing Hyponymic Information

Juan Carlos Gil-Berrozpe
Translation Centre for the Bodies of the European Union (CdT)
juan-carlos.gil_berrozpe@cdt.europa.eu

Abstract

Terminology has evolved from static and prescriptive theories to dynamic and cognitive approaches. Thanks to these modern approaches, there have been significant advances in the design and elaboration of terminological resources, resulting in the creation of tools such as terminological knowledge bases (TKBs). For instance, they can show how concepts are interrelated through different semantic relations. Of these relations, hyponymy is the most relevant to terminology work because it deals with concept categorization and term hierarchies. In this line, this paper analyzes the representation of hyponymy in terminology work and presents a new enhancement related to hyponymy for EcoLexicon, a TKB on environmental science. It is known as HypoLexicon and it is a stand-alone module for EcoLexicon in the shape of a terminological resource designed to describe and represent hyponymic information of environmental concepts. It includes definitional, relational, ontological and contextual information about specialized hypernyms and hyponyms. The materials and methods used for the creation of HypoLexicon are described, involving the EcoLexicon English Corpus, Sketch Engine, and Lexonomy. Finally, the use and features of HypoLexicon are shown by analyzing the terminological entry of a geological concept, SEDIMENT, whilst also explaining all the hyponymic elements displayed using its characteristic hierarchical structure.

Introduction

Terminology is the study of specialized language, namely, the terms and phrases used in scientific and technical domains. Though interpreted in different ways (Sager, 1994), Terminology is an interdisciplinary domain that includes not only linguistic but also extralinguistic aspects, such as elements of human perception and computational processes. Terminology arose from the need to unify concepts and terms in specialized subject fields in order to facilitate professional communication and knowledge transfer (Cabré, 2000).

Most Terminology theories have practical applications, such as encyclopedias, specialized dictionaries, knowledge bases and other terminological or translation resources (Faber, 2012), which are the flagship for their approach. These resources ideally display their information so that it can be easily retrieved and used by different user profiles (Sager, 1990). This practice-based facet of Terminology, aimed at

systematically describing and representing previously collected terminological data, is also often referred to as Terminography (Temmerman, 2000).

Terminology has evolved from static and prescriptive theories (Wüster, 1968, 1979) to dynamic and cognitive approaches (Cabré, 1999; Faber, 2009). Thanks to these modern approaches, there have been significant advances in the design and elaboration of terminological resources. Over the years, traditional paper-based glossaries and dictionaries have been gradually replaced by electronic or digital versions, which can also be easily updated and modified. In recent years, terminological knowledge bases (TKBs) have become an important linguistic resource, showing a wide range of linguistic and non-linguistic information through intuitive interfaces (Meyer *et al*., 1992).

An example of a modern TKB is EcoLexicon (Faber *et al*., 2016). It is a multidimensional and dynamic TKB on environmental science that provides conceptual, linguistic, phraseological, and multimodal data in each entry. EcoLexicon, apart from its ontological approach, is characterized by its visualization of conceptual networks, showing how concepts are interrelated through different semantic or conceptual relations – generic-specific, part-whole, and non-hierarchical relations. Of these relations, generic-specific or hyponymic relations are particularly relevant to terminology because they deal with concept categorization and term hierarchies (Murphy, 2006). For this reason, the description of concepts and terms can be greatly improved by highlighting their hyponymic information.

Hyponymy

Hyponymy is the conceptual or semantic relation between a hypernym (i.e., a term referring to a generic, superordinate or parent concept) and a hyponym (i.e., a term referring to a specific, subordinate or child concept). Accordingly, the hyponyms of a same hypernym that are located at the same hierarchical level are regarded as co-hyponyms (i.e., terms referring to sibling concepts). The inverse relation of hyponymy is hyperonymy. It is central to many models of the lexicon for the following reasons (Murphy, 2003): (i) its inference-invoking nature; (ii) its importance in definition; and (iii) its relevance to selectional restrictions in grammar.

Hyponymy is defined in terms of inclusion, but the content that is inherited is dependent on whether hyponymy is viewed in terms of extensions (i.e., the categories that the words refer to), or in terms of intensions or senses (i.e., the semantic content associated with the words). Following Murphy & Koskela's (2010) example of birds, from the extensional perspective the category BIRD includes all the members of the category SWAN. However, from an intensional perspective, the inclusion relation is reversed and thus the hyponymic sense includes the sense of the hypernym. This means that, if BIRD is defined as "a winged animal that lays eggs", then SWAN would include all of these characteristics plus a few others (e.g., having a long neck, being usually white). Since this property inheritance does not happen in reverse, hyponymy gives rise to transitivity or unilateral entailment, by which the hypernym entails the hyponym, but not vice versa (Murphy & Koskela, 2010).

Hyponymic relations tend to be represented in hierarchical or tree structures, which reveals their relevance towards conceptual organization. For instance, Murphy (2006) illustrates this kind of visual representation with a summarized version of the hyponymic relations in the lexical field of FRUIT (Figure 1).

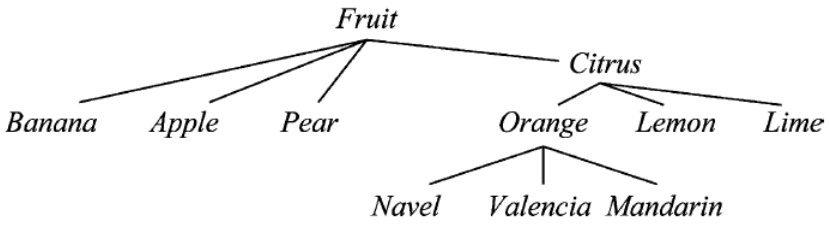


Figure 1. Hierarchy and hyponymic relations in the lexical field of FRUIT (Murphy, 2006)

The conceptual hierarchy and its hyponymic relations are asymmetrical. This means that any word may have many hyponyms, but in most cases, only one immediate hypernym. When a hyponym has more than one hypernym, it is multidimensional. However, this is more frequent when a concept belongs to different contextual domains.

For instance, a MORAINE is a type of SEDIMENT because of its composition, but it is also a type of LANDFORM because of its impact on the terrain. The example shown in Figure 1, however, represents a monodimensional conceptual system. In it, ORANGE has various hyponyms (i.e., NAVEL, VALENCIA, MANDARIN), but each of them has only one hypernym (ORANGE). ORANGE has only one immediate hypernym (CITRUS), which is then associated with the most general hypernym, FRUIT.

In relation to multidimensionality, phenomena that affect hyponymy are facets and microsenses (Cruse, 1995, 2002). Facets are dimensions or aspects of a concept that show a high degree of autonomy, and which make it possible to describe that concept from any of those perspectives. For instance, Cruse (2002) highlights two facets or dimensions in the hyponyms of BOOK, and divides them into two sets: physical object (e.g., HARDBACK, PAPERBACK) and abstract text (e.g., NOVEL, BIOGRAPHY). In these cases, the co-hyponyms of the same hypernym display within-set incompatibility, but between-set compatibility (a certain BOOK can be simultaneously a NOVEL and a HARDBACK, but a HARDBACK cannot be a PAPERBACK at the same time).

In contrast, a microsense is a specific meaning of a concept (i.e., regarding its properties, attributes or functions) which is only activated in a certain context. For example, although KNIFE generally has a single sense, it can be classified in different domains under a variety of hypernyms (WEAPON, TOOL, SURGICAL INSTRUMENT, etc.).

On the other hand, apart from the term itself, another essential element in any terminological resource is the definition. As the natural language explanation of the location of a concept in the conceptual structure of the specialized domain (Faber, 2022), definitions not only specify the properties of concepts, but also link them to other realities (Antia, 2000). Since the most basic way of associating concepts is by alluding to their hypernym, hyponymy is always present in all terminological resources with

intensional definitions. However, this indirect way of representing hyponymy does not fully exploit all its possibilities since it does not reflect all its complexity.

1.15 Hyponymy in Traditional Resources

Traditional terminological resources mainly include dictionaries and encyclopedias. Dictionaries list lexemes from the vocabulary or terminology of a language, or more languages in the case of bilingual and multilingual dictionaries. They often arrange this information alphabetically and include data regarding definitions, usage contexts, etymologies, pronunciations, and other elements. One of the most common classifications is the distinction between general language dictionaries and specialized language dictionaries, but here both types are reviewed from the perspective of terminology work. The emphasis is thus on terms instead of words.

General language dictionaries also include terms, but with very concise definitions. A good example of a general language dictionary is the *Oxford Dictionary of English* (Oxford University Press, 2010). Since the entries in this and similar dictionaries focus more on meanings, various definitions are displayed. The linguistic data in these resources include usage contexts, etymological (i.e., origin or historical development of the term), phonetic (i.e., pronunciation of the term), and collocational information (e.g., compound nouns with the term acting as subject). Furthermore, electronic versions of general dictionaries, such as *OED Online* (Oxford University Press, 2022), may even include multimedia information (e.g., audio files to check the pronunciation of the term according to different diatopic variants). However, there is no mention of hyponymy or any other conceptual relation. Neither is there any information regarding complex nominals or compound nouns whose head is the term given. This is another way to codify hyponymy by specifying characteristic attributes with hyponymic nuances (Gil-Berrozpe, 2020). Therefore, the only way to identify hyponymy in this type of resource is through the intensional definition displayed, which alludes to the hypernym of the concept (e.g., BACTERIUM – ORGANISM).

Specialized language dictionaries, such as *A Dictionary of Biology* (Hine, 2019), provide less linguistic information and focus on more detailed definitions of the terms. In contrast to general dictionaries, each entry has a single definition. Not surprisingly, the amount of specialized knowledge in these resources is greater than in general language dictionaries, and thus contain a wider range of more specific terms. Interestingly, there are certain specialized language dictionaries, such as the *Dictionary of Geology and Mineralogy* (McGraw-Hill, 2003), which also specify the subdomain (e.g., Geophysics, Paleontology, Mineralogy) to which the term belongs. This makes it possible to delimit the concept and to better differentiate its microsenses in relation to contextual domains. Nevertheless, yet again there is no explicit allusion to hyponymy, but it is indirectly present in intensional definitions, as well as in extensional explanations that allude to other interrelated concepts and which are so common in these resources.

On the other hand, encyclopedias do not have either a linguistic or a definitional approach, because their objective is to provide as much information (e.g., chronological, cultural, social, technical) as possible about a certain topic or domain.

Terminological information can be extracted from their entries, but it is more difficult to structure than in the case of dictionaries. Furthermore, they usually contain graphical information, not only pictures but also diagrams or flow charts, which facilitate comprehension and knowledge acquisition. For instance, in the *Encyclopedia of Biology* (Rittner & McCabe, 2004), it is possible to see that entries are very similar to those in specialized language dictionaries, since it begins with an intensional definition (e.g., "Bacteria are microscopic, simple, single-cell organisms") followed by an explanation of the relation of the concept with other entities and processes (e.g., AEROBIC DECOMPOSITION, COLONIES, GRAM'S STAIN, etc.). Encyclopedias thus combine both an intensional and an extensional description of the concept. In this case, hyponymy is again indirectly reflected only in the definition.

1.16 Hyponymy in Contemporary Resources

Contemporary terminological resources are digital or electronic tools such as term banks and TKBs. On the one hand, term banks provide direct access to terms as well as to their related linguistic information. Each entry may include data fields such as definition, correspondences in one or various languages, synonyms, abbreviations, status of each term (e.g., preferred, reliable, not recommended, etc.), usage contexts (and their corresponding references), and even the domain and subdomain to which the concept belongs. On the other hand, many TKBs go beyond term banks by implementing a wide range of features that enhance terminology. Such features include a dynamic knowledge representation, a graphic visualization of conceptual relations between concepts, and the integration of multimedia information, inter alia.

A paradigmatic example of a term bank is IATE[1] (Zorrilla-Agut & Fontenelle, 2019), the official terminology database of the European Union (EU), developed and supervised by the Translation Centre for the Bodies of the EU in collaboration with other European institutions. It is the largest multilingual term bank in the world, with around 900,000 concept entries and eight million terms in the 24 official languages of the EU. Within each of its entries, the following main elements are distinguished: (i) term; (ii) term reference; (iii) term reliability; (iv) definition; (v) definition reference; and (vi) creation and modification dates. These terminological entries can contain usage contexts and observation notes as well. However, since no mention is made of any kind of relation between terms or concepts, there is no direct way of accessing hyponymic information, as is also the case of traditional resources. IATE thus does not represent hyponymy. In fact, there are not even any hyperlinks within the entries that redirect users to other associated concepts. Therefore, once again, hyponymy representation is only indirectly reflected through intensional definitions of terms.

As an example of the second typology of contemporary resources indicated above, EcoLexicon[2] (Faber *et al*., 2014, 2016) is a multidimensional, multimodal, and dynamic TKB on the environment developed by the Lexicon Research Group of the University of Granada. To date, it has over 4,500 concepts and over 24,500 terms in seven different

[1] Available at: https://iate.europa.eu/

[2] Available at: http://ecolexicon.ugr.es/

languages: English, Spanish, German, French, Modern Greek, Dutch, and Russian. EcoLexicon has an intuitive visual interface with a series of modules that provide conceptual, linguistic, and multimodal information. Instead of viewing all this information simultaneously, users can browse and select the data that are most relevant to their needs. Once users select a concept, it is represented in the center of an interactive map representing its conceptual system. Also displayed are the multilingual terms for that concept, as well as different conceptual relations between all the concepts belonging to the same contextual domain (Figure 2).

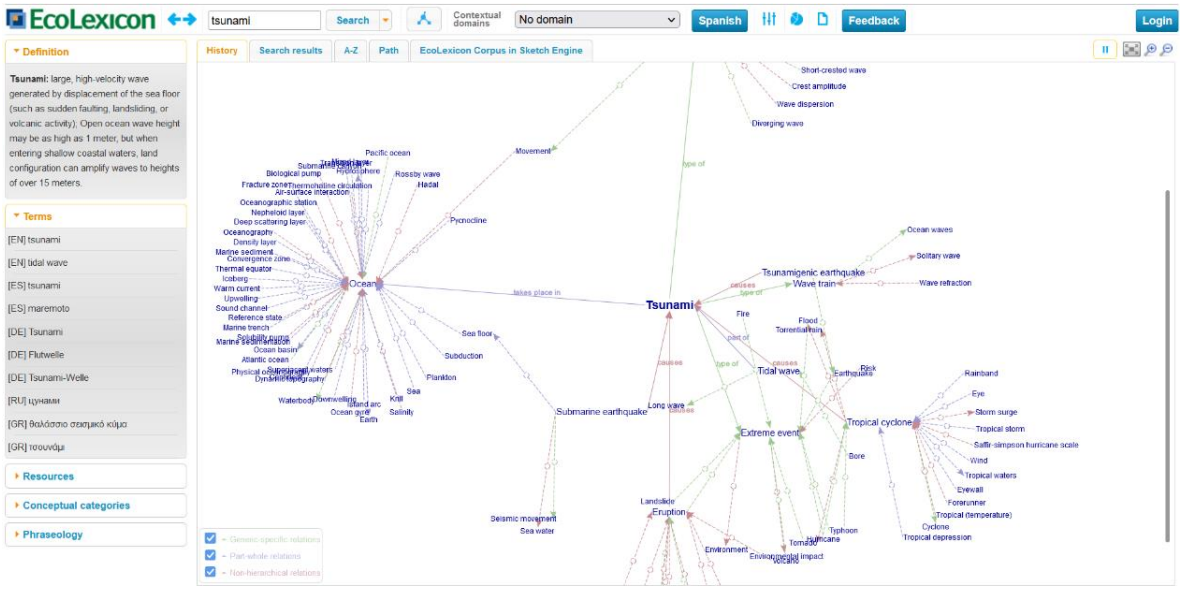


Figure 2. Visual interface of EcoLexicon (conceptual system: TSUNAMI)

Conceptual systems in EcoLexicon are cloud-like representations of the interrelations between a given concept and related concepts. Since EcoLexicon can represent up to 18 different types of conceptual relation, the visual map offers the possibility of filtering the relationships so that only those of interest are displayed with a view to reducing information overload. Figure 3 shows the conceptual system of BACTERIA in EcoLexicon represented with a single hyponymic relation (*type of*) as follows.

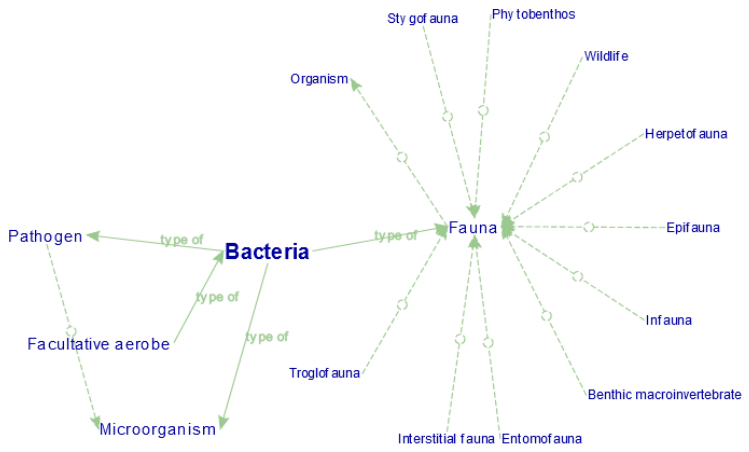


Figure 3. Conceptual system in EcoLexicon with a single hyponymic relation (*type of*)

However, these conceptual systems in cloud-like networks can be confusing when a large number of related concepts are involved. For this reason, EcoLexicon also offers the possibility of representing this information in conceptual hierarchies or tree-like representations with different levels of the hypernyms and hyponyms of a given concept. Figure 4 shows the conceptual system of ROCK in EcoLexicon represented with a single hyponymic relation (*type of*).

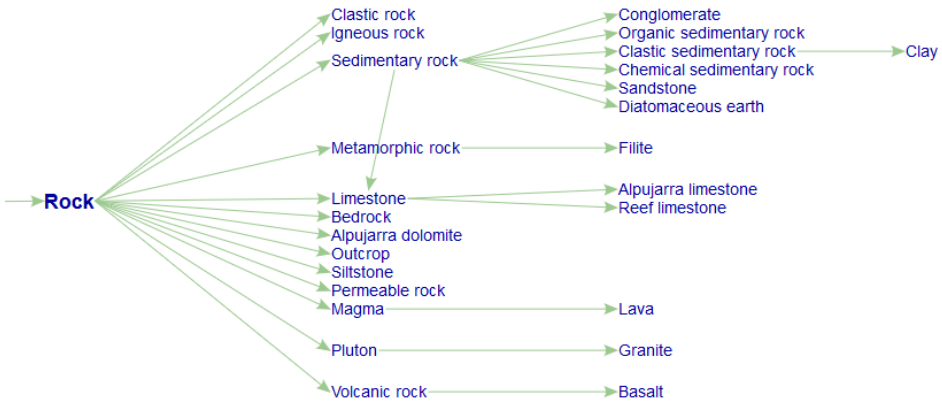


Figure 4. Conceptual hierarchy in EcoLexicon with a single hyponymic relation (*type of*)

Filtering conceptual relations so that only hyponymic relations are selected and generating tree-like hierarchies with this information is an excellent way of representing hyponymy in a terminological resource. Therefore, of all the resources shown, EcoLexicon has the most effective hyponymy representation. However, there is still room for improvement. In this regard, Gil-Berrozpe (2016) and Gil-Berrozpe *et al*. (2018) detected a number of problems such as the visualization of dimensions of co-hyponyms at the same level without any distinction, noise, information overload,

redundancy, and transitivity problems regarding property inheritance. For this reason, it was necessary to find an alternative way to accurately represent hyponymy.

This paper introduces a new approach to the description and representation of hyponymy based on hierarchical structure, intensional definitions, conceptual categories, hyponymy subtypes, and hyponymic contexts. The aim was thus to create the entries and to design the template of a hyponymy-based terminological resource: HypoLexicon.

Materials and Methods

The materials used in this research include four specialized environmental subcorpora from the EcoLexicon English Corpus. The software used was the EcoLexicon internal application, Sketch Engine, and Lexonomy. The methodology mainly involved information extraction (including corpus analysis to compile, identify, and select all relevant information regarding hypernyms and hyponyms) and design of the terminological template for the hyponymy-based terminological entries.

The EcoLexicon English Corpus (EEC) is a 23.1-million-word corpus of contemporary environmental texts compiled by the LexiCon Research Group (León-Araúz *et al*., 2018, 2019). The EEC was processed and compiled in an internal application of the research group, but it was also recompiled in Sketch Engine[3] with the Penn Treebank tagset (TreeTagger version 3.3) and with the EcoLexicon Semantic Sketch Grammar (ESSG) (León-Araúz *et al*., 2016; León-Araúz & San Martín, 2018). The ESSG is a customized sketch grammar that extracts semantic word sketches based on the most common conceptual relations (i.e., hyponymic, meronymic, locative, causal, functional). The hyponymic or generic-specific word sketches were used as the main method for hyponymic information extraction.

For the sake of delimiting the scope of the study, and of analyzing and comparing hyponymy across microdomains, four subcorpora were extracted from the EEC: a Biology subcorpus (BIO: 6,217,032 words), a Chemistry subcorpus (CHEM: 2,984,197 words), a Civil Engineering subcorpus (CIV: 4,491,909 words), and a Geology subcorpus (GEO: 3,975,045 words).

1.17 Hyponymic Information Extraction

Hypernym extraction, identification, and selection was based on the *Keywords* function in Sketch Engine, which extracts the most relevant single-word terms (SWTs) and multi-word terms (MWTs) from a corpus. We thus identified the three most representative terms of each subcorpora, which became the candidate hypernyms for each domain (BIO, CHEM, CIV, and GEO). Three was considered the optimal number because the objective was to create twelve terminological entries with sufficient conceptual, relational, and contextual information for the ontological categories in each environmental domain.

The advanced query of the *Keywords* function was used. This option allowed us to apply different criteria (e.g., focus corpus, reference corpus, rarity, minimum

[3] Available at: https://www.sketchengine.eu/

frequency, maximum frequency, etc.) to refine the query. The results are then given in two tabs (SWTs and MWTs) in the form of columns where it is possible to filter the information (e.g., hits in the focus corpus, hits in the reference corpus, keyness score, etc.). The four environmental subcorpora were processed by comparing them with the English Web 2020 (enTenTen20) general corpus in Sketch Engine. Moreover, empty words and non-terms were excluded. The four queries were performed according to different rarity levels (BIO: 50; CHEM: 10; CIV: 10; GEO: 100) so as to obtain the three best hypernyms because the texts in each subcorpus had different levels of specificity.

The three candidate terms of each subcorpus with the highest keyness score were selected as the three hypernyms. They were BACTERIUM, REEF, and CELL in the BIO subcorpus; SLUDGE, NITROGEN, and MAIZE in the CHEM subcorpus; WASTEWATER, BREAKWATER, and POLLUTANT in the CIV subcorpus; and EARTHQUAKE, SEDIMENT, and SOIL in the GEO subcorpus. Below, Table 1 shows as an example the hypernym identification and selection processes in the GEO subcorpus.

	TERM	FOCUS	REFERENCE	KEYNESS SCORE
1	**earthquake**	6,292	734,916	12.3
2	**sediment**	4,698	346,627	10.2
3	**soil**	6,659	2,469,621	9.8
4	wave	5,901	2,975,621	8.1
5	earth	7,226	5,210,790	7.6
6	surface	7,220	5,472,604	7.4
7	rock	6,424	5,011,777	6.9
8	water	14,184	17,935,266	6.2
9	ecosystem	3,034	937,880	6.2
10	velocity	2,776	591,949	6.1

Table 1. Hypernym identification and selection in the GEO subcorpus

Afterwards, hyponym extraction and identification was based on the *Word Sketch* (WS) function in Sketch Engine, which provides summaries of a term's grammatical and collocational behavior. This selection was validated and expanded by CQL queries performed with the *Concordance* function. The WS queries were performed for each of the twelve hypernyms extracted in the previous step of the corpus analysis. Furthermore, a customized CQL search using the *Concordance* function was employed to validate and expand the hyponym lists extracted with the WSs:

[tag="N.*|JJ.*|RB.*|VVN.*|VVG.*"]{1,}[lemma="HYPERNYM"]

This CQL query searches for any lemmatized hypernym ([lemma="hypernym"]) preceded one or more times ({1,}) by any noun (N.*), adjective (JJ.*), adverb (RB.*), verb in past participle form (VVN.*), or verb in gerund or present participle form (VVG.*).

Hyponym selection then involved classifying and comparing all the hyponymic information retrieved in the previous extraction and selection process. The data from the WS queries were validated and further expanded with the data from the customized CQL queries. As an illustrative example, Table 2 shows a segment of the hyponym selection of the hypernym SEDIMENT. In accordance with the *SEDIMENT is the generic of* WS extraction and identification, four hyponyms (e.g., SAND, SILT, CLAY) were selected. The combination of the MWT WS, the *modifier* WS, and the MWT CQL query produced 27 hyponyms (e.g., COHESIVE SEDIMENT, SUSPENDED SEDIMENT, STREAM SEDIMENT).

	SEDIMENT hyponyms (*is the generic of* WS) [total frequency = 4,698]		
	TERM	**FREQUENCY**	**SCORE**
1	sand	14	10.4
2	silt	10	10.5
3	clay	8	9.8
4	gravel	5	9.3

	SEDIMENT hyponyms (MWT WS, modifier WS & MWT CQL) [total frequency = 4,698]			**FOUND WITH**		
	TERM	**FREQUENCY**	**SCORE**	**MWT WS**	**MODIFIER WS**	**MWT CQL**
1	cohesive sediment	196	10.8	X	X	X
2	suspended sediment	100	9.9	X		X
3	stream sediment	68	9.3	X	X	X
4	lake sediment	53	9.0	X	X	X
5	marine sediment	50	8.9	X	X	X
6	bottom sediment	36	8.4	X	X	X
7	coastal sediment	35	8.4	X	X	X
8	fine sediment	33	8.3	X	X	X
9	coarse sediment	23	8.7		X	X
10	fine-grained sediment	22	7.7	X	X	X
	…					

Table 2. Segment of the hyponym selection of SEDIMENT

1.18 Hyponymy-based Template Design

Once all the data regarding both hypernyms and hyponyms had been selected, the next step was to create the conceptual hierarchies for the hyponymy-based terminological entries. These terminological entries were to portray four main elements: (i) terminological definitions; (ii) conceptual categories (Gil-Berrozpe *et al*. 2019); (iii) hyponymy subtypes in a hierarchical structure (Gil-Berrozpe *et al*. 2017); and (iv) hyponymic contexts (Gil-Berrozpe *et al*. 2017). The entries were based on the

information contained in EcoLexicon (regarding hierarchical structure, terminological definitions, conceptual categories, and hyponymy subtypes), which was enhanced and nurtured by the corpus analysis carried out to specifically extract hyponymic information.

Therefore, the final step was the design of the terminological template for the twelve hyponymy-based terminological entries. The software Lexonomy[4] was used for this purpose. Since the terminological entries in Lexonomy are written in XML, they can be designed from scratch to meet the needs of the terminological resource in question.

The elements included in the design of the terminological template for the hyponymy-based terminological entries were the following: (i) parent or superordinate concept (represented by a hypernym); (ii) child or subordinate concepts (represented by hyponyms); (iii) up to six hyponymy levels; (iv) terminological definitions; (v) conceptual categories; (vi) hyponymy subtypes; and (vii) hyponymic contexts.

As required by Lexonomy, the entry structure of this terminological template was designed in XML (Table 3), following a hierarchical structure for the representation of the hyponymic relations. Parent concepts and hypernyms, child concepts and hyponyms, definitions, hyponymy subtypes, and hyponymic contexts were introduced as elements (represented between angle brackets, < >), whilst conceptual categories were introduced as attributes (preceded by @).

```
<parentconcept>
    <hypernym>
        @conceptualcategory
    <definition_hyper>
    <hyponymiccontext_hyper>
    <childconcept-1>
        <hyponymysubtype_hypo1>
        <hyponym-1>
            @conceptualcategory
        <definition_hypo1>
        <hyponymiccontext_hypo1>
        <childconcept-2>

    <hyponymysubtype_hypo2>
            <hyponym-2>

    @conceptualcategory
            <definition_hypo2>

    <hyponymiccontext_hypo2>
            <childconcept-3>
                …
```

[4] Available at: https://www.lexonomy.eu/

Table 3. Segment of the entry format of the hyponymy-based terminological template in XML

HypoLexicon: A Hyponymy-based Terminological Resource

HypoLexicon[5] is a terminological resource focused on the description, categorization, and representation of hyponymy in environmental concepts. It is designed as a stand-alone module for EcoLexicon, since it is also one of its by-products. It includes definitional, relational, ontological, and contextual information about specialized hypernyms and hyponyms of environmental terminology. It is thus the main result and the practical application of this study, because it is the resource in which the hyponymy-based terminological entries were compiled and shared. HypoLexicon is publicly available on the Lexonomy platform and was published using the Creative Commons (CC) Attribution 4.0 International license.

The home view shows the main menu in HypoLexicon on the Lexonomy platform (Figure 5). This section is composed of the following elements: (i) resource title; (ii) resource description; (iii) search bar to perform queries; and (iv) list of random entries. The search bar and the list of random entries are designed for terminological resources published in Lexonomy with a large number of entries. However, since the number of entries in HypoLexicon is still rather small, these features are less relevant, other than providing direct access to all entries from the list of random entries. In the upper right corner of the home view, users can log in to Lexonomy if an account on this platform is available (e.g., access as an administrator to manage this resource, access as a contributor to add or modify entries in this resource). The font size can also be increased or reduced for better accessibility.

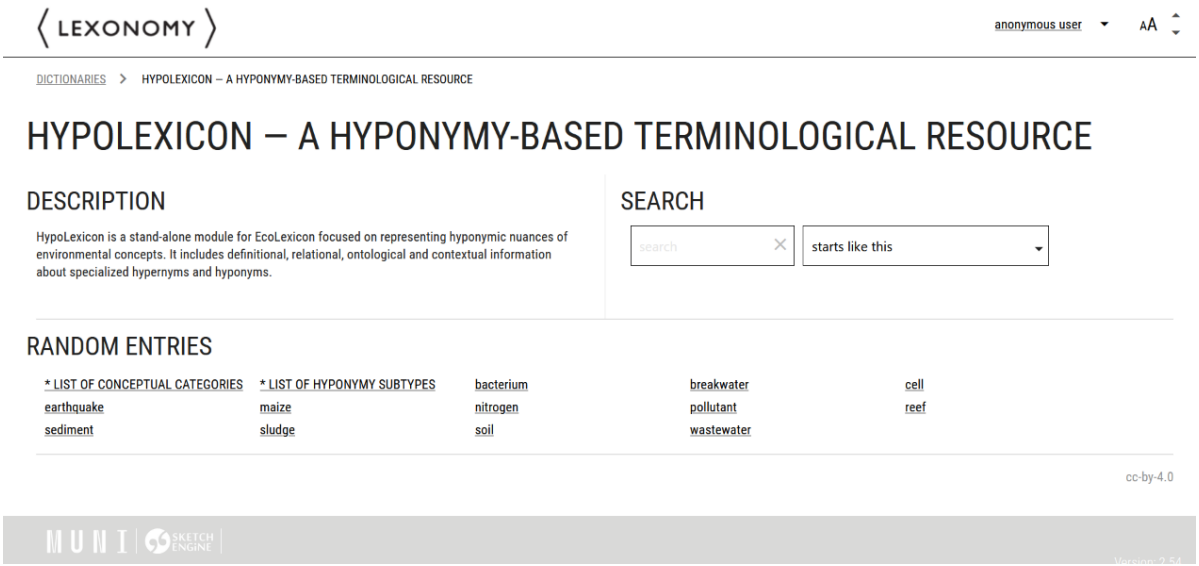


Figure 5. Main menu in HypoLexicon

[5] Available at: https://www.lexonomy.eu/hypolexicon

The entry view shows the contents of HypoLexicon, which are the following: (i) the twelve hyponymy-based terminological entries; (ii) list of conceptual categories; and (iii) list of hyponymy subtypes. Because of the importance of conceptual categories and hyponymy subtypes in HypoLexicon, legends were added so that users could access the complete inventories within the same platform, without searching for the information in EcoLexicon.

Figure 6 shows a segment of one of the twelve hyponymy-based terminological entries in HypoLexicon. It corresponds to the hypernym SEDIMENT. On the left side of the entry view, the search bar is followed by the full list of entries in HypoLexicon in alphabetical order. On the right side of the entry view, the contents of each terminological entry are displayed.

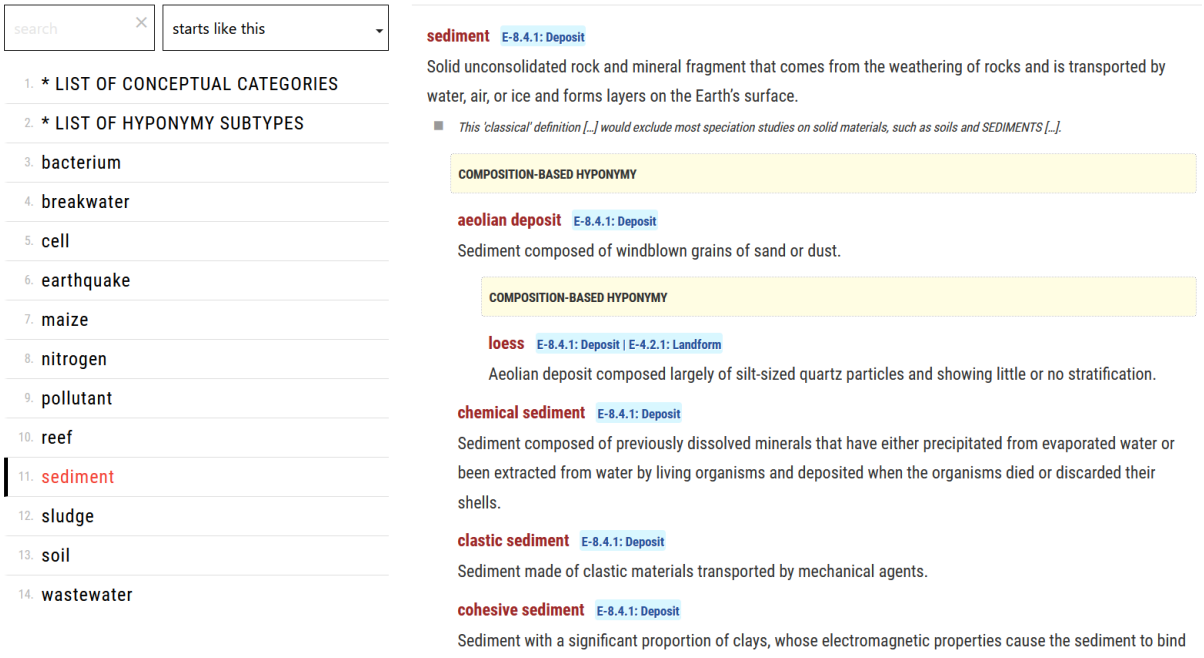


Figure 6. Segment of the SEDIMENT terminological entry in HypoLexicon

The SEDIMENT terminological entry in HypoLexicon has the following information: 48 concepts with their definitions; 71 terms designating those concepts; five conceptual categories (i.e., *deposit*, *landform*, *mineral*, *particle*, and *rock*); eight hyponymy subtypes (i.e., *composition-based*, *effect-based*, *function-based*, *location-based*, *movement-based*, *origin-based*, *size-based* and *status-based* hyponym); up to four hyponymy levels; and 17 concepts with hyponymic contexts. This is the richest entry not only of the GEO entries, but of all entries in relation to the number of concepts and terms. Therefore, this generates a wider variety of conceptual categories and hyponymy subtypes.

Of the conceptual categories in this entry, the predominant one is *deposit*, which is present in all concepts (e.g., COHESIVE SEDIMENT, FINE SEDIMENT, TERRIGENOUS SEDIMENT). However, as concepts are further specified in the conceptual hierarchy through hyponymy subtypes, this also generates more specific categories, such as

landform (e.g., CENTRAL MORAINE, LATERAL MORAINE, LOESS), *mineral* (e.g., DRY-SCREEN SAND, FINE SAND, ORIGINAL SAND), particle (e.g., COARSE SAND, FINE SEDIMENT, INTRACLAST), and *rock* (CLAY, DIAMICTITE, SILT). Therefore, thanks to the structure of the entry, it is possible to see and identify that the *deposit* category, which is more general, also acquires more nuances as the conceptual hierarchy progresses.

On the other hand, the most relevant hyponymy subtypes in this entry are *composition-based*, *location-based*, and *origin-based* hyponymy. The *composition-based* hyponyms (e.g., CARBONATE SEDIMENT, INTRACLAST, TERRIGENOUS SEDIMENT) are differentiated from their hypernyms because of the materials they are made of. The *location-based* hyponyms (e.g., ALLUVIAL SEDIMENT, GROUND MORAINE, SHALLOW SEDIMENT) are determined by the place where the sediment is deposited or where it is typically formed. Finally, the *origin-based* hyponyms (e.g., BIOGENIC SEDIMENT, ORIGINAL SAND, RECESSIONAL MORAINE) are characterized by the process that creates or originates them. It is thus clear that, by explicitly stating the hyponymic nuance that makes a hyponym more specific than its hyperonym, the conceptual and terminological understanding of these concepts is improved.

Interestingly, the only concept at the fourth and last hyponymy level of this entry is DIAMICTITE, which has the following schema: DIAMICTITE is a *composition-based* type of TERRIGENOUS SEDIMENT, which is a *composition-based* type of SHALLOW SEDIMENT, which is a *location-based* type of MARINE SEDIMENT, which is a *location-based* type of SEDIMENT. In this sequence, *location-based* hyponymy is at the first and second levels, whereas *composition-based* hyponymy is at the third and fourth levels. However, *composition-based* hyponymy is at the highest hyponymy levels of other sequences (e.g., those of CHEMICAL SEDIMENT, COHESIVE SEDIMENT, and INTRACLAST), and *location-based* is at the lower hyponymy levels of other sequences (e.g., those of DEPOSITED SEDIMENT, FLUVIAL SEDIMENT, and BEACH SEDIMENT). Therefore, sometimes it can be difficult to establish a common pattern regarding which levels are characteristic of certain subtypes.

Finally, in relation to hyponymic contexts, there are many examples in this entry that show different ways of expressing hyponymic knowledge patterns (e.g., "unusual types of gravel and conglomerate include tillites […] and diamictite"; "bed load material such as gravel and sand"; "stream sediments, soils, and other near-surface materials"). Revealing this type of grammatical and syntactic information allows us to better understand the collocational behavior of hypernym and hyponym pairs, as well as to detect the most common language forms for encoding hyponymic relations.

Similarly, the remaining eleven terminology entries of which HypoLexicon is composed also provide the equivalent information on the remaining environmental hyperonyms and hyponyms. This also allows users to check, for example, which conceptual categories or subtypes of hyponymy are the most characteristic depending on different specialized domains.

Conclusions

HypoLexicon is the convergence point of four resources: (i) EcoLexicon, for the basic structure and information of the terminological entries; (ii) the EcoLexicon English

Corpus and the four specialized subcorpora, for the population and enhancement of the terminological entries; (iii) Sketch Engine, for the extraction of hyponymic and contextual information through corpus analysis; and (iv) Lexonomy, for the design of the terminological template and for the implementation of all data in the form of an actual terminological resource.

The hyponymy-based terminological entries in HypoLexicon are a successful approach to the description of hyponymic information because of their hierarchical structure and graphical classification of information based on definitional and corpus analysis. Moreover, the visualization of hyponymic information in them permits the identification of dynamic phenomena regarding generic-specific relations (e.g., hyponymic nuances in the verticality and horizontality of the conceptual hierarchies, different dimensions or microsenses of co-hyponyms, changes in characteristics of concepts through the addition of conceptual categories at more specific hyponymy levels, etc.).

Therefore, this paper has presented a new way of representing hyponymy in terminological resources. This methodology is also applicable to any other specialized domain, and may even provide an accessible way of dealing with hyponymy in general language resources as well. Basically, the objective of the methodology and resource proposed is to facilitate knowledge acquisition at all level.

Future work in this research line will take two paths. On the one hand, HypoLexicon can continue to grow and be nourished with more content by creating additional terminological entries with all kinds of hyponymic information extracted from corpus techniques. These new entries, moreover, could belong to the same environmental subdomains or to new ones so as to extend the range of conceptual categories and hyponymy subtypes. However, perhaps the most innovative idea would be to seek the total integration of HypoLexicon in EcoLexicon. In this way, it would cease to be a stand-alone module or a by-product, and would become an integral part of the original resource.

Acknowledgements

This research was carried out as part of project PID2020-118369GB-I00, *Transversal Integration of Culture into an Environmental Terminological Knowledge Base* (TRANSCULTURE), funded by the Spanish Ministry of Science and Innovation; and as part of project A-HUM-600-UGR20, *La cultura como módulo transversal en una base de conocimiento terminológico medioambiental* (CULTURAMA), funded by the European Regional Development Fund (FEDER).

References

Antia, Bassey E. 2000. *Terminology and Language Planning: An Alternative Framework of Practice and Discourse*. Amsterdam/Philadelphia: John Benjamins.

Cabré, María Teresa. 1999. *La terminología: Representación y comunicación*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Cabré, María Teresa. 2000. Elements for a Theory of Terminology: Towards an Alternative Paradigm. *Terminology*, 6(1), 35–57. Amsterdam/Philadelphia: John Benjamins.

Cruse, D. Alan. 1995. Polysemy and Related Phenomena from a Cognitive Linguistic Viewpoint. In P. Saint-Dizier & E. Viegas (eds.), *Computational Lexical Semantics*. Cambridge: Cambridge University Press.

Cruse, D. Alan. 2002. Hyponymy and Its Varieties. In R. Green, C. A. Bean, & S. H. Myaeng (eds.), *The Semantics of Relationships: An Interdisciplinary Perspective*, 3–22. Dordrecht/Boston/London: Kluwer Academic Publishers.

Faber, Pamela (ed.). 2012. *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin/Boston: De Gruyter Mouton.

Faber, Pamela, Pilar León-Araúz, and Arianne Reimerink. 2014. Representing Environmental Knowledge in EcoLexicon. In E. Bárcena, T. Read, & J. Arús (eds.), *Languages for Specific Purposes in the Digital Era*, 19, 267–301. Berlin/Heidelberg: Springer.

Faber, Pamela, Pilar León-Araúz, and Arianne Reimerink. 2016. EcoLexicon: New Features and Challenges. In I. Kernerman, I. Kosem Trojina, S. Krek, & L. Trap-Jensen (eds.), *GLOBALEX 2016: Lexicographic Resources for Human Language Technology in Conjunction with the 10th Edition of the Language Resources and Evaluation Conference*, 73–80. Portorož, Slovenia.

Faber, Pamela. 2009. The Cognitive Shift in Terminology and Specialized Translation. *MonTI (Monografías de Traducción e Interpretación)*, 1, 107–134. Valencia: Universitat de València.

Faber, Pamela. 2022. Frame-based Terminology. In P. Faber & M. C. L'Homme (eds.), *Theoretical Perspectives on Terminology: Explaining terms, concepts and specialized knowledge, Terminology and Lexicography Research and Practice*, 23, 353–376. Amsterdam/Philadelphia: John Benjamins.

Gil-Berrozpe, Juan Carlos, Pilar León-Araúz, and Pamela Faber. 2017. Specifying Hyponymy Subtypes and Knowledge Patterns: A Corpus-based Study. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, & V. Baisa (eds.), *Proceedings of the eLex 2017 Conference – 5th International Conference on Electronic Lexicography in the 21st Century*, 63–92. Brno: Lexical Computing CZ s.r.o.

Gil-Berrozpe, Juan Carlos, Pilar León-Araúz, and Pamela Faber. 2019. Ontological Knowledge Enhancement in EcoLexicon. In I. Kosem, T. Zingano Kuhn, M. Correia, J.P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek, & C. Tiberius (eds.), *Proceedings of the eLex 2019 conference – 6th International Conference on Electronic Lexicography in the 21st century*, 177–197. Brno: Lexical Computing CZ, s.r.o.

Gil-Berrozpe, Juan Carlos, Pilar León-Araúz, and Pamela Faber. 2018. Subtypes of Hyponymy in the Environmental Domain: Entities and Processes. In C. Roche (ed.), *TOTh 2016 – Terminology & Ontology: Theories and Applications, Terminologica*, 39–54. Chambéry: Éditions de l'Université de Savoie Mont Blanc.

Gil-Berrozpe, Juan Carlos. 2016. *Extending the Conceptual Systems in EcoLexicon to Enhance Multidimensionality*. BA Thesis. Granada: Universidad de Granada.

Gil-Berrozpe, Juan Carlos. 2020. Attribute-based Approach to Hyponymic Behavior in Botanical Terminology. In C. Roche (ed.), *TOTh 2019 – Terminology & Ontology: Theories and Applications, Terminologica*, 93–108. Chambéry: Éditions de l'Université de Savoie Mont Blanc.

Hine, Robert (ed.). 2019. *A Dictionary of Biology*. 8th ed. Oxford: Oxford University Press.

León-Araúz, Pilar, and Antonio San Martín. 2018. The EcoLexicon Semantic Sketch Grammar: from Knowledge Patterns to Word Sketches. In I. Kerneman & S. Krek (eds.), *Proceedings of the LREC 2018 Workshop Globalex 2018 – Lexicography & WordNets*, 94–99. Miyazaki: Globalex.

León-Araúz, Pilar, Antonio San Martín, and Arianne Reimerink. 2018. The EcoLexicon English Corpus as an open corpus in Sketch Engine. In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (eds.), *Proceedings of the 18th EURALEX International Congress*, 893–901. Ljubljana: EURALEX.

León-Araúz, Pilar, Antonio San Martín, and Pamela Faber. 2016. Pattern-based Word Sketches for the Extraction of Semantic Relations. In *Proceedings of the 5th International Workshop on Computational Terminology*, 73–82. Osaka, Japan.

León-Araúz, Pilar, Arianne Reimerink, and Pamela Faber. 2019. EcoLexicon and by-products: integrating and reusing terminological resources. In A. Alcina, R. Costa, & C. Roche (eds.), *Special issue of Terminology and e-dictionaries, Terminology*, 25(2), 222–258. Amsterdam/Philadelphia: John Benjamins.

McGraw-Hill (ed.). 2003. *Dictionary of Geology & Mineralogy*. 2nd ed. New York: McGraw-Hill.

Meyer, Ingrid, Lynne Bowker, and Karen Eck. 1992. COGNITERM: An Experiment in Building a Knowledge-based Term Bank. In *Proceedings of the 5th EURALEX International Congress*, 159–172. Tampere, Finland.

Murphy, M. Lynne, and Anu Koskela. 2010. *Key Terms in Semantics*. London/New York: Continuum.

Murphy, M. Lynne. 2003. *Semantic Relations and the Lexicon: Antonymy, Synonymy and Other Paradigms*. Cambridge: Cambridge University Press.

Murphy, M. Lynne. 2006. Hyponymy and Hyperonymy. In K. Brown (ed.), *Encyclopedia of Language and Linguistics*, 1, 446–448. New York: Elsevier.

Oxford University Press. 2010. *Oxford Dictionary of English*. 3rd ed. Oxford: Oxford University Press.

Oxford University Press. 2022. *OED Online*. Oxford: Oxford University Press. Available at: https://www.oed.com

Rittner, Don, and Timothy Lee McCabe (eds.). 2004. *Encyclopedia of Biology*. New York: Facts On File.

Sager, Juan Carlos. 1990. *A Practical Course in Terminology Processing*. Amsterdam/Philadelphia: John Benjamins.

Sager, Juan Carlos. 1994. Terminology: Custodian of Knowledge and Means of Knowledge Transfer. *Terminology*, 1(1), 7–15. Amsterdam/Philadelphia: John Benjamins.

Temmerman, Rita. 2000. *Towards New Ways of Terminology Description: The Sociocognitive Approach*. Amsterdam/Philadelphia: John Benjamins.

Wüster, Eugen. 1968. *The Machine Tool. An Interlingual Dictionary of Basic Concepts*. London: Technical Press.

Wüster, Eugen. 1979. *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*. Vienna: Springer.

Zorrilla-Agut, Paula, and Thierry Fontenelle. 2019. IATE 2: Modernising the EU's IATE terminological database to respond to the challenges of today's translation world and beyond. *Terminology*, 25(2), 146–174. Amsterdam/Philadelphia: John Benjamins.

Using bitext mining to identify translated material: practical assessment and new applications

Zhilu Tu[1]
Hong Kong Baptist University
tuzhiluke@life.hkbu.edu.hk

Minghao Wang[1]
Hong Kong Baptist University
19454414@life.hkbu.edu.hk

Mark Shuttleworth[1]
Hong Kong Baptist University
markshut@hkbu.edu.hk

Zhiwen Hua[1]
Hong Kong Baptist University
20462786@life.hkbu.edu.hk

Abstract

Locating translated text can be thought of as a way of "reverse engineering" a complex translation job. By utilising tools from bitext mining, this study attempts to facilitate pre-trained models using techniques from natural language processing (NLP) within translation studies research as an example of how the approach may be applied in other contexts. Starting with a brief review of the development and application of text alignment, this study further substantiates the feasibility of bitext-mining in the case of Wikipedia's translation and multilingual practice via a practical assessment of the reliability of auto-aligned results. The assessment method involves a study and in situ observations of Shuttleworth's (2018) and more recent on-going work of finding translation fragments. The paper then describes new applications for bitext mining facilitated by improvements to alignment tools including a language model selection that may increase sensitivity to semantically or structurally close sentences; granular sentence segmentation that helps to reveal smaller translation units; and an interactive front-end design for highlighting the distribution of the alignments for users' reference. The study thus provides an outlook on possible new applications of bitext mining.

Background

Bitext mining has great potential as a tool for locating translation pairs, and the sentence alignment techniques on which it is based have already been used in areas such as corpus construction (Wu, 1994) and statistical machine translation (Brown et al., 1993). Deriving from Gale-Church alignment (1993), a sentence length-based algorithm, traditional alignment techniques represented by Hunalign (Varga et al., 2007) and Champollion (Ma, 2006) assume related sentence length and rely on bilingual cross-reference resources, and their alignment tends to be monotonic (Wu, 2010). Besides, the approach relies heavily on an initial building on a bilingual lexicon or algorithm, which is not usually easy to obtain or construct (Couto, 2017); thus, their performance would be largely restricted to certain language pairs.

Since the introduction of Word2Vec (Le & Mikolov, 2014), a distributed word vector, in 2013, cross-lingual word/sentence embedding has emerged from the success of word embedding, intending to align embedding spaces rather than lexicons (Couto, 2017). As one of the downstream tasks of multilingual sentence embedding, several of the best alignment methods are currently implemented based on large-scale pre-trained models, such as LASER (Schwenk, 2018) and LaBSE (Feng et al., 2022). They are usually trained with enormous parallel datasets, achieving multilingual embedding in the same vector space. When different languages can be mapped into the same vector space, bitext mining can be implemented in a margin-based manner (Artetxe & Schwenk, 2019). Bitext mining has thus succeeded in further improving accuracy by utilizing a pre-trained deep learning model while featuring non-monotonicity.

1 Equal contributions

The performance of bitext mining utilizing language models has proved itself to be very impressive using the BUCC benchmark (Reimers & Gurevych, 2020). A variety of big data applications have now been created using bitext mining in both academia and the industry (Schwenk, Chaudhary, et al., 2019; Schwenk, Wenzek, et al., 2019), but it is rarely mentioned in the context of real-world data applications with small sample sizes and high accuracy requirements. The paper investigates the use of bitext mining tools within a small-scale dataset setting as an example of how they might be exploited in translation studies and other real-world scenarios.

Objectives

The translated material in Wikipedia is difficult to locate, and researchers have so far struggled to map it out (Shuttleworth, 2017). In Shuttleworth's previous research (2018), translation fragments have been found in Wikipedia articles about the murder of the Russian politician Boris Nemtsov from different multilingual revisions, which has led to discrepancies and changes in points of view. During the research, no targeted and sustainable tool was available to locate translations in Wikipedia consistently, thus the research materials had to be manually checked one by one, which compromises the efficiency of the research and the possibility of enlarging the study scale.

From the perspective of translation studies, the need for bitext mining lies in its ability to quickly locate the translation pairs present in a project sample. Therefore, the hypothesis is that bitext mining can help reduce the researcher's workload while opening up the possibility of analysing more data. With the known performance of bitext mining, the objective is to refactor the coding so that it can be optimized for (Wikipedia) translation research scenarios. The new features include but are not limited to automatic fetching and segmentation of Wikipedia articles, interfacing different language models and a front and back-end integrated system with multi-functionality. These improvements will be considered as new applications of bitext mining, and the results will be further qualitatively assessed for practicality.

Methodology

To achieve the stated goal, this paper will be dedicated to developing a novel interactive bitext mining tool, namely WikiAligner, within the context of Wikipedia research. The flowchart of the tool is shown in Figure 1, which illustrates the front-end on the left and the back-end on the right. It simulates every step of how the parameters are passed from the front end to the back end and reflow to the front end for visualizing outputs.

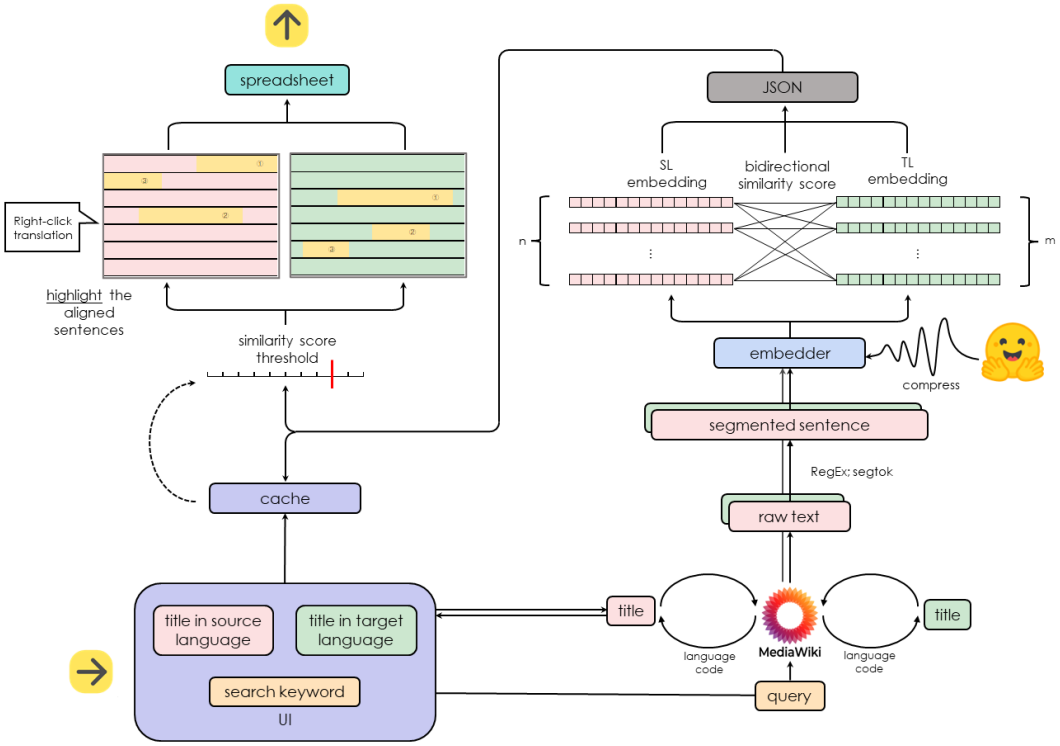


Figure 1: WikiAligner workflow

When using the tool, the front-end UI will guide the user through the whole bitext mining process. As with other web-based services, potential users do not need to attend the coding of the bitext tool before using, which could help the tool reach a wider audience. After entering the keywords of SL and TL's title from Wikipedia in the front-end UI, the tool will proceed with the raw texts from Wikipedia's database once the titles are typed and selected from the search prompt. The mentioning process is supported by the MediaWiki API, which in a way ensures ethical and textual integrity while allowing users to be seamlessly directed to Wikipedia's official search prompt.

Passing the requests from the front end, then the back end starts to step in, which is considered the second phase in the flowchart. The article pairs acquired by the user, if marked up, will firstly be uniformly formatted as plain text. Then depending on the language of the articles, syntok[1] or related regular expressions will be used to split the plaintexts into separate sentences for further sentence embedding. According to Reimer and Gurevych's evaluation (2020), LaBSE is currently one of the best methods for bitext mining and therefore set as the default pre-trained model. Taking the performance of the server into account, LaBSE will be compressed to improve loading speed before embedding. During the embedding process, all sentences from SL and TL will be encoded to a vector value which can be used for a margin-based similarity calculation (Artetxe & Schwenk, 2019). The process requires a k candidates-based bi-directional computation on the cosine distance between sentences, which represents the similarities. After finishing the computation, the program will generate a JSON file based on the sentence sequence IDs from SL and TL, in which the IDs will be combined as translation

[1]A freely available Python library that supports most of the Indo-European languages: https://github.com/fnl/syntok

pairs, appending with a bi-directional similarity score to each of them. The information in this JSON file will be further displayed by the front end for visualization and other advanced features.

When the JSON file is cached to the server, the front end aims to insert the sentence IDs to the texts and distribute them with a highlight visualization. Inspired by modern MT UI design, the corresponding TT will be highlighted when the cursor hovers over an ST segment, with a similarity score attached for the user's reference. Catering to numerous potential ST-TT segments with similarity scores, tags and thresholds will be placed with front-end events, and they will be brought out when overviewing all the potential translation outputs from the back-end. Tags that follow each sentence are aligned from ST to TT, which is designed to distinguish different highlights, while threshold filters similarities score to rule out low quality translation pairs. Together they help the user scrutinize and locate the translation segments that might be significant for the research. A right-click event is also set to redirect a designated sentence to an MT result, which facilitates users reconfirming the alignment and also researching uncharted languages across Wikipedia articles. Finally, in cooperation with the back end, the front end supports the output of bitext mining results as a spreadsheet, which complement the output scheme and offer the user a method to save and look up the result locally. To overview how the front end works in this phase, Figure 2 shows the current stage of development.

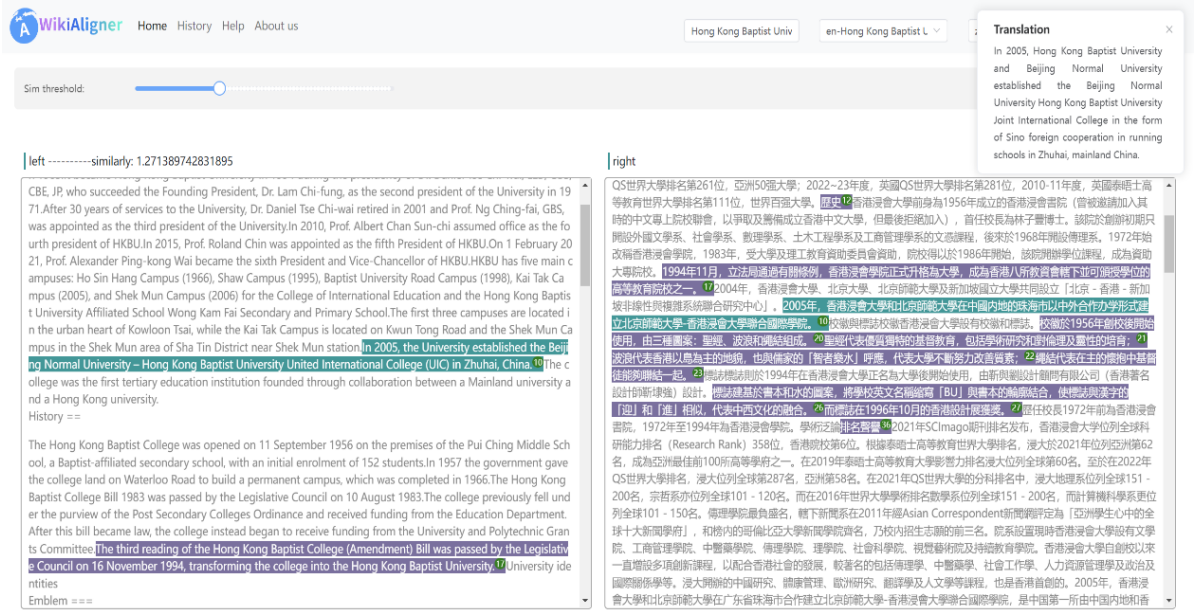


Figure 2: WikiAligner - Front end re-presentation of the Wikipedia article *Hong Kong Baptist University* (English-Chinese)

Discussion

It is surprising but reasonable to find all alignments that have been extensively analyzed by Shuttleworth (2018) are also located by using the bitext mining (LaBSE) method. The method has also shown a promising performance in Shuttleworth's subsequent work. It sustains the feasibility of using bitext mining within an academic setting. Based on the interactive GUI, the WikiAligner complements bitext mining in its accessibility and manifestations. When reviewing the context that Shuttleworth focuses on, the tool helps to reveal the distribution of translation and researchers will also be able to scrutinize every translation pair by using the highlight feature. From another point of view, the highlight feature has also exposed those

isolated segments that have not been translated into another language, which can also be significant to trace the shifting of POV within multilingual Wikipedia articles. While exporting the statically aligned mined bitext output provides an incredibly valuable research resource, the tool also offers a more dynamic representation. The tool's static output presents candidate pairs of aligned sentences indicate their place in the text and can be ordered by their certainty. On the other hand, the dynamic highlighting of paired sentences within the two texts can provide an excellent visualization how the content of the texts can be presented in orders that differ significantly from each other. At the same time, the excellent results that are generally obtained from using the tool will free up researchers' time to allow them to focus on other, more qualitative areas of analysis.

The tool has also scheduled several improvements regarding the limitations found during in-house usage. Firstly, granularized segmentation can be introduced to reveal smaller translation units within long sentences. This is because translation pairs can be taken as units of meaning groups. Prolonged or unstructured sentences from the SL or TL may result in sloppy tokenization. A granular design could bring more translation pairs to the surface and thus harvest more potential bitext mining results.

Apart from sentence segmentation, the tool will also try to adapt the PARSE API from MediaWiki that supports calling for article revisions and outputting as a parsed HTML. Not only can it help avoid parsing complicated wikitext that may cause the loss of textual data, but it can also give the tool the ability to align historical articles from different languages' timeframes. On top of that, accessing revisions would also let intralingual alignment become possible. Instead of harvesting text from two languages, comparing the similarities between revisions of a single-language article can help reveal how the text has evolved and been reframed.

Finally, although LaBSE has performed well in most of the bitext mining scenarios that this study has covered, the drawback of the model may reveal itself when expanding the research scope. As the back-end of the tool is having native support from the Python framework sentence transformer (Reimers & Gurevych, 2019), up to 700 different models can be utilized seamlessly for the sentence embedding, which includes dedicated bilingual models that may perform better on high-resource languages, and models that have higher STS (Semantic Textual Similarity) scores to distinguish different forms of translations (Reimers & Gurevych, 2020). These improvements have been experimentally implemented in the tool and justified the revisiting of Shuttleworth's works.

Conclusion

Standing on the shoulders of giants from NLP, this paper presents a method to utilize bitext mining in the context of a small-scale dataset. Inspired by Shuttleworth's research, the paper has brought WikiAligner to investigate the "dark matter", i.e., the translated segments that exist in Wikipedia articles. Bitext mining has therefore further been concreted as a Browser/Server solution, connecting with a back-end for computing and a front-end representation. Nevertheless, it is also noted that the performance of bitext mining can be compromised in some corner cases as mentioned in the discussion. This may indicate that the current bitext mining method is not perfect, but the significance of it is to facilitate the alignment process rather than diminishing human effort.

That said, WikiAligner's practicality has also been proved, though not extensively, through a qualitative assessment of revisiting Shuttleworth's work. This new application of bitext mining

is by no means exclusive to researching Wikipedia translations. Instead, WikiAligner can be an entry point for applying bitext mining from big data to real-word alignment scenarios with new features, i.e., easier accessibility and novel manifestation. The support for aligning uploaded files is indeed on the way, and together with the back-end front-end integrated system, it can bring more insights by expanding the usage.

References

Artetxe, M., & Schwenk, H. (2019). Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3197–3203. https://doi.org/10.18653/v1/P19-1309

Couto, J. (2017, December 12). *Deep Learning for NLP: Advancements & Trends | Tryolabs*. https://tryolabs.com/blog/2017/12/12/deep-learning-for-nlp-advancements-and-trends-in-2017

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). Language-agnostic BERT Sentence Embedding. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 878–891. https://doi.org/10.18653/v1/2022.acl-long.62

Gale, W. A., & Church, K. W. (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, *19*(1), 75–102. https://aclanthology.org/J93-1004

Le, Q., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. In E. P. Xing & T. Jebara (Eds.), *Proceedings of the 31st International Conference on Machine Learning* (Vol. 32, Issue 2, pp. 1188–1196). PMLR. https://proceedings.mlr.press/v32/le14.html

Ma, X. (2006, May). Champollion: A Robust Parallel Text Sentence Aligner. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. http://www.lrec-conf.org/proceedings/lrec2006/pdf/746_pdf.pdf

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. https://arxiv.org/abs/1908.10084

Reimers, N., & Gurevych, I. (2020). Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. https://arxiv.org/abs/2004.09813

Schwenk, H. (2018). Filtering and Mining Parallel Data in a Joint Multilingual Space. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 228–234. https://doi.org/10.18653/v1/P18-2037

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., & Guzmán, F. (2019). WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 1351–1361. https://doi.org/10.48550/arxiv.1907.05791

Schwenk, H., Wenzek, G., Edunov, S., Grave, E., Joulin, A., & Fan, A. (2019). CCMatrix: Mining Billions of High-Quality Parallel Sentences on the WEB. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 6490–6500. https://doi.org/10.48550/arxiv.1911.04944

Shuttleworth, M. (2017). Locating foci of translation on Wikipedia. *Translation Spaces*, *6*(2), 310–332. https://doi.org/10.1075/TS.6.2.07SHU

Shuttleworth, M. (2018). Translation and the Production of Knowledge in "Wikipedia": Chronicling the Assassination of Boris Nemtsov. *Alif: Journal of Comparative Poetics*, *38*, 231–263. https://www.jstor.org/stable/26496376

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., & Trón, V. (2007). Parallel corpora for medium density languages. In N. Nicolov, K. Bontcheva, G. Angelova, & R. Mitkov (Eds.), *Recent Advances in Natural Language Processing IV: Selected papers from RANLP 2005* (pp. 247–258). John Benjamins Publishing Company. https://doi.org/10.1075/CILT.292.32VAR

Wu, D. (1994). Aligning a parallel English-Chinese corpus statistically with lexical criteria. *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics* -, 80–87. https://doi.org/10.3115/981732.981744

Wu, D. (2010). Alignment. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of Natural Language Processing* (2nd ed.). CRC Press.