Abigail R. Kaplan

# Suitability of Neural Machine Translation for Producing Linguistically Accessible Text

## Exploring the Effects of Pre-Editing on Easy-to-Read Administrative Documents

Directrice : Silvia Rodríguez Vázquez
Jurée : Pierrette Bouillon

Mémoire présenté à la **Faculté de traduction et d'interprétation** (Département de traduction, Unité d'anglais) pour l'obtention de la **Maîtrise universitaire en traduction, mention Traduction et communication spécialisée multilingue**

Université de Genève
Janvier 2021

# Déclaration attestant le caractère original du travail effectué

J'affirme avoir pris connaissance des documents d'information et de prévention du plagiat émis par l'Université de Genève et la Faculté de traduction et d'interprétation (notamment la *Directive en matière de plagiat des étudiant-e-s,* le *Règlement d'études des Maîtrises universitaires en traduction et du Certificat complémentaire en traduction de la Faculté de traduction et d'interprétation* ainsi que *l'Aide-mémoire à l'intention des étudiants préparant un mémoire de Ma en traduction*).

J'atteste que ce travail est le fruit d'un travail personnel et a été rédigé de manière autonome.

Je déclare que toutes les sources d'information utilisées sont citées de manière complète et précise, y compris les sources sur Internet.

Je suis consciente que le fait de ne pas citer une source ou de ne pas la citer correctement est constitutif de plagiat et que le plagiat est considéré comme une faute grave au sein de l'Université, passible de sanctions.

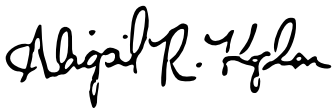
Au vu de ce qui précède, je déclare sur l'honneur que le présent travail est original.

Nom et prénom :

**KAPLAN Abigail Rose**

Lieu / date / signature :

**Aix-en-Provence**
**7 janvier 2021**



# Acknowledgements

First, my sincere gratitude to Silvia Rodríguez Vázquez, for your guidance, enthusiasm, attention, and care. You did not give up hope despite the many curveballs life threw at us along the way. Thank you for inspiring this project through your own work in the fields of translation technologies and accessibility, and for believing in me enough to orchestrate the incredible opportunity of presenting a portion of this research at the 2019 Klaara Conference.

I would also like to thank Pierrette Bouillon, who opened my eyes to the fascinating and complex world of machine translation, and all of the other smart and dedicated FTI professors who have shared their knowledge and helped shape me into a capable budding professional.

Special thanks to my six colleagues who graciously agreed to give up their time and post-editing energy to participate in my (long) study, and to all of my FTI friends and classmates. You awe me with both your creativity and your compassion, and you motivate me to strive for excellence in our field.

To my mom, dad, Jake, and Allan, thank you for knowing when to nudge and knowing when to sympathize. Thank you for showing me nothing but love and patience. Thank you for being my readers, my listeners, and always my biggest cheerleaders.

# Abstract

Governments have the potential to improve the civic inclusion of people with intellectual disabilities living in multilingual societies by providing administrative documents in an Easy-to-Read format and in the preferred language of the target audience. However, in the case of Switzerland and France, Easy-to-Read is not widely used for French administrative communication, and even less so for English. This study aims to address the obstacles that could be keeping the Swiss and French administrations from publishing more translated Easy-to-Read documents, or in other words, the barriers to successful accessible communication with their English-speaking citizens with disabilities. To this end, it investigates the suitability of a free and public neural machine translation system, DeepL, for generating linguistically accessible English from Easy-to-Read French source texts. With the goal of increasing the linguistic accessibility of the texts produced by the system, it proposes the introduction of a pre-editing step, in which formatting is removed from the French text prior to translation. A four-part study examined the issue from three different angles of linguistic accessibility: translation quality, accessibility, and readability. In the first three parts of the study, translation quality and accessibility were manually assessed by way of DQF-MQM error annotation, measurements of post-editing effort, and Easy-to-Read guideline violation annotation. The fourth part featured an automatic evaluation of readability. Findings allow us to conclude that adding a pre-editing step, which clarifies sentence boundaries for the machine, does in fact improve two of the three factors of linguistic accessibility examined, translation quality and accessibility, and make this type of tool more suitable for producing Easy-to-Read English translations.



Keywords: neural machine translation; controlled language; Easy-to-Read; accessibility; accessible communication

# Table of Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| AOA | Age of Acquisition |
| CL | Controlled Language |
| CNSA | *Caisse Nationale de la Solidarité pour l'Autonomie* |
| DQF-MQM | Dynamic Quality Framework-Multidimensional Quality Metrics |
| E2R | Easy to Read |
| EBMT | Example-Based Machine Translation |
| FAHQT | Fully Automatic High-Quality Translation |
| FALC | *Facile à Lire et à Comprendre* |
| HAMT | Human-Aided Machine Translation |
| HOCL | Human-Oriented Controlled Language |
| ID | Intellectual Disability |
| IFLA | International Federation of Library Associations and Institutions |
| MAHT | Machine-Aided Human Translation |
| MOCL | Machine-Oriented Controlled Language |
| MT | Machine Translation |
| MTPE | Machine Translation Post-Editing |
| NMT | Neural Machine Translation |
| PEE | Post-Editing Effort |
| RBMT | Rule-Based Machine Translation |
| SIMPLES | *Simplification des Langues Écrites* |
| SMT | Statistical Machine Translation |
| TAUS | Translation Automation User Society |
| TQA | Translation Quality Assessment |
| TS | Text Simplification |
| TTE | Time-to-Edit |
| W3C | World Wide Web Consortium |
| WCAG | Web Content Accessibility Guidelines |

# Chapter 1: Introduction

## 1.1 Motivation

In a study on administrative communication, Felici and Griebel (2019, p. 168) suggest that "one of the main administration duties is to serve citizens and to speak their language; therefore, information accessibility should always be a priority when it is not a public duty." The Oxford English Dictionary offers two distinct but related definitions of "language" ("language, n.," 2008):

> 1.a. The system of spoken or written communication used by a particular country, people, community, etc.
>
> 2.a. The form of words in which something is communicated; manner or style of expression.

When Felici and Griebel (2019) put forward the assertion that a government has an obligation to speak the language of its citizens, they engage both definitions. They focus not only on the role of interlingual translation in a multilingual society, translation between different natural languages, but also on the role of intralingual translation, how the language surrounding complex topics can be brought to a level that average citizens can understand and utilize – plain language. However, that idea can and should be taken a step further if a government is to truly serve its citizens. Just as not all people in a society have the same knowledge of legislation, people in the same society also have a range of different reading and communication abilities. Even in plain language, what one person considers a straightforward text could present major comprehension difficulties to another, for instance someone with an intellectual disability (ID). So, how can administrative texts be translated into a language that reaches the most diverse readership of citizens possible? One possibility is another controlled language (CL), Easy-to-Read (E2R), which is designed to be just that: easy to understand by as many people as possible, including those with particular reading and processing difficulties.

Historically, authors have employed the terms "textual accessibility" or "text accessibility" to talk about inclusivity via simplified language (Drndarevic et al., 2012; Hassell, 2018; Rodríguez Vázquez, 2013, 2016), but in order to explicitly include the dual definition of language we will use the term "linguistic accessibility." The first component of linguistic accessibility, further developed in *Chapter 3*, can be mapped to OED definition (1.a.). Without this, the other factors of linguistic accessibility, which deal with the second definition – lexical, syntactic, and structural aspects of language, or form – do not matter. Consider a French text that is written using the simplest version

of the language possible. If a reader has no knowledge of French, the text is still not accessible. Linguistic accessibility should be prioritized in order for an administration to effectively deliver information in a way that is inclusive of *all* of those whom it governs – not just those who speak the country's official language fluently, and not just those who have what could be considered "average" reading and communication styles and abilities. Nearly every country in the world still has a long way to go if that is to be achieved. We focus our investigation on two countries in particular: Switzerland and France.

First, let us examine linguistic accessibility with regard to definition (1.a.). Information published by the Swiss government is generally provided in three of the four official languages: French, German, and Italian. Yet in 2018, English was the primary language of 6.6% (471,056 speakers) of the population in Switzerland, spoken by nearly as many residents as Italian (593,646 speakers) and roughly 13 times more residents than the fourth official language, Romansh (36,709 speakers) (Office fédérale de la statistique, 2020). In 2017, according to statistics published by the *Institut National de la Statistique et des Études Économiques* (INSEE), roughly 225,000 immigrants from primarily English-speaking countries were living in France (2020). France is also the world's top tourist destination, welcoming many English-speaking visitors each year (United Nations World Tourism Organization, 2020). The researcher's personal experience living in Switzerland and France revealed that despite prevalent and growing Anglophone populations, few administrative documents are made available in the English language. Professional translation can be costly, and governmental organizations do not always have room in their budgets to justify ordering translations for non-official languages. Consequently, free, online neural machine translation tools like Google Translate and DeepL are starting to play an increasingly important role in a foreign resident's inclusion in the civic life of their host country. Machine translation has made enormous strides of progress since its inception in the 1940s, and the newest generation, which makes use of neural networks and deep learning, is often able to produce results that are of good enough quality to be of use to average readers for informational purposes.

But sometimes "good enough" is not good enough for all. With respect to our second definition, (2.a.), Article 21 of the United Nations Convention on the Rights of Persons with Disabilities, ratified in France in 2010 and in Switzerland in 2014, ensures access to information for all "through the form of communication of their choice," which includes simplified language. Despite such legislation, individuals with disabilities are still often marginalized and numerous barriers to their full participation in society still exist, even in developed countries such as Switzerland and France (World Health Organization & World Bank Group, 2011). Felici and Griebel (2019) investigated

the extent to which plain language guidelines are respected in Swiss administrative texts in three of the four official languages in Switzerland and the role that human translation may or may not play in accessible communication in multilingual countries and contexts. They found that although plain language is said to be a priority in institutional communication, in practice, Swiss insurance leaflets feature far from optimal readability. And Easy-to-Read, though it has the potential to provide value to more citizens, is even less widespread than plain language. Whether real or perceived, the obstacles to producing E2R, which include money, time, awareness, and training, mean that few Swiss and French government publications are provided in Easy-to-Read, and those that are, are almost never translated into English.

In light of these facts, we will investigate the suitability of DeepL, a neural machine translation (NMT) system as a tool for generating linguistically accessible English versions of French Easy-to-Read texts. If it requires fewer resources for public and private organizations, associations, and companies to produce Easy-to-Read materials, it is reasonable to expect that more information would be made available in this accessible format. We therefore anticipate that English-speaking adults with intellectual disabilities living in these two countries, and others that produce information primarily in French but have large English-speaking populations, could benefit from the research carried out for this thesis. With this study, we will shed light on some of the challenges that arise from using free and public NMT technology to translate administrative documents published in an Easy-to-Read format from French into English. We also aim to examine one potential solution to these challenges, introducing a specific pre-editing step into the translation workflow, with the ultimate goal of improving barrier-free communication and inclusion for adults with intellectual disabilities.

## 1.2 Research context

With a few notable exceptions, there is generally a dearth of research combining controlled language (see *Chapter 2*), machine translation (see *Chapter 2*), and the different factors that make up linguistic accessibility (translation quality, readability, and accessibility; see *Chapter 3*). The machine translatability of controlled language using machine learning technology has been explored in healthcare (Rossetti, 2019) and technical communication (Marzouk and Hansen-Schirra, 2019) contexts, but never with a focus on accessibility for readers with disabilities, to the researcher's knowledge. Research on automatic text simplification using neural networks has also proliferated in recent years (Chen et al., 2017; T. Wang et al., 2016). The SIMPLES (*Simplification des Textes* Écrites) project addresses the issue of accessibility-driven E2R text production using machine

learning technology as well, proposing an automatic text summarization tool and an authoring tool for *Français facile à lire et à comprendre* (FALC), the French version of Easy-to-Read (Jacquet & Poitrenaud, 2019; Chehab et al., 2019). However, even these closely related studies do not address the multilingual component that we are interested in.

The observation of this research gap prompted Kaplan, Rodríguez Vázquez, and Bouillon (2019) to begin delving into the topic with their exploratory study on NMT of Easy-to-Read, which is expanded upon in *Section 3.5.2*. Their work, in addition to the researcher's personal family connection to and interest in the rights and inclusion of people with disabilities, was a direct inspiration to this thesis. In summary, the exploratory study compared English output from three different machine translation systems for: translation quality, measured using a manual translation error typology; accessibility, measured manually based on the Inclusion Europe E2R guidelines; and readability, measured automatically via the indices calculated by the Coh-Metrix tool. DeepL, a generic neural machine translation system, greatly outperformed its challengers, Google Translate and Yandex.Translate, but nevertheless produced some curious errors, including highly ungrammatical constructions. It is known from previous studies that NMT technology tends to score high in fluency ratings, meaning that it is unlikely to make such grammatical mistakes (Castilho et al., 2017; Neubig et al., 2015). Upon closer examination of the results of the exploratory study and of the *Information for All* standards published by Inclusion Europe, one E2R guideline in particular stood out as the possible culprit of a number of unusual phenomena in the target translations (2009, p. 17):

> 19. Keep your sentences short.
>
> Where possible, 1 sentence should fit on 1 line. You could do this by
>
> - writing only 1 idea per sentence
> - using a full stop before starting a new idea, instead of using a comma or an "and".
>
> **If you have to write 1 sentence on 2 lines, cut the sentence where people would pause when reading out loud.**

Indeed, translation errors and E2R violations in source sentences that had been split onto multiple lines manually with either a soft (↵) or hard return (⁋), in compliance with that guideline, were often varied and unpredictable. This discovery prompted us to narrow our focus, leading to the research goals and questions laid out in the following section.

## 1.3 Research goals, questions, and hypotheses

The main goal of this study is to continue the investigation into whether a free and public NMT system could be a worthwhile tool for generating linguistically accessible English translations from French source texts. We do so by targeting the sentences that are impacted by the line break guideline described in *Section 1.2* and exploring the effects of a pre-editing step on their quality, accessibility, and readability.

A four-step empirical process, outlined in *Section 1.4*, was designed in an attempt to answer three research questions:

> **Research Question (RQ) 1**: *How does the removal of line breaks from Easy-to-Read French-language administrative documents during the pre-editing process influence the* ***translation quality*** *of English output produced by a generic NMT system?*

This question aims to evaluate how well DeepL handles French-to-English translation with and without the peculiar E2R formatting challenge of line breaks. Predictions about the impact of pre-editing on translation quality were made and formulated as one general hypothesis statement and four more granular hypotheses:

> **Hypothesis 1.0**: *Removing forced line breaks from French Easy-to-Read texts will improve the quality of English NMT output.*
>
> > **Hypothesis 1.1**: *The segments that were pre-edited to remove manual line breaks prior to translation with DeepL NMT will contain* ***fewer translation errors*** *than the segments that contain manual line breaks.*
> >
> > **Hypothesis 1.2**: *The segments that were pre-edited to remove manual line breaks prior to translation with DeepL NMT will contain* ***less serious translation errors*** *than the segments that contain manual line breaks.*
> >
> > **Hypothesis 1.3**: ***Fluency*** *and* ***style*** *will be the two categories most positively affected by this pre-editing process.*
> >
> > **Hypothesis 1.4**: *The segments that were pre-edited (i.e. manual line breaks were removed) prior to translation will require* ***less post-editing effort*** *to achieve publishable quality than the segments that were not pre-edited, and that therefore contained manual line breaks, when translated with DeepL NMT.*

These hypotheses were tested in *Steps 1* (**H1.1-3**; *Section 4.3.1.1.1*) and *2* (**H1.4**; *Section 4.3.1.1.2*) of the study.

The second question we were interested in exploring relates to how accessible translated texts are for readers with ID, or in other words, to what degree they are E2R compliant.

> **Research Question (RQ) 2**: *How does the removal of line breaks from Easy-to-Read French-language administrative documents during the pre-editing process influence the* ***accessibility*** *of English output produced by a generic NMT system?*
>
> > **Hypothesis 2.0**: *Removing forced line breaks before performing NMT with DeepL will improve text accessibility.*
> >
> > > **Hypothesis 2.1**: *The segments that were* ***pre-edited*** *to remove manual line breaks prior to translation with DeepL NMT will contain* ***fewer violations of E2R guidelines*** *than the segments that contain manual line breaks.*

*Step 3* (*Section 4.3.1.2.1*) aims to test **H2.0** and **H2.1**.

Finally, our third research question and corresponding hypothesis statement (**H3.0**) deals with readability, measured quantitatively in *Step 4* (*Section 4.3.2.1*).

> **Research Question (RQ) 3**: *How does the removal of line breaks from Easy-to-Read French-language administrative documents during the pre-editing process influence the* ***readability*** *of English output produced by a generic NMT system?*
>
> > **Hypothesis 3.0**: *Removing forced line breaks before NMT will* ***improve readability****.*

## 1.4 Materials and methods

Since no Easy-to-Read corpora that would fit the needs of our study existed at the time of writing, a corpus of 41 French-language Easy-to-Read documents was assembled. Continuing on the same trajectory as Kaplan et al.'s (2019) exploratory study, and because of the importance of clear and accessible administrative documents for civic inclusion highlighted by Felici and Griebel (2019), all of the texts selected were published by or in collaboration with Swiss and French federal agencies and relate to the rights and inclusion of adults with disabilities in their community.

We analyzed the effect of pre-editing on NMT output of these documents, with a particular focus on the line break guideline of Inclusion Europe's *Information for All* standards. DeepL Translator, a free and public online neural machine translation tool, was selected to translate the corpus from French to English, because findings from the aforementioned exploratory study show that this tool produced the best results in terms of number of translation errors and Easy-to-Read guideline violations (Kaplan et al., 2019). Each sentence in the corpus containing one or more line breaks was translated twice: once with the E2R line formatting intact and once with line breaks removed. Then, three manual analyses were performed on the resulting translations. One relied on the DQF-MQM error typology model to measure translation quality. A second, which invited a small group of translators to participate in a post-editing task, allowed us to evaluate translation quality from another perspective. A third used Inclusion Europe's set of guidelines to identify accessibility violations. Finally, one automatic analysis was performed to measure various indices of readability.

## 1.5 Thesis structure

Following this introduction (*Chapter 1*), an overview of the main topics addressed in this thesis, as well as a review of the literature, will be provided in order to situate this work within the broader research context. *Chapter 2* surveys the fields of controlled language, which can be mapped to definition (2.a.) of the OED definition of language, and machine translation, which deals with definition (1.a.). It demonstrates how the two subfields studied in this thesis, Easy-to-Read and neural machine translation, came to be, how they relate to the other types of CL and MT that exist, and why they were chosen for this work. The third broad field incorporated in this work is linguistic accessibility, which is broken down into three measurable factors and defined in *Chapter 3*, along with the current state of the art of the intersections of these three areas of research.

*Chapter 4* details the methods and materials employed in the four-step experiment that tested the translation quality, accessibility, and readability of French-to-English neural machine translations of administrative documents written in *Français facile à lire et à comprendre*, the French version of Easy-to-Read, before and after forced line breaks were removed from the texts.

The results from the aforementioned experiment are presented and discussed in *Chapter 5*. Finally, a brief conclusion in *Chapter 6* summarizes the work and its intended contributions to the field, and introduces related avenues of research that could be explored in the future.

# Chapter 2: Controlled Language and Machine Translation

## 2.1 Introduction

This thesis brings together three broad fields of research in Translation Studies that, to our knowledge, have rarely before been linked within a single project. In this chapter, the concepts of controlled language and machine translation will be explored and defined in order to provide a solid theoretical foundation grounded in scholarly literature for *Chapter 3*, which introduces components of the third field, linguistic accessibility, and *Chapter 4*, which describes the methodological basis of this thesis. Following this short introduction (*Section 2.1*), we present the idea behind and primary purposes of controlled language; synthesize the main machine- and human-oriented controlled languages that have emerged in English- and French-speaking contexts up to this point; and outline the development, characteristics, applications, and limitations of Easy to Read (or *Facile à lire et à comprendre* in French), the focus of our investigation (*Section 2.2*). *Section 2.3* provides a brief history of machine translation, establishes some key features of neural machine translation, and introduces DeepL Translator, the system that was used to generate our bilingual corpus. In *Section 2.4*, we present the current state of research combining the two fields explored throughout the rest of the chapter.

## 2.2 Controlled language

This section provides a comparative view of the history and evolution, goals, usage, and specifications of several restricted subsets of the two natural languages relevant to this study: French and English. Due to the exhaustive nature of controlled languages (CL), particularly in English, we only address a small selection in order to provide context for Easy to Read, which is examined more thoroughly in *Section 2.4.3*.

### 2.2.1 Defining controlled language

It is difficult to pin down a single definition of CL, as well as what does and does not qualify as one. However, researchers seem to agree on one overarching goal of CL – improved comprehension – and two ways by which that goal is achieved: reducing ambiguity by limiting the language on a lexical level (e.g. vocabulary) and reducing complexity by imposing syntactical constraints (e.g. sentence length, accepted grammatical constructions). There is an important

distinction to be made between natural sublanguage, spontaneous restricted usage of a language in a specific field or area of specialization, and controlled language, language restrictions that are designed and applied deliberately (Kittredge, 2003). Since a controlled language is consciously created, Wyner et al. (2010) identify three levels of properties – generic, design, and linguistic – which translate to different questions that developers must ask themselves, and which can help us further define a given CL. Questions which range from the very broad, such as: "Who are the intended users?" and "What are the purposes?" to the extremely precise, for instance: "What sorts of subordinate clauses are supported?" (Wyner et al., 2010, pp. 283–286). To answer those first generic questions, CL designers must think about whether they intend to enable human-human communication or human-machine communication (Wyner et al., 2010).

### 2.2.2 Branches and applications of controlled language

Both Huijsen (1998) and Nyberg et al. (2003) note that classifying CLs as one or the other can be challenging because a controlled language that was designed for one purpose often fulfills the other. In fact, one study's findings suggest that it is possible to develop a CL that improves both translatability and readability (Reuther, 2003). This thesis further blurs the lines by applying machine translation technology to a CL that has thus far been strictly human-oriented. Nevertheless, attempting to characterize CLs can be helpful for situating them within the landscape of language tools and understanding the possibilities and limitations of their applications.

#### 2.2.2.1 Machine-oriented controlled language (MOCL)

The primary goal of machine-oriented controlled language (MOCL) is to improve "understanding" by machines, i.e. translatability and other computational processing.[1] Take for instance Jurafsky and Martin's (2009, p. 4) example of an ambiguous sentence: "I made her duck." Processing this seemingly simple sentence is a complex task for a machine, because it has at least five distinct possible meanings in English. Applying a controlled language that implements restrictions such as a one-to-one correspondence between form and meaning could reduce, if not eliminate completely, the risk of an incorrect interpretation. Whether the result is a translation that is nearly ready to deploy as soon as it comes out of the machine – in other words, fully automatic, high-

[1] This is not to say that translatability cannot be the primary goal of some human-oriented controlled language (HOCL); take for instance Minimal English, which has roots in natural semantic metalanguage and breaks English down into "semantic primes," the most basic universal ideas that cannot be further reduced, to ensure optimal intercultural exchange (Goddard, 2018). Other HOCL with primary or secondary goals of translation are explored in *Section 2.2.2.2*. and *Section 2.4* presents one study that focused specifically on a CL and how it was handled by a neural machine translation system.

quality translation, introduced in *Section 2.3.1* – or a translation that needs full post-editing often depends on how strictly or loosely defined the CL used to author the text is. KANT Controlled English is a well-known example of a strictly defined CL, which was developed along with its own machine translation (MT) system and has been used for technical translation (Mitamura, 1999). A lesser-known strict CL is PENG (Processable English), which was designed not for interlingual translation but rather for formal representation, and "designed for writing unambiguous and precise specifications" (Schwitter, 2002). On the other end of the spectrum, loosely defined CLs like Perkins Approved Clear English (PACE), which features much more vague rules, such as, "Order the parts of the sentence logically," but which still resulted in significantly faster post-editing when compared to "conventional translation," have also been successfully deployed (Pym, 1990 as cited in Nyberg et al., 2003, p. 256).

#### 2.2.2.2 Human-oriented controlled language (HOCL)

In contrast to machine-oriented controlled languages, which generally serve one main purpose – reducing linguistic ambiguity for more accurate computational processing – human-oriented controlled languages (HOCL) are more focused on comprehension for humans and may be further categorized by the goal that developers created them to achieve. Some of these goals include **assisting foreign-language learners** and promoting intercultural communication via international auxiliary languages; **reducing misunderstanding in technical writing and translation** to ensure safety; **facilitating interactions in business and administration**, particularly in government-to-citizen communication; and **improving accessibility** for people with disabilities and other conditions that cause reading difficulties.

The way that controlled language came about, and perhaps one of its most intuitive objectives, is within the context of **language learning and literacy**. The earliest documented CL is thought to be Charles Kay Ogden's Basic (British American Scientific International Commercial) English, described in his book that was published in the UK in 1930. Ogden argued that everything one could possibly want to express can be achieved with a very limited amount of English vocabulary, just 600 to 1,000 words, only 18 of which are verbs. Basic English blossomed from a desire to give people around the world a simple, and quick-to-learn basis for communication in English, an international auxiliary language of sorts (Kuhn, 2014; Nyberg et al., 2003; Ogden, 1930). Kuhn (2014) and Nyberg et al. (2003, p. 250) seem to disagree on the actual application of the CL, the former asserting that it is still in use today and citing several published Basic English texts and the later calling it "a mere curiosity, unsuitable for any practical purpose," and claiming that it has "never been widely used". In any case, it has inspired many more controlled languages; Kuhn

(2014) mentions just under 50 primarily HOCL in his 2014 survey and classification, and over 50 more MOCL. Not mentioned in his Anglo-centered classification is *Français élémentaire*, which later become *Français fondamental*, developed in the 1950s due to a "concern over the declining role of French as a world language" (Stern, 1983, p. 55). This French CL is considered an offshoot of Basic English – though chiefly a way to teach the language, not a means of international communication – no matter how ardently the authors opposed the association in response to the harsh criticism they received when they described it as such (López, 2006). Another interesting feature of *Français fondamental* is that it is originated from a corpus of spoken French rather than written texts (Cortier, 2006).

Kittredge's (2003, p. 441) definition of controlled language – one of the many attempts to delineate the concept – specifically mentions the second of those four goals, asserting that restricted language is often used for "**writing technical documentation** for non-native speakers of the document language." He (as well as Nyberg et al. (2003)) cites safety and liability concerns as one advantage to using clear and simplified language, as well as ease of translation – when translation is even required, since another advantage is that CLs in the manufacturing and service industries can sometimes eliminate this need by simplifying language for non-native English-speaking workers. Reducing the amount of time and effort it takes to get a product to market can also have a positive impact on cost and competitiveness (Huijsen, 1998). Two of the most influential and well-documented are Caterpillar Fundamental English (CFE), which later evolved into Caterpillar Technical English (CTE), and AECMA (from the French *Association Europeene des Constructeurs de Materiel Aerospatial*) Simplified English, which became ASD Simplified Technical English (ASD-STE) following a merger. CFE was first deployed by the American equipment manufacturer in 1971, making it likely the oldest CL curated specifically for industries (Wojcik & Hoard, 1997 as cited in Kuhn, 2014).

Like Basic English, by which it was greatly influenced, CFE was mainly targeted toward Caterpillar's non-native English-speaking mechanics, but was abandoned in 1982 because it was no longer practical. About a decade later, the company reevaluated the purpose of CL in their processes and shifted their focus to reducing translation costs as opposed to eliminating the need for it altogether by trying to make English the default language of communication. The result was CTE, which was more similar to KANT Controlled English than Basic English (Kuhn, 2014). ASD-STE was introduced in 1983 with the same main goals as CFE and CTE, and is still in use today, primarily by aerospace manufacturers and airlines (Kittredge, 2003). These are major industries in France, so it is not surprising that a version of this CL was also implemented for the

French language. GIFAS (*Groupement des Industries Françaises Aéronautiques et Spatiales*), a French partner in the original AECMA project, began working on *Français rationalisé* (Rationalized French) in 1985 to facilitate translation to and from the English simplified language and to increase clarity for employees. But the task of developing one CL on the basis of another was more complex than the authors anticipated, and the project took 12 years to complete. Barthe et al. (1999, p. 228) raise an interesting question about Simplified English (SE) and *Français rationalisé* that is pertinent to Easy to Read since it, too deals with multiple EU languages: "Was working backward from the SE guide a satisfactory method, or should we have worked only on the basis of the corpus of French documents we collected, independently of SE?"

HOCL are also employed for **business and administration** purposes, and some people have even advocated for controlled language in legal contexts, a field that has traditionally produced non-lay-friendly writing.[2] The terms "plain language" and "plain English" have been discussed in various English-speaking governmental contexts since the 1940s, beginning with *The Complete Plain Words* (later *Plain Words*), a style guide intended to eradicate "officialese," written by a civil servant based at the request of the UK Treasury Department (E. Gowers, 2014). Presidents Nixon and Carter introduced the idea in the US in the 1970s, however their executive orders lacked substance, calling for the use of "layman's terms" and "plain English" but not providing guidance as to how that should look or how to go about it. The Clinton administration (1993-2001) built on its predecessors' progress and specified four very broad components of plain language (common words, active voice, direct address, and short sentences). It also began rewarding government employees who complied particularly well with these components in their written communication. The vague idea of plain language was finally fleshed out into more concrete standards following the adoption of the Plain Writing Act of 2010 under President Obama.[3]

In the Francophone world, ministries of the French and Quebecois governments also published similar, and quite extensive guides to writing simplified (i.e. in language that the general population can understand) administrative documents in the 2000s (COSLA, 2008; Savard et al., 2003). Despite these and other initiatives such as the Plain English Campaign in the UK, Harper and Zimmerman (2009) found that interpretations of what "plain language" actually means, even among people within the same organization, still vary wildly. However, this may have improved in recent years, since Kuhn (2014) does consider US Federal Plain Language Guidelines (although,

[2] See Bryan Garner's *Legal Writing in Plain English* (2002).
[3] https://plainlanguage.gov/ Last accessed: August 3, 2020

curiously, not the "How to Write Clearly" set by the European Commission, from the same year) to be a strict CL in his classification.

The final overarching goal of CL that we will address is **accessibility**. Although most HOCL aims to reduce ambiguity and complexity, and improve understanding between the person or body authoring a text and the people receiving it, one subset of the population that is particularly affected by these very factors is often left out of discussion around CL: people with different reading needs and other disabilities. This could be because the line between style guide and controlled language can be fuzzy; according to Kuhn (2014, p. 124), qualification "depends on whether the style guide defines a new language or whether it merely describes good practices that have emerged naturally." Nevertheless, a multitude of guides on accessible communication have been set forth, particularly within the last 20 years, which some would consider accessibility-focused CLs.

### 2.2.3 Easy to Read/*Facile à lire et à comprendre*

Although it does not appear in the predominant literature on the topic of controlled language, we take the position that Easy to Read falls into that final, narrow branch because it meets all four of the conditions in Kuhn's (2014, p. 123) long definition of controlled natural language:

> 1. It is based on exactly one natural language (its "base language").
> 2. The most important difference between it and its base language (but not necessarily the only one) is that it is more restrictive concerning lexicon, syntax, and/or semantics.
> 3. It preserves most of the natural properties of its base language, so that speakers of the base language can intuitively and correctly understand texts in the controlled natural language, at least to a substantial degree.
> 4. It is a constructed language, which means that it is explicitly and consciously defined, and *is not* the product of an implicit and natural process (even though it is based on a natural language that *is* the product of an implicit and natural process).

First, we can consider that each translation of the Inclusion Europe guidelines are indeed based only on one natural language; Easy to Read is based on English and has slightly different rules than *Facile à lire et à comprendre*, which is based on French (which in turn differs somewhat from *Lectura fácil* (Spanish), *Leichte Sprache* (German), *Leitura fácil* (Portuguese), *Facile da leggere* (Italian), etc.). Next, it dictates restricted lexicon (there is no dictionary of acceptable/unacceptable terms, but it does stipulate no difficult words, no words in other languages, no abbreviations, no contractions, etc.), syntax (e.g. present tense, active voice, no ambiguous pronouns), and semantics (e.g. natural line breaks, bullet points). Third, because it was created for English speakers with

specific reading needs, we can assume that "speakers of the base language can intuitively and correctly understand" it, despite some research showing negative reception by typical adults (*Section 2.2.3.2*) (Kuhn, 2014, p. 123). And lastly, the "good practices" Kuhn (2014) refers to were consciously developed, although as some scholars have pointed out there is a notable lack of empirical evidence for various forms of easy-read, and guideline designers often do not cite references or provide information about how they came to the conclusions they did (Fajardo et al., 2014; Sutherland & Isherwood, 2016).[4]

#### 2.2.3.1 History and development of Easy to Read

The international non-profit International League of Societies for the Mentally Handicapped – European Association (ILSMH-EA) was founded in 1988 in Brussels to represent Europe as a full member branch of Inclusion International, a disability advocacy network created in 1960.[5] ILSMH changed its name to Inclusion Europe in 2000,[6] but not before publishing "Make it Simple: European Guidelines for the Production of Easy-to-Read Information for People with Learning Disability" in 1998. The guide outlined best practices for writing and presenting information in a such a way that as many people as possible are able to understand. In it, the authors contend that implementing Easy to Read can contribute to breaking down the "information rich"/"information poor" barrier that exists in our society because of the way that information has typically been presented (Freyhoff et al., 1998). Despite this and other initial steps taken by European and international organizations such as Mencap and the International Federation of Library Associations and Institutions (IFLA) in the 1990s and early 2000s, it was found that the current work existed "without any common ground, regulation or quality standards" (CARDET, 2014, p. 28).

In response to a need for more formal and standardized guidance for producers of accessible education materials, and within the framework of the first phrase of the Pathways to Adult Education for People with Intellectual Disabilities project, funded by the European Commission as part of the Lifelong Learning Programme, Inclusion Europe led nine partners from eight different countries in the creation of European standards for Easy-to-Read informative documents (2009). The project also "developed and tested a methodology on how to involve people with intellectual disabilities in preparation and quality control of accessible adult education material" (CARDET, 2014, p. 28). The result of the first phase of Pathways was four brochures outlining

[4] Cf. Ruel et al. 2011, a Canadian accessibility guide that cites extensively.
[5] http://www.siwadam.com/hmm/euie.htm Last accessed: August 3, 2020
[6] https://www.inclusion-europe.eu/about-us/ Last accessed: August 3, 2020

best practices not only for producing Easy-to-Read information ("Information for all"), largely based on ILSMH's original publication, but also for training lifelong learning staff about writing (Training lifelong learning staff") and teaching accessibly ("Teaching can be easy"), and for involving people with disabilities in the writing process ("Do not write for us without us").

Our research is focused on the guidelines set forth in the "Information for all" brochure, and the definition of Easy-to-Read (E2R) that we have adopted for this thesis is based on that document. The guidelines have been published in 16 of the 24 official EU languages, and a handful of guidelines are language specific, an important point to consider when analyzing the results of the accessibility study presented in *Section 3.3.2*. Two guidelines appear in English but not in French, which recommend writing in the present tense instead of the past tense where possible and avoiding contractions (Inclusion Europe, 2009, pp. 22–23). Due to the slight differences in the two sets of guidelines, we make the distinction between French texts written specifically in accordance with the French rules ("*Facile à lire et à comprendre*," or FALC) and English texts as ("Easy to Read," or E2R). The first two sections of the document provide general standards and standards for written information, but the brochure also addresses electronic, video, and audio formats. The guidelines in these two sections can be further broken down into several broad categories: general (e.g. appropriate language for the target audience), structural (e.g. bullet points instead of comma-separated lists), word-level (e.g. no figures of speech, consistent terminology throughout), and sentence-level (e.g. active instead of passive constructions when possible). The full list that was used for the annotation described in *Chapter 3* can be found in *Appendix A*. In addition to providing instructions for how to write in Easy-to-Read, unlike the 1998 ILSMH version, the 2009 publication itself is written in Easy-to-Read, allowing the target population to access the standards that impact them and participate more easily in the document creation process.

The line break guideline that we have chosen to focus on (introduced in *Section 1.2*) was also present in the original version, but the wording changed slightly to reflect E2R standards, from: "Try to put one sentence on one line. If this is not possible, try to have separate clauses on separate lines or break the sentence into separate lines at the points where people would naturally pause," to: "Where possible, 1 sentence should fit on 1 line. / If you have to write 1 sentence on 2 lines, / cut the sentence where people would pause / when reading out loud," where forward slashes represent line breaks (Freyhoff et al., 1998; Inclusion Europe, 2009). Unlike some other recommendations that seem to be universally accepted in other accessibility style guides and CLs, for instance to use well-known words, the line break guideline only appears in a handful of other publications. For

example, the "Information for All" booklet produced by the Norah Fry Research Centre in the UK: "If a sentence has to run onto a second line, try to break it after words like 'and' or 'but' so it is not read as two separate sentences. Or better still, make a new sentence" (Mears et al., 2004). A similar suggestion was put forth by Nomura et al. (2010, p. 11) in their revised version of the IFLA guidelines: "Words of a single phrase should fit on a single line, i.e. each sentence should be broken off at a natural speech break."[7] Ruel et al. (2011) echo this sentiment: "Mettre les mots d'une phrase sur la même ligne et la même page. Si vous ne pouvez pas, essayez de diviser la phrase après les conjonctions « et, mais », car elles indiquent une pause naturelle." However, most other notable publications on accessible communication by organizations like People First New Zealand, CHANGE, NHS England, and Mencap, do not include any restrictions on line breaks, which raises the question of the influence of this text feature on readability (CHANGE, 2016; Mencap, 2000; NHS England, 2018; People First New Zealand, 2017).

FALC publications have been created for a growing number of cultural and touristic spaces in France and Switzerland, such as the Centre Pompidou and the Musée d'Orsay in Paris (Lamotte & Therwath, 2016) but also smaller museums like the Musée d'Evreux and the Musée International de la Croix-Rouge et du Croissant-Rouge in Geneva, and the Clermont-Ferrand rugby stadium. *Figure 2.1* shows an example of what FALC can look like for cultural contexts. Some health professionals have also adopted FALC, most notably SantéBD, which has produced over 50 fact sheets (as well as about 25 in English E2R) about a variety of medical topics, and Santé Très Facile.[8] The focus of this thesis is on FALC and E2R used for official communication by governmental agencies in compliance with Article 21 of the United Nations Convention on the Rights of Persons with Disabilities (ratified in France in 2010 and in Switzerland in 2014), which ensures access to information for all "through the form of communication of their choice," and includes "plain language."[9] The corpus of documents studied is described in detail in *Section 3.2*.

[7] We were unable to obtain a copy of the 1997 original guidelines to verify whether they were added before or after the Inclusion Europe set was published.
[8] http://www.santetresfacile.fr/infos_sante Last accessed: August 3, 2020
[9] https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html Last accessed: August 3, 2020


*Figure 2.1: Excerpts from the Musée D'Orsay's FALC visitors' guide to the "Paris au 19e siècle" exhibition*

#### 2.2.3.2 Limitations of E2R/FALC

Application of E2R has the potential to improve the inclusion and quality of life not only of people with disabilities but also of foreign-language learners, people with low literacy levels, and senior citizens across European countries and the world. Yet despite the (albeit limited) empirical evidence suggesting that people with disabilities are indeed aided by easy language (see Karreman et al. (2007) user study on adapted web pages[10] and Fajardo et al. (2014) user study on reading comprehension[11]), and the fact that clear communication benefits everyone in society, information in E2R and FALC is still **not widely available**, particularly at the federal level. Very few reports that we are aware of have measured the prevalence of E2R or FALC specifically. One 2016 report by the European Union Agency for Fundamental Rights on disability and the EU migration situation found "little evidence of accessible information, for example, in easy-read format for persons with intellectual disabilities" in the seven Member States it examined (FRANET, 2016). Luce (2018) backs up this finding in her report on asylum seekers and refugees with intellectual disabilities (ID) in Europe, noting that her research did not turn up any E2R handouts, leaflets, or other documents related to asylum application. The most comprehensive report was produced by Holken Consultants & Partners as a first step in the *Simplification des Langues Écrites* (SIMPLES) project (see *Section 3.5.2*), and found that France produces few FALC documents, noting, "Bien que la simplification du langage administratif fasse l'objet de plusieurs initiatives, nous remarquons, cependant, qu'elle n'est pas encore démocratisée" (Chehab et al., 2019, p. 15). Our observations

[10] This study was carried out in 2007 before the Inclusion Europe guidelines were developed, however it used the very similar ILSMH guidelines.
[11] This study was performed with Spanish texts written according to IFLA guidelines, which feature a large amount of overlap with E2R guidelines.

during the compilation of documents for the corpus used in this thesis (see *Section 3.2*) also support this conclusion. For instance, all French adults with disabilities who have a legal guardian, and who had previously been denied the right to vote because of it, gained that right in 2019 (Swiss citizens under full guardianship are still not allowed to vote, despite Switzerland's 2014 ratification of the UN CRPD). Yet in the 2019 European Parliament election, although all political parties running to represent France were encouraged to provide a FALC version of their campaign platform (*profession de foi*) (Ministère de l'intérieur, 2018), we observed that only nine out of 36, or one quarter, did so.[12,13]

**Cost of production** both in terms of human and financial capital – or even misconceptions about it – is one factor that could impact the limited amount of E2R information currently available. A representative from the France Insoumise party, which did have its campaign materials translated into FALC, told reporters that she believed it would be much more expensive than it was (reportedly less than 500€ for three documents) (Hennequin, 2019). On the other hand, over 60 people with intellectual disabilities participated in the "Orsay facile" project to publish two FALC guides for the Musée d'Orsay art museum (Lamotte & Therwath, 2016). Due to the extensive steps involved, e.g. target audience analysis, selection of most important information, simplification, formatting, testing and validation by members of the target population, one prominent Swiss translator charges a minimum of CHF 4.00 per 55 characters (roughly one line or 8-11 words) to produce FALC versions of standard texts, CHF 5.50 for more complex texts, and custom pricing for technical documents.[14] Consequently, access to resources, both money and labor, could be a valid limitation for smaller companies or organizations that are interested in producing accessible content. When it comes to federal agencies, implementation of E2R/FALC may rely on prioritization of and advocacy for accessibility in budget and policy negotiations.

Another barrier to the dissemination of E2R is **mixed or poor reception** and the **stigma** that seems to surround it, primarily by people who do not rely on this type of simplified language to consume information. In Germany, where easy language is much more widely established than in France or Switzerland, a 2017 attempt to generalize the communication about an upcoming election in easy language only (rather than provide it upon request to those who need or prefer it) failed remarkably, provoking responses of "upheaval, incomprehension and alienation" (German

[12] *Profession de foi* documents for this election were available at https://programme-candidats.interieur.gouv.fr/, last accessed May 20, 2019, however the page has since been updated to reflect information about the most recent local elections (for which it appears no candidates published FALC statements).
[13] Although outside the scope of this study, it is worth mentioning that easy language has been considerably further developed and deployed in some other countries, such as Germany and Finland.
[14] https://www.textoh.ch/traduction/ Last accessed: August 3, 2020

State Parliament website, 2017, as cited in Maaß, 2019, p. 5). A user study on web accessibility that found that people with and without ID both benefitted from an E2R version of a webpage also found that the adapted page was negatively perceived by the group without ID. That group happened to be made up of people in community with individuals with disabilities, making these findings on satisfaction all the more surprising and significant (Karreman et al., 2007). Another experiment comparing Easy-to-Read, Elaborated Plain Language, and standard medical informed consent procedures found the opposite of what they hypothesized: younger adults actually understood standard information better than E2R, and performed best with Elaborated Plain Language (Schatz et al., 2017). Findings by Vollenwyder et al. (2018) did not corroborate either of these studies, showing no significant decrease in either text "liking" or understanding of E2R as compared to standard language. However, it is important to note that of those three studies, only one (Karreman et al., 2007) included participants belonging to the main population for whom E2R was founded: people with cognitive or intellectual disabilities. Interestingly, another scholar criticized the CL for being *too* inclusive, which can reinforce stigmatization. A corpus study on Spanish newspaper articles, drafting manuals, and websites promoting *Lectura fácil* (LF) found that immigrants were often grouped together with people with intellectual disabilities as potential recipients of LF (Becker, 2019). Especially considering that they are not included as part of the text production or validation process, Becker (2019, p. 9) believes this to be detrimental to their status in the receiving society:

> In my view, the inclusion of 'immigrants' among the users of Easy-to-Read can do more harm than good to the representatives of this heterogeneous group. […] the propagation of 'immigrants' as Easy-to-Read users reinforces the already rather negative and homogenised image of this group with other attributes of marginalised communities […] The progressive degradation of the image of immigrants serves the motives of exclusion, otherisation and marginalization.

Christiane Maaß (2019) points to layout as a major cause of stigma around Easy Language (EL), the term she proposes for "maximally comprehensibility enhanced" language such as E2R. Text design conventions such as pictures help readers perceive, or recognize, a text that is written in easy-to-understand language. But at the same time, they identify, or "out" readers who have a disability that affects communication. The author also cites a reception study performed with four target user groups of EL; the study found that German people ages 65 and older reacted mostly negatively out of all of the groups to even being considered readers of Easy Language, despite clearly benefitting from this form of accessible communication (Gutermuth, 2020 as cited in Maaß, 2019).

Finally, a **lack of training and awareness** of the standards that exist could also partially explain the scarcity of accessible texts. The SIMPLES preliminary analysis also reported that some of the organizations that were contacted cited difficulty in adhering to all of the rules of FALC text production as a reason why they do not use it (Chehab et al., 2019). Though initiatives like the European Commission European Disability Strategy reference the need for web accessibility standards to be applied and information in easy-to-read formats to be provided, they generally do not prescribe actionable steps (European Commission, 2017). A systematic literature review on the topic of easy-read for people with ID, conducted in 2013-14, neglected to even include the latest Inclusion Europe guidelines, despite them being presented by the creators as a European standard, citing instead the ILSMH version (as well as several of the other resources mentioned in *Section 2.2.3.1*). The review critiqued existing guidelines based on the lack of transparency of the methodology used to develop them, the absence of a hierarchy that would place more importance of the recommendations that most impact accessibility, and inconsistencies across guidelines (Sutherland & Isherwood, 2016).[15] That being said, a small number of translation agencies and disability advocacy groups, such as UNAPEI, specialize in producing and translating E2R and FALC publications, and projects like SIMPLES, discussed in more detail in *Section 3.5.2*, are working to create tools to make it easier for individuals to write texts that comply with E2R guidelines. If neural machine translation of E2R/FALC ever becomes a viable option for text production – a possibility that we begin to explore with this research – it could address all of the previously mentioned factors: financial burden, perception and normalization of accessible communication formats, and resource limitations.

## 2.3 Machine translation

The second major theme of this thesis is machine translation. In this section we will define and discuss the major developments in machine translation, from rule-based to statistical to neural, as well as present the NMT system used for this research (*Section 2.3.4.1*).

### 2.3.1 Defining machine translation

Hutchins (2005, p. 501) defines machine translation as "computerized systems responsible for the production of translations with or without human assistance." The process of interlingual translation – the transfer and rendering of meaning from source language to target language – can

[15] This and other literature raises the question of the effectiveness and worth of E2R itself, of which there is a notable dearth of scientific research (Hurtado et al., 2014; Ignacio Madrid et al., 2012; Sutherland & Isherwood, 2016), although that is outside of the scope of this thesis.

be considered on a spectrum, as shown in *Figure 2.2*: at one extremity lies pure human translation, at the other lies fully-automated high-quality machine translation (FAHQT), and in between are various degrees of human and machine interaction.

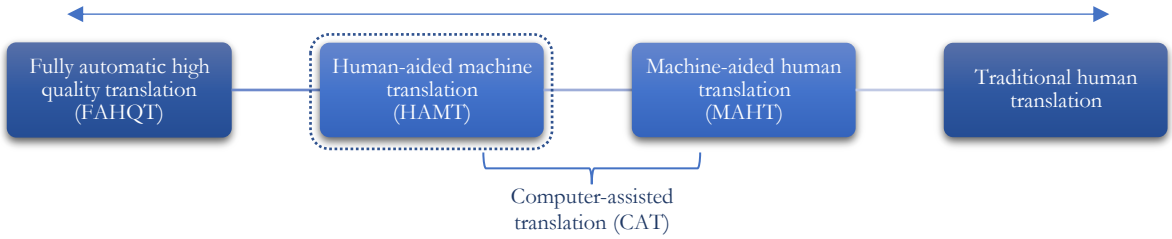



*Figure 2.2: Continuum of human-machine cooperation in translation from Hutchins & Somers (1992, p. 148)*

Both extremes are quite uncommon in practice today. In fact, Castilho et al. (2018, p. 27) go so far as to assert that "the traditional separation of human and machine is no longer valid, and drawing an arbitrary line between human translation and MT no longer serves us in research, teaching and professional practice." In 2020, rare is the translator who does not employ the spelling and grammar check feature in their word processor or an electronic dictionary or glossary, at the very least. Similarly, despite linguist Yehoshua Bar-Hillel's prediction in the 1950s that FAHQT would never be possible and his subsequent argument that researchers should not make it their ultimate goal, very few systems, such as the one developed in 2013 to translate the Swiss avalanche bulletin or the TAUM-Météo system that has been translating daily weather reports in Quebec since 1976, have gotten very close to – some would argue that they have reached – the opposite end of the aforementioned spectrum (W. J. Hutchins & Somers, 1992; Quah, 2006; Winkler et al., 2014). The middle of the continuum is often called computer-assisted (or aided) translation (CAT). It encompasses machine-aided human translation (MAHT), translation performed by humans with support from computerized tools such as translation memory software, and human-aided machine translation (HAMT), translation performed by machines with some amount of human intervention, for instance preparing a text before it is run through the machine (pre-editing) or cleaning a text up after it comes out of the machine (post-editing) (Hutchins & Somers, 1992). As indicated by the dotted outline in *Figure 2.2*, this thesis focuses on the latter.

### 2.3.2 Machine translation history

It is useful to have an understanding of the historical backdrop against which the developments in machine translation outlined in *Section 2.3.3* occurred. Arnold et al. (1994) propose a playful analogy summarizing the ups and downs of the first four decades of machine translation:


*Figure 2.3: "Machine Translation and the Roller Coaster of History" from Arnold et. al (1994)*

Most translation historians agree that the idea of (or at least serious discussion about) using machines to perform translation tasks originated in the late 1940s when Warren Weaver, an American scientist and mathematician working for the Rockefeller Foundation, penned the "Translation" memorandum. The memorandum described the possibility of "decoding" text written in a foreign language using computers, and it spurred the first steep incline in the roller coaster track. Major investments were made and research groups in the US and Europe began working on the topic, resulting in the very first rule-based machine translation (RBMT) systems (see *Section 2.3.3.1* for more information on this type of MT). However, the US government, which had dedicated multiple millions of dollars to research efforts, and other financial backers soon began feeling dissatisfied with the amount of progress being made and the prospects of getting a return on their investment. This suspicion, coupled with vocal criticism within the research community, notably Bar-Hillel's remarks invoked in *Section 2.3.1*, led to the crest of the first roller coaster hill: a commissioned report written by the Automatic Language Processing Advisory Committee (ALPAC). As Arnold et al. (1994, p. 13) put it,

> The ALPAC Report was damning, concluding that there was no shortage of human translators, and that there was no immediate prospect of MT producing useful translation of general scientific texts. This report led to the virtual end of government funding in the USA. Worse, it led to a general loss of morale in the field, as early hopes were perceived to be groundless.

Support was pulled and research ground to a near halt, represented by the roller coaster's precipitous descent around 1964. Interestingly, some of the research that did persist over the following decade was funded by the Mormon Church in the US state of Utah for the religious sector, which also happens to be one of the first recorded contexts for *human* translation in history (Ballard, 1992; Slocum, 1985). Hutchins (2005) cites the TAUM group's project in Montreal and the adoption of the SYSTRAN system by NASA in 1974 and the European Commission in 1976 as turning points in MT history – the roller coaster's second ascent. In the 1980s, major developments were made in RBMT systems, which rely on dictionaries and linguistic information. Japanese companies capitalized on a growing personal computer market to develop integrated systems, and near the end of the decade computerized tools like translation memories helped increase translator productivity (Hutchins, 2005). The 1990s brought about a new approach: corpus-based models. In contrast to the rule-based systems that dominated the 1980s, corpus-based systems, including statistical machine translation (SMT; see *Section 2.3.3.2*), use data from existing source text-target text pairs called parallel corpora (Somers, 2005). Both RBMT and SMT will be explored in more depth in *Section 2.3.3*. As the ride enters "The Future," leaving the confines of Arnold et al.'s (1994) track, its trajectory only steepens. The last six years in particular, which have brought many developments in neural machine translation technology – a new generation of statistical machine translation and the topic of *Section 2.3.4* – have convinced us that the figurative roller coaster of MT shows no sign of descent any time soon.

### 2.3.3 The evolution of machine translation

This section will present a brief explanation of the characteristics of the main MT models that exist, direct machine translation systems (*Section 2.3.3.1*), rule-based machine translation (*Section 2.3.3.2*) and corpus-based machine translation (*Section 2.3.3.3*), in order to show how we have arrived at the systems currently dominating MT research and development: a branch of corpus-based MT that employs neural networks (*Section 2.3.4*). It should be noted that this information has been simplified for concision, and many hybrid approaches also exist.

The first and second generations of machine translation technology first emerged in the 1950s and peaked in the 1980s and 1990s, as explained in *Section 2.3.2*. The first generation is known as direct machine translation; the second is known as indirect or rule-based machine translation and comprises two main classifications of system design: transfer and interlingua. All three system architectures operate using varying amounts and types of linguistic information that must be coded by humans. In other words, the systems must be programmed to "know" the rules and have the

necessary linguistic knowledge that a human translator would use to first understand and then produce a translation of a given text (i.e. informed decisions regarding ambiguity, such as whether "I made her duck" means that the author cooked waterfowl for a woman, caused a woman to quickly lower her head, or something else). French scientist Bernard Vauquois, considered a pioneer in the field of machine translation, designed a helpful diagram for understanding the depth and complexity of these different RBMT approaches: the Vauquois triangle, pictured in *Figure 2.4* (Trujillo, 1999). The Y-axis represents the level of analysis (i.e. lexical, syntactic, semantic, discourse) that is performed by a system to try and resolve the same types of ambiguities that controlled language tackles. The X-axis represents the level of comparative bilingual knowledge needed to move a sentence from one language to another.

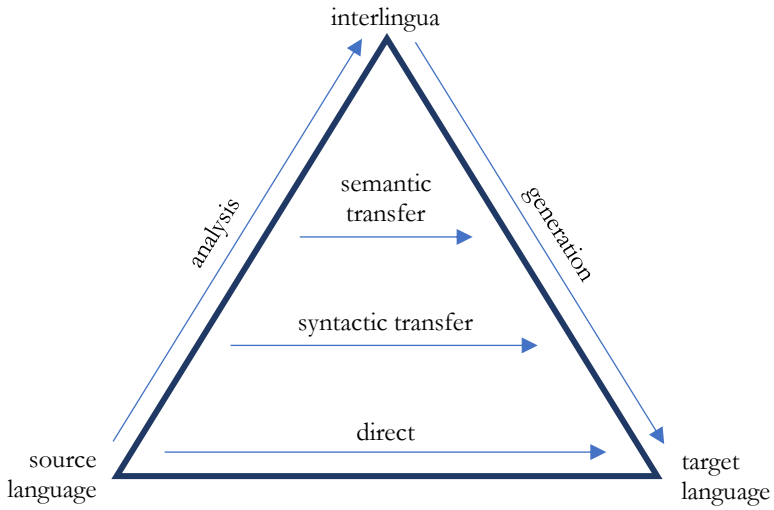



*Figure 2.4: Vauquois triangle (Trujillo, 1999, p. 6)*

#### 2.3.3.1 Direct systems

**Direct systems**, appearing at the bottom of the pyramid in *Figure 2.4*, are the simplest in terms of depth of analysis, which is only performed at the lexical/morphological level. This means that it only takes into account basic linguistic properties, such as parts of speech or singular/plural, and rules about how to treat certain sentences structures. Words are parsed in the source language and generated in the target language by way of a bilingual dictionary. It essentially results in word-for-word translation, sometimes rearranged using comparative grammar rules (Arnold et al., 1994). A disadvantage of direct MT engines is that once a system is built for a certain language combination it is limited to that pair; the bilingual nature of these systems does not allow for much carry-over into other languages (Hutchins, 2005; Jurafsky & Martin, 2009). They work best on languages that are structurally similar, for instance two languages that follow the same subject-verb-object word order (Quah, 2006). As Arnold et al. (1994) note, an advantage of these systems is that they are

robust, meaning that they do not break or return errors easily. However, a downside of this robustness is that they are susceptible to producing "word salad" output because when they encounter an unknown word, i.e. a word that is not in the bilingual dictionary, or a structure for which a specific transformation rule has not been coded, they simply return (fully or partially) untranslated and/or ungrammatical sentences (Arnold et al., 1994, p. 64).

#### 2.3.3.2 Indirect systems

In contrast to direct systems are **indirect systems**, collectively known as **rule-based machine translation**, which include two main approaches: transfer architectures and interlingua architectures. As can be seen in *Figure 2.4*, these models appear higher up on the triangle and therefore are more sophisticated when it comes to depth of analysis and are less specific-language dependent. While direct systems generally only perform analysis at the lexical and morphological level, indirect systems go at least one level deeper (up the pyramid, in this case), carrying out syntactic and semantic analysis. Dictionaries can use semantic information, such as categories of meaning (e.g. animal, plant, material) for each word, and syntactic information, such as whether an adjective appears before or after the noun it modifies, which theoretically helps resolve ambiguity.

Whereas direct systems produce translations in two steps (parsing and transformation) using a single bilingual dictionary and a small set of transformation rules, **transfer systems**, featured in the middle section of the Vauquois triangle, include three steps – analysis, transfer, and synthesis – and subsequently three dictionaries: one for each the target language and the source language, and one for the bilingual transfer step. They also rely on contrastive language knowledge. In the first analysis step, the system extracts syntactic and semantic information from the source text and turning it into a source-language representation, in the second step it transfers that representation into a target-language representation, and in the final step it generates a target text based on the information in that representation. Therefore, unlike direct systems, the word order of the source text has less influence on the target language output (Arnold et al., 1994; Quah, 2006).

**Interlingua systems**, at the top of the triangle in *Figure 2.4*, perform the deepest level of analysis of the three types of architectures we have seen so far. As Hutchins (2005, p. 503) explains, these systems work under the assumption "that it is possible to convert SL texts into semantico-syntactic representations which are common to more than one language (but not necessarily 'universal' in any sense)." Rather than focusing on words, these systems focus on *meaning* in two steps: analysis, in which an interlingua abstraction (the "semantico-syntactic representation" that Hutchins refers

to) is developed from the source text, and synthesis, in which that abstraction is used to produce a target text. This feature of abstraction in indirect systems, and interlingua architectures in particular, can be considered an advantage because it results in systems that are less language dependent. Whereas direct systems are built for two specific languages and essentially have to be rewritten from scratch to include a new language, indirect systems can more easily be adapted to a multilingual environment. However, it can also be a disadvantage, because focusing on meaning is not as simple as it sounds. This is because, as Arnold et al. (1994, p. 78) so aptly put it, "different languages 'carve the world up' differently." No two languages have a perfectly one-to-one equivalence between all of their vocabulary. Translation would be quite a boring task if that were so! For instance, in English one could have a "turkey" sandwich or a pet "turkey"; in French, while one could have a pet "dinde" or a pet "dindon," depending on whether the animal is female or male, one would only ever eat a sandwich made of "*dinde*" (regardless of the sex of the bird the lunch meat came from). Distinctions like these make it difficult to reduce language down to its most basic concepts.

#### 2.3.3.3 Corpus-based systems

Although work on these linguistic systems did not stop and some rule-based technology is in fact still in use today, a new, empirical paradigm emerged in the 1990s, becoming the third generation of MT: **corpus-based machine translation**. In fact, the idea was first explored beginning in the 1960s at IBM but was abandoned for about the next two decades (Quah, 2006). The basis of these systems is translation itself, in contrast to rule-based systems, which are underpinned by linguistic knowledge. Corpus-based systems make use of existing translations in the form of aligned parallel corpora, collections of texts that are broken into corresponding bilingual segments. There are three major architectures of this type; two are fairly similar in nature, example-based systems and statistical systems, and the third and most recently developed, neural systems, represents a significant departure into new territory and therefore we dedicate *Section 2.4* to this technology.

Somewhat of an intermediary between rule-based and statistical systems, the architectures we explore next are **example-based systems**. In this model, the pairs of bilingual segments we mentioned are called "examples." Example-based machine translation (EBMT) systems work in three steps: matching, alignment, and recombination. In the matching step, the corpus is searched for examples that are similar to the source segment. Unless of course the full source sentence is already in the corpus, fragments or strings are then pulled from multiple different examples in the second step (alignment) and put back together in the right order using a target language grammar in the third and final step (recombination) (Quah, 2006; Somers, 2005).

Evidently, some target language knowledge is still needed in these models, which is not necessarily the case in the next approach to machine translation that we will address: **statistical machine translation** (SMT). As Quah (2006, p. 77) describes it, "The premise of this approach is that a translation can be modelled with a statistical process. Bayes' theorem deals with probability inference and defines how to *combine knowledge of prior events*, for example past translations with new data (new source-language texts) *to predict future events* (in this case, new translations)." Essentially, the goal of these systems is to find the target sentence with the highest likelihood of being the translation of the source sentence. Two models are used to ensure accuracy – that the correct meaning is transferred – and fluency – that it is transferred in a way that reads well in the target language. First, a parallel corpus must be trained to develop what is called the translation model; the probabilities that a word or group of source words is the right translation of a word or group of target words are calculated based on the data from existing translations. The training step also includes the development of a language model, determined using sequences of words of varying lengths, or n-grams, in the monolingual target-language corpus. During decoding, all of the possible – no matter how improbable – translations are identified and the one with the highest probability based on the two models is output (Hearne & Way, 2011; Quah, 2006; Somers, 2005). The simplest way that this operation can be expressed is:

$$Translation = argmax_T P(S|T) \cdot P(T)$$

where $T$ is all of the target sentence candidates for a given source sentence, $S$. *Translation*, the target sentence that is output, is the candidate that receives the best score, calculated as the product of the translation model ($P(S|T)$) and the language model ($P(T)$) (Hearne & Way, 2011, p. 206).

These approaches often produce more fluid, natural-sounding translations since the target language model probabilities can act as fluency "tie-breakers" for two translations that are equally accurate (Bouillon, 2017). Nevertheless, corpus-based machine translation is not without drawbacks. The computational resources needed to calculate and store the translation and language models is significant, however calculations only need to be performed once unless a corpus is modified. Another major disadvantage of these two approaches is of course, the need for large amounts of bilingual data (and reliable bilingual data, at that). Parallel corpora are not always easy to come by, as we establish in *Section 4.2.1*. Furthermore, though these models are generally quite robust, they are often not as well equipped to deal with unknown words (those not seen during training) as the newest type of corpus-based machine translation, which we will outline in the following section.

### 2.3.4 Neural machine translation

The latest paradigm to emerge is neural machine translation (NMT). Although NMT systems have risen to prominence only within the last five to six years, researchers were already exploring neural networks for machine translation in the 1990s. However, the computing power available at the time was a major limiting factor, and they were essentially set aside in favor of the development of SMT (Koehn, 2017). By 2016, the resources had finally caught up with the theory and key players in the technology industry, including Google, Microsoft, and SYSTRAN, began commercializing NMT systems (Crego et al., 2016; Microsoft Translator, 2016; Wu et al., 2016).

Castilho et al. (2017, p. 110) propose the following description of neural machine translation: "Neural models involve building an end-to-end neural network that maps aligned bilingual texts which, given an input sentence X to be translated, is normally trained to maximise the probability of a target sequence Y without additional external linguistic information." In this definition, we can observe a few similarities between NMT and SMT. First, both statistical and neural models are corpus-based and are trained using large amounts of "aligned bilingual" data. Second, both approaches generally do not rely on "external linguistic information," or the semantic and syntactic knowledge that forms the basis of RBMT. And third, like SMT, NMT produces translation based on "probability." The method by which probability is calculated is where the two approaches diverge. The neural network that Castilho et al. (2017) refer to is a machine learning technique made up of layers of interconnected nodes, or neurons. They receive stimulation from other nodes and, depending on the positive or negative weight of this input and the strength of the connection, produce an output by applying an activation function that either excites or inhibits the nodes they are connected to (Forcada, 2017). So-called hidden layers, meaning that "we can observe inputs and outputs in training instances, but not the mechanism that connects them," are another important feature of NMT (Koehn, 2017, p. 8). Connections are formed and strengthened during a neural system's lengthy training phase, using the data from massive parallel corpora, as described above; input weights are continuously updated and fine-tuned until the output generated is as close to the reference translation as possible (Forcada, 2017; Koehn, 2017). The result is a distributed representation of vocabulary, or a word embedding.[16] In simplified terms, this means that each word in the vocabulary is assigned a unique numerical representation and words that appear in similar contexts are mapped to similar vectors. The example word embedding shown in *Figure 2.5*

[16] As mentioned in the previous section, NMT can also deal with unknown or rare vocabulary by using representations of sub-word character sequences, or parts of words (Koehn, 2017).

has just two dimensions to help us visualize the concept, but real distributed representations in NMT feature many more (Forcada, 2017; Koehn, 2017).


*Figure 2.5: Two-dimensional model of word embeddings, a visualization of the semantic similarity of words represented by physical proximity, from Koehn (2017)*

One of the main NMT architectures in use at the time of writing is the encoder-decoder approach, which relies on two recurrent neural networks. The encoder portion of the architecture builds a vector representation of the source sentence by recursively combining all of its parts. From this representation, the decoder generates a target sentence by choosing the most likely word at each position, taking into consideration everything that has already been generated, until an end-of-sentence marker is the next most likely word (again, based on probabilities). Forcada (2017, p. 296) likens this process to a text prediction feature, such as the one built into most smartphones: "… the decoder provides, at each position of the target sentence being built, and for every possible word in the target vocabulary, the likelihood that the word is a continuation of what has already been produced." Cho et al. (2014) found that this approach worked well, but identified an issue with longer sentences, to which Bahdanau et al. (2015, p. 4) responded by introducing an alignment model, also called *attention*:

> The decoder decides parts of the source sentence to pay attention to. By letting the decoder have an attention mechanism, we relieve the encoder from the burden of having to encode all information in the source sentence into a fixed-length vector. With this new approach the information can be spread throughout the sequence of annotations, which can be selectively retrieved by the decoder accordingly.

Koehn (2017) cites this research from 2015, when attention mechanisms were first added to pure NMT, as the turning point when the technology truly became competitive, and since then it has been even further refined, definitively becoming the MT state of the art. Advantages and disadvantages of NMT technology are outlined in the following section about the generic NMT tool chosen for this thesis.

#### 2.3.4.1 DeepL Translator

DeepL Translator, the neural machine translation system chosen for this thesis, was first launched in August 2017 by German start-up Linguee GmbH (now DeepL GmbH) and offered free machine translation to and from seven different languages. Four more were added between 2018 and 2020, including languages with non-Latin writing systems, plus regional varieties for English and Portuguese. Unlike its main competitors such as Google Translate and Yandex.Translate, which began as purely statistical models and later evolved to incorporate neural networks, DeepL Translator has been an NMT system since its inception.[17] Since DeepL Translator is owned by a private company, not much information has been made public about the inner workings of its proprietary software or where it obtains the language data fed into its system for training purposes. However, we do know that it was developed by the same company that created Linguee, an online dictionary and multilingual parallel corpus, or concordance tool, featuring over 1 billion translations.[18] A February 6, 2020 press release by DeepL announced a completely redesigned neural network, that is "far superior to previous technologies."[19]

**Advantages**

When Quah published his book *Translation and Technology* in 2006, he concluded that generic online MT systems were most valuable only for personal use – consuming information or producing writing in a language they do not know (known as assimilation in Translation Studies), and that does not necessarily have to be held to high quality standards. But technology has come a long way since then. In fact, neural networks were only mentioned two times in the entire book, as a potential area of knowledge that was lacking (Quah, 2006). Two major benefits of general-purpose online systems such as DeepL that have *not* changed since 2006 are that they are free and public. As *Section 2.2.3.2* revealed, a business or organization might not have the means or justifications to invest in a customized system to handle the small number of E2R publications that are currently

[17] https://www.deepl.com/press.html Last accessed: September 16, 2020
[18] https://www.linguee.com/ Last accessed: September 16, 2020
[19] https://www.deepl.com/blog/20200206.html Last accessed: September 16, 2020

being produced. An advantage specific to the DeepL system is that the results it produces are very good compared to much of the technology Quah praised over a decade ago, and even relative to other comparable NMT systems on the market today. At the same time that they introduced their semi-automatic MT quality evaluation framework, TQ AutoTest, Macketanz et al. (2018) tested it on five different MT engines for the German-to-English language pair. They found that DeepL achieved better results than Edinburgh (NMT) and Google (SMT) across the board, better results than Lucy (RBMT) in all but two (non-verbal agreement and negation) of the categories they measured, and better results than Google Translate in 11 out of 14 categories (with the exception of punctuation, negation, and subordination) (Macketanz et al., 2018). Findings by Kaplan et al. (2019) for NMT of FALC from French into English, introduced in *Chapter 1* and further developed in *Section 3.5.2* also support these assertions. In addition to these optimistic findings, we also know that artificial intelligence (AI) and NMT technology are constantly and rapidly evolving and expect the DeepL tool to continue to undergo improvements. For instance, two features that were not available when the experiments for this thesis were carried out have the potential to resolve several of the translation quality issues highlighted in *Section 5.2*. A glossary feature was released on May 6, 2020, which gives the user added control over terminology,[20] and as of August 27, 2020 DeepL offers the possibility to specify American English or British English as a language variation preference.[21]

**Drawbacks**

With headlines such as "DeepL schools other online translators with clever machine learning,"[22] and "Intelligence artificielle et traduction : DeepL pulvérise ses « concurrents » !" (*Artificial intelligence and translation: DeepL pulverizes "competition"!*),[23] blogs and news media seem to tout DeepL as a near miracle cure. However, like any technology, it is not without limitations. Chiefly, unlike RBMT and SMT systems (and customized NMT systems, though they require large amounts of monolingual bilingual training data),[24] DeepL is nearly impossible to tailor to one specific field or application. A collaboration between Swiss Post, the national postal service in Switzerland, and researchers at the University of Geneva found that the "off-the-shelf" DeepL tool outperformed even a customized (i.e. trained on translation memories and field-specific glossaries) SMT system

[20] https://www.deepl.com/blog/20200506.html Last accessed: September 14, 2020
[21] https://www.deepl.com/blog/20200824.html Last accessed: September 14, 2020
[22] https://techcrunch.com/2017/08/29/deepl-schools-other-online-translators-with-clever-machine-learning/ Last accessed: September 14, 2020
[23] https://www.xavierstuder.com/2020/02/intelligence-artificielle-et-traduction-deepl-pulverise-ses-concurrents/ Last accessed: September 14, 2020
[24] https://omniscien.com/custom-mt/ Last accessed: September 14, 2020

for the German-French language pair, both in terms of the automatic BLEU metric and human post-editing time and human-targeted translation error rate (HTER) (Volkart et al., 2018). However, when compared to a customized NMT system on the other hand, DeepL did not fare so well; a follow-up study with the same business client found that end-user satisfaction was higher for both raw and post-edited translations produced by a trained, in-house system. Limited data security is another drawback to generic translation technology such as DeepL, a concern voiced by participants in the same Swiss Post study (Girletti et al., 2019); however, it is unlikely to pose much of a problem for organizations wishing to publish E2R materials destined for public consumption, like those that made up our corpus (*Section 4.2*). The way that words are represented in NMT can also present limitations, chiefly: terminological inconsistency, explored further in our results (*Section 5.2.1*), and lack of transparency, making it more difficult to figure out where errors could have been introduced in the training corpus (L. Wang et al., 2017). Finally, NMT is generally regarded as less predictable than its predecessors, provoking errors that can sometimes be overlooked due to the grammatical correctness and generally more fluent nature of target translations (Neubig et al., 2015).

## 2.4 Controlled language and neural machine translation

As mentioned in *Section 2.2.2.1*, translatability is one of the main goals that has been identified for some controlled languages. Until 2019, we believe that the research tying these two fields together was centered exclusively around CLs and RBMT[25] or SMT.[26] To our knowledge, Marzouk and Hansen-Schirra (2019) and Rossetti (2019) have carried out the only published English-language research on controlled language as applied to neural machine translation at the time of writing.

In the former study, a technical-writing CL developed by tekom was applied to German texts, which was consider a pre-editing step, and results were analyzed for five different MT systems. Interestingly, through error annotation, human evaluation, and automatic evaluation, they found a positive impact on the content and style of English translations for all of the MT systems they tested *except* the NMT system. Google NMT was the only system out of all five to produce better results for the non-controlled language output than for the output produced after a CL was applied. Their hypothesis was therefore rejected, and the results suggest that text that is easier for a human to understand is not necessarily "easier" for NMT to deal with. They even go so far as to

[25] See Mitamura (1999).
[26] See Aikawa et al. (2007), Doherty (2012).

call this approach, the application of CLs before machine translation as a strategy for improving output quality, "obsolete" for new NMT architectures (Marzouk & Hansen-Schirra, 2019, p. 200).

One portion of the latter study, Rossetti's (2019) research, focused on the machine translatability of simplified health content. Unlike Marzouk and Hansen-Schirra (2019) who compared MT output from different systems, Rossetti (2019) was interested in comparing NMT quality of plain language summaries (PLS) that were verified using the semi-automated Acrolinx CL checker as part of the authoring step and those that were not. A similar controlled authoring tool, tailored to E2R, could help ensure consistency in authoring and resolve some of the Easy-to-Read violations that were already present in the French source texts that we studied for this thesis and that were transferred into the English translation (see *Section 5.3* for examples and discussion regarding these instances). One such system, designed to facilitate FALC text production, is described in *Section 3.5.1*. Although the author found no significant improvement in quality between the non-automated and the semi-automated CL source texts, she did observe promising overall results in terms of NMT quality: "While the style of the MT output was often described as unnatural by the evaluators, the content of the source English PLS was often translated fully and accurately into the Spanish output" (Rossetti, 2019, p. 212).

Marzouk and Hansen-Schirra (2019) did not produce particularly encouraging findings with regards to the neural machine translatability of controlled language, however both studies differ in significant ways from the research carried out for this thesis – language combinations (German-English and English-Spanish vs. French-English), NMT systems (Google vs. DeepL), CLs and aims (tekom technical communication and PL health information vs. E2R administrative communication), target audiences (lay readers vs. readers with intellectual disabilities) and most notably the comparison that was explored (CL/non-CL and non-automated CL/semi-automated CL vs. pre-edited CL/pure CL) – making it problematic to draw close parallels between these two studies and this thesis. Thus there remain major gaps in the current research that this thesis can attempt to fill.

# Chapter 3: Factors of Linguistic Accessibility

## 3.1 Introduction

Controlled language and machine translation, the areas of research introduced in *Chapter 2*, could be considered the left-hand side of an equation. The vehicles that, when operated together, either lead us to our goal of successful accessible multilingual communication or lead us astray. On the right-hand side, behind the equal sign (or perhaps the ≠ sign, as discussed in *Chapter 6*) is the aforementioned goal: linguistic accessibility. In this thesis we will explore three factors that fall under the umbrella of this term and that constitute the dependent variables we studied using the methodology described in *Chapter 4*. Two of these variables, **translation quality** (*Section 3.2*) and **readability** (*Section 3.3*), are general in nature and apply to a wide range of translated texts, while the third, **accessibility** (*Section 3.4*), relies on conditions unique to our target audience. It is vital for the researcher and the reader to have a mutual understanding of these key definitions that are adopted going forward. Other measurements of linguistic accessibility that are not explored in this thesis but that could be the focus of future related work include satisfaction or acceptability, comprehensibility, and usability. *Section 3.5* provides an overview of the research most closely related to this thesis so that we may identify how this thesis address the gaps in our current knowledge.

## 3.2 Translation quality

The first component of our definition of linguistic accessibility that we must define is translation quality, a key issue and highly debated topic in Translation Studies. As such, many attempts have been made to take quality out of the "eye of the beholder" and somehow render it quantifiable through translation quality assessment (TQA). We will now survey the predominant approaches to quality assessment of machine translation in particular, acknowledging that this vast subject cannot be addressed fully in the limited space we have for this thesis.

### 3.2.1 Automatic measures of translation quality

The automatic approach to TQA consists of calculations performed by machines that essentially compare a model, written by a human, to the machine translated output. The closer the machine translation is to the reference, the better the score the system obtains. Common measures of

translation quality assessment carried out by machines include Word Error Rate (WER), Translation Error Rate (TER) and Human-Targeted TER (HTER), Bilingual Evaluation Understudy (BLEU), and METEOR (Castilho et al., 2018). Although automatic evaluation of machine translation quality often has the advantage of being relatively inexpensive and impartial, the current technology does not allow for the same level of analysis that human evaluators can bring to the table. As Castilho et al. (2018) observe, whereas automatic measurements must compare machine translation output to a reference – which in itself introduces a degree of subjectivity since it assumes that what the human translation produced was the "best" translation, or the gold standard – humans have the analytical capabilities to acknowledge that for any given source sentence there can exist many different but equally valid target sentence possibilities.

### 3.2.2 Manual measures of translation quality

Human evaluators can also determine what makes a particular translation better quality than another. There are several different methods of manual TQA. Most commonly, evaluators are expected to assess quality based on two factors: fluency, how well a target text reads in and complies with the norms of the target language; and adequacy, the extent to which a target text faithfully relays the information or meaning present in the source text. In practice, evaluators often rate these two factors on a Likert scale for a given translation, but sometimes two or more translations can be compared and ranked against one another (Castilho et al., 2018).

Because we were interested in a more granular view of the types of errors in the NMT of E2R texts, *Step 1* of our experiment used another type of manual evaluation: error annotation. In order to produce a more realistic evaluation of whether or not generic NMT could be a suitable way to produce E2R that is fit for publication, we supplemented that annotation with a post-editing task, another measure of translation quality often employed by translation agencies. These two types of evaluation are explored in the sections that follow.

#### 3.2.2.1 DQF-MQM error typology

In the 1980s, the need was identified for a way for language service providers (LSPs) to hold translators to a certain objective and consistent standard of quality and at the same time to provide them with more concrete feedback. Various solutions and standards appeared and evolved without overwhelming success, until two until projects emerged in the early 2010s that would eventually become one integrated metric for TQA: DQF-MQM. Four main principles formed the foundation of the MQM project: a flexible catalog of error types ("MQM does not define a single metric, but rather a common vocabulary for declaring metrics"); compatibility with existing systems; multilayer

specificity; and an approach based on the requirements or brief for any given translation (Lommel, 2018, pp. 113–114). Eight primary "branches" were identified, each with its own set of more granular issues, as well as four levels of error severity. Simultaneously, the Translation Automation User Society (TAUS) was developing their own system for quality assessment called the Dynamic Quality Framework (DQF), which included, among other things, an error typology based on recommendations from LSPs and translation buyers (Lommel, 2018). In 2015, the two were harmonized and have since become the leading industry standard, pictured in *Figure 3.1* (not including the five severity categories: critical, major, minor, neutral, kudos) (TAUS, 2016, p. 16). The taxonomy of DQF-MQM is made up of eight main branches: accuracy, which deals with translation-related issues such as addition/omission, under-/over-translation, and flat out mistranslation; fluency, which encompasses "issues related to the form or content of a text"; terminology, which is domain-specific; style, which may be unidiomatic, awkward, inconsistent within the text, or in violation of the style guides specified by the translation commissioner; design, which moves away from linguistic factors and into layout and formatting issues; locale convention, which considers the specific conditions and context in which a target text is received; verity, which deals with culture specificities in the content; and other.[27]

Lommel (2018, p. 109) proposes a rather optimistic position on this shared method: "By bringing together disparate strands of quality assessment into a unified systematic framework, it offers a way to escape the inconsistency and subjectivity that have so far characterized TQA." For these reasons, we use this tool in our translation quality evaluation (refer to *Section 4.3.1.1.1*). However, it is not without drawbacks. It has been critiqued as a method, including by Bawa Mason (2019, p. 272) who acknowledges that while it is valuable as a thorough, common lexicon for talking about translation quality and errors, the application is impractical: "Reducing translation quality to this form of box ticking, however sophisticated, sidesteps many of the key elements essential to guarantee the usability of an actual translation in the real world." Additionally, as with any type of manual evaluation, there is always the possibility for bias to be introduced and simple human error to occur no matter how meticulously designed the framework is.

#### 3.2.2.2 Post-editing

DQF-MQM error typology and the two frameworks it was derived from are often paired with machine translation post-editing in research contexts. For instance, Zaretskaya et al. (2016)

[27] Descriptions come from the DQF-MQM error typology template used for error annotation in this thesis (*Section 3.3.1.1.1*), available for download at: https://info.taus.net/dqf-mqm-error-typology-templ Last accessed: September 21, 2020

measured correlations between different error types from the MQM taxonomy and post-editing time and effort to determine whether certain statistical, rule-based, and hybrid machine translation mishaps are more challenging to fix. Unlike in our post-editing study, described in *Section 4.3.1.1.2*, post-editors were alerted to the specific error to be fixed; this was done in order to control the experiment's variables, though a downside is that it provides less realistic post-editing conditions. Some research on the translatability of controlled language (CL) has also involved post-editing, discovering that the application of CL authoring rules to a text prior to MT can indeed reduce the amount of time that editors spend correcting and preparing the output for publication (O'Brien, 2004). However, time is not the only measure of post-editing effort, and the same study also found that CL did not necessarily have a positive effect on technical and cognitive indicators of effort. This leads us to a sort of crossroads: "One of the major concerns in the translation industry is how to quantify the amount of effort that is necessary for MT PE, based on the initial quality of the raw output, in relation to the final needs and expectations of the end-users" (Castilho et al., 2018, p. 29). Evaluating post-editing effort can be helpful for performing a cost-benefit analysis for using machine translation and hiring a post-editor versus hiring a human translator to perform the entire job, but the various ways in which it can be measured do not always align.

Since neither of the TQA methods we have just explored are perfect, in *Steps 1 and 2* of the study carried out for this thesis, we use both. We complement an error analysis using the DQF-MQM framework with a two-perspective post-editing study (temporal and technical) in hopes of reaching a more conclusive understanding of translation quality.

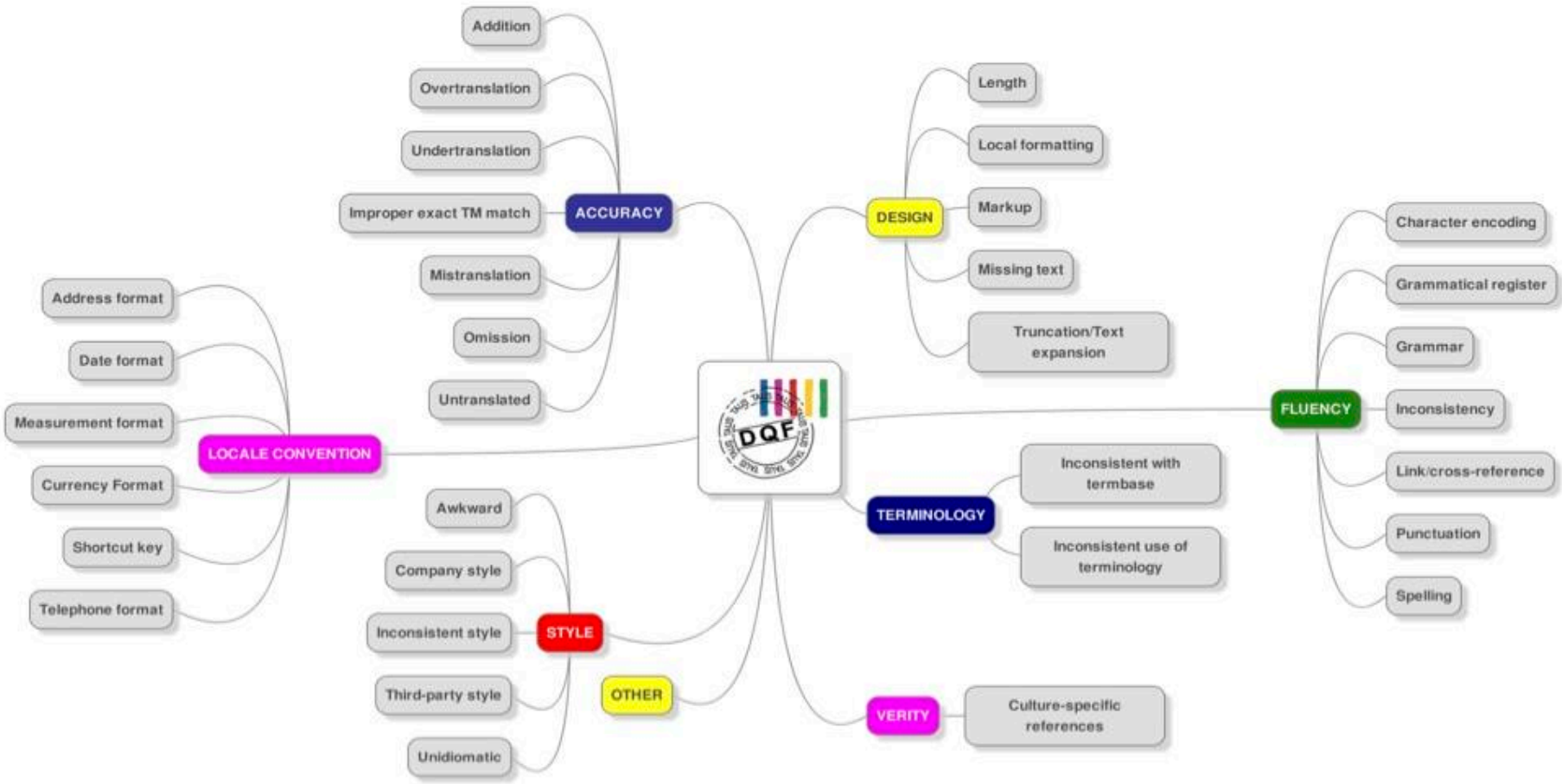



*Figure 3.1: A visual representation of the DQF-MQM harmonized error typology (TAUS, 2016)*

## 3.3 Readability

Readability has been studied extensively for more than a century; according to Dubay (2004), as of the 1980s, over 200 different readability formulas had been proposed and at least five times that many studies testing their validity had been performed. Yet there still is not a single definition or metric that everyone can agree on.

### 3.3.1 A brief review of the literacy and readability literature

In the US, adult literacy testing began in a military context (where it has continued to be prevalent throughout US history and into present day) in the early 1900s and quickly led to a more widescale exposure of the overestimation of reading skills in American adults. This realization, and the need for a system for accurately matching readers with level-appropriate reading materials that ensued, prompted the research that made way for the first readability formulas. Dubay (2004) identifies two main branches of literacy research that contributed to the formulas that are still used to this day: *sentence length* and *word frequency*. L.A. Sherman, who performed historical data analysis on sentence length, discovered that the average sentence length has been cut by over half, from 50 words in the Pre-Elizabethan era to 23 in the 1890s when he performed the research.[28] Based on this analysis, Sherman proposed that readability is positively impacted by shorter sentences. The second branch, word frequency and familiarity of vocabulary, began developing in the US in the 1920s, beginning with psychologist Edward Thorndike's English-language frequency list, *Teacher's Word Book*. It was based on the assumption that the more frequently a word is used, the more familiar it is to readers, and also the easier it is; humans generally learn "easy" words first, and then build a more advanced vocabulary through reading and education. The first readability formulas, or **text-based measurements**, that were proposed were based on these principles (Dubay, 2004). Three of the most well-known and widely tested calculations from this period, which motivated and moved readability research forward, are:

- the Flesch Reading Ease formula, introduced in 1943 and revised in 1948, based on average sentence length and average number of syllables per word, sometimes called shallow metrics of readability (Flesch, 1948). Rudolf Flesch was particularly influential in journalism and worked with the Associated Press to make major news stories more accessible, lowering them to an 11$^{th}$ grade reading level[29] (Dubay, 2004). Readability studies have historically been much less developed in French-speaking contexts than in the

[28] It has dropped even further since that research was carried out, sitting at around 20 words in the early 2000s (Dubay, 2004).
[29] Students in 11$^{th}$ grade are typically 16-17 years of age.

Anglosphere, however de Landsheere (1963) investigated the possibility of applying the Flesch formula to the French language and the adjustments that would need to be made for an accurate transposition;

- the Dale-Chall formula (1948), a calculation formulated by Edgar Dale and Jeanne Chall, which takes into account average sentence length and the presence of difficult words, those not found on their 3,000-word frequency list (Dale & Chall, 1948);
- and the 1952 Fox Index by Robert Gunning, which also relies on sentence length and "hard words" but adopts a different definition of "hard" – words featuring more than two syllables (Dubay, 2004, p. 24; Gunning, 1952).

Despite the flaws that many – including the author himself – expose (Davison & Kantor, 1982), Dubay (2004, p. 3) insists that readability formulas do provide, at the very least, "an objective prediction of text difficulty." In general, once a prediction made or a hypothesis formed, the logical next step is to test it. This is especially true for written material, since readability is intrinsically linked to the user; the reason most people write is so that others will read or consume the product. Another type of readability metric emerged after this primary period of research to fill that gap: **user-based measurements**. Researchers realized that many other factors besides the purely linguistic aspects of a text, such as how long a sentence is or how many syllables a word has, can affect how well a person comprehends a text. These factors can include how familiar with the topic a reader is, or even how motivated they are to read (Klare, 1976). User-based measurements can take the form of assessments of cognitive effort, such as in eye-tracking studies, or of usability – whether a person can successfully complete a task based on what they have read. Usability is a metric that is sometimes touted as superior or complementary to readability. Another popular form of user testing is via cloze tests, an assessment technique developed by Wilson Taylor in the 1950s, which test a user's ability to fill in blanks in a text based on context (Dubay, 2004).

### 3.3.2 Readability and disability

Traditional definitions and calculations of readability have been criticized for their one-dimensional nature and are often ill-suited when it comes to readers with particular needs. As Nietzio et al. (2014) point out in their work on E2R as a component of web accessibility, certain features of E2R can interfere with traditional principles of readability. They provide two relevant examples to illustrate that the length of a sentence is not the only factor that impacts readability. In their first example, the first sentence is certainly more concise, however it is also more ambiguous than the second and does not comply with the E2R guideline that recommends clear pronoun use:

She helped her. (sentence with personal pronouns)

The teacher helped the student. (sentence with nouns)

The second focuses on the E2R guideline that advises the use of the active voice whenever possible. The length of the two sentences is the same, so a traditional readability formula would indicate the same level of readability, even though the second is more accessible by E2R standards (Nietzio et al., 2014, p. 346):

The plan will be changed. (passive sentence)

We will change the plan. (active sentence)

Redish (2000) brings up several other critiques of readability formulas that happen to make them incompatible with E2R, notably that they do not work on non-traditional prose paragraphs, and that they do not take into account all (or even most) of the factors that make a text "usable," and that they assume a homogenous readership. First, as mentioned in *Section 2.2.3.1*, Inclusion Europe's guidelines advocate for lists in bullet point form rather than standard comma-separated lists in order to make relationships between ideas clearer. On paper, lists receive poor readability scores because they are calculated as long sentences, whereas in reality they have been shown to have a positive impact on usability (Grudniewicz et al., 2015). Second, readability often goes beyond what can easily be measured quantitatively (Redish, 2000); the many formatting, structural, and content-based guidelines that exist in addition to the rules on word difficulty and sentence length attest to the multiple dimensions that must be considered if a text can be labeled as Easy to Read.

Her third point, that readability formulas wrongly assume all readers are the same, is also especially relevant to our study (Redish, 2000). Reading and comprehension level of adults with mild intellectual disabilities is extremely difficult to generalize, as different conditions provoke drastically different challenges. O'Brien (2010, p. 144, emphasis mine) aptly situates this issue within the conflicting definitions of readability: "One can, therefore, view readability as being primarily dependent on the properties of text, *or as being a function of understanding or as being determined by the reader and his or her level of education and processing capabilities*." Reading abilities can vary between people with the same diagnosis and even from text to text for the same person, depending on factors such as familiarity of and interest in the topic (Feng et al., 2009; Yaneva, 2015). For instance, in their comparison of manually simplified and non-simplified texts in a user study to determine the cognitively motivated factors specific to adults with disabilities, studied as the first step in trying to create an automatic readability metric, Feng et al. (2009) found that traditional features of readability such as syllable count and word frequency might not be as significant in

determining the readability of a text for adults with ID, but entity density and other factors that impact working memory might. One of the main takeaways of their research was that texts written for children, which made up the corpora the study was based on, – although they contain more simplified ideas and language – might not be suitable for adults with ID or accurately predict readability because the groups have different challenges (Feng et al., 2009). This is not to mention that secondary target readerships of E2R documents – low literacy groups, the elderly, and learners of English as a foreign language, to name a few – have other readability needs still. That idea is even acknowledged in our Easy-to-Read guidelines. Guideline 1 in Section 1 states, "Always find out as much as you can about the people who will use your information and about their needs." Guideline 3 also addresses this question, "Always use the right language for the people your information is for. For example, do not use language for children when your information is for adults" (Inclusion Europe, 2009, p. 9).

For these reasons, the definition of readability that we have adopted for this thesis is directly based on an unweighted combination of features of Easy to Read (see *Section 3.4*), which include but are not limited to some of the same metrics considered in traditional formulas, like sentence and word length.

## 3.4 Accessibility

Accessibility, even when narrowed down to the context of text on the web, is a broad term that encompasses many principles, tools, and techniques, and that can mean different things to different people. Someone who is blind or visually impaired might use a screen reader or a braille display, and therefore define an accessible webpage as one that is compatible with the assistive technology that they use to consume information – one that features meaningful text alternatives for images that enhance the content and descriptive headings that make it easier to navigate, for instance. On the other hand, someone in the Deaf community may require captions or a sign language alternative to video or audio media on a webpage, and someone with a learning or intellectual disability may need adapted text in order to take full advantage of all a webpage has to offer.

Another factor that makes the term even more difficult to define is the fact that perceived accessibility, or how the user experiences a website, and accessibility as defined in design and development guidelines or recommendations such as the World Wide Web Consortium (W3C) Web Content Accessibility Guidelines (WCAG) are not always perfectly aligned (Aizpurua et al., 2016). Nevertheless, the WCAG are currently the most comprehensive and widespread set of

international accessibility recommendations for web content and their use is mandated by accessibility laws and policies in many countries (Kirkpatrick et al., 2018). In Switzerland, *P028 – Directives de la Confédération pour l'aménagement de sites Internet facilement accessibles* (2016) mandates compliance up to the AA level of WCAG 2.0, the second of three levels of conformance (A-AAA), meaning that all A-level and AA-level success criteria must be met. They were also recently incorporated into the internationally recognized standard, ISO/IEC 40500:2012 (ISO/IEC 40500:2012 Information Technology – W3C Web Content Accessibility Guidelines (WCAG) 2.0, 2012). WCAG 2.1 are broken up into four main principles of accessibility; web content must be perceivable, operable, understandable, and robust. In this thesis, we focus on linguistic factors of accessibility, so we are primarily interested in guideline 3.1.5 – Reading Level, an AAA-level success criteria that falls under the Understandable principle (Kirkpatrick et al., 2018):

> When text requires reading ability more advanced than the lower secondary education level after removal of proper names and titles, supplemental content, or a version that does not require reading ability more advanced than the lower secondary education level, is available.

The explanation of guideline 3.1.5 appears to equate accessibility with text difficulty, education level, and readability.[30] Though as we have shown in *Section 3.3.2*, traditional measures of readability are not necessarily a reliable way to predict how well a person with a disability will understand a given text. The editors of WCAG 2.0 themselves acknowledge in the document's abstract that not all user needs could be met through the application of the guidelines, particularly with regard to intellectual disability; they explain that conformance provides "accommodations for blindness and low vision, deafness and hearing loss, limited movement, speech disabilities, photosensitivity, and combinations of these, and *some* accommodation for learning disabilities and cognitive limitations" (Kirkpatrick et al., 2018). It would even seem as though we are making backward progress when it comes to barrier-free communication standards aimed at people with ID. Reading Level, an A-level requirement in the first version of WCAG, was demoted to an AAA-level requirement when version 2.0 was published due to concerns about how to test it objectively (Hassell, 2018), meaning that it no longer falls under Swiss mandates of accessibility. With these challenges in mind, the definition of accessibility that we adopt for our study – and that we believe should be integrated into future versions of WCAG since it is a clear testable and verifiable objective – is full compliance with the Inclusion Europe Easy-to-Read standard.

[30] https://www.w3.org/WAI/WCAG21/Understanding/reading-level.html Last accessed: October 10, 2020

## 3.5 State of the art: Intersections of CL, NMT, and accessibility

In the previous chapter (*Section 2.4*), we addressed the current state of the research concerning both controlled language and neural machine translation. In this section, we will present some of the most recent investigations into the relationships between CL, NMT, and the factors of linguistic accessibility that we defined at the beginning of this chapter (*Section 3.5.1*), as well as an in-depth look at the exploratory study that directly inspired this thesis (*Section 3.5.2*).

### 3.5.1 Controlled language, neural networks, and linguistic accessibility

Less common are intersections of all three of our primary topics: controlled language, machine translation, and linguistic accessibility. Automatic text simplification (ATS) is one tangentially related field of research that has recently produced findings on controlled language for accessibility purposes using neural networks (Chen et al., 2017; T. Wang et al., 2016). The *Simplification des Langues Écrites* (SIMPLES) project, a French-led initiative, is another. That endeavor has very similar ambitions to this thesis, though it approaches the issue in a slightly different way. Like our research, it operates on the assumption that if E2R text production can be made easier with technology, more E2R documents will be made available to people who need them (Chehab et al., 2019). Two deliverables are expected to come out of the SIMPLES project. The first, which has already been released, is a web-based open-source tool for FALC production called LIREC.[31] The second, which is in development with a team from the machine translation company SYSTRAN, is an automatic text simplification or summarization tool that relies on deep learning and differs from other similar initiatives in that it focuses on producing simplified text that meets FALC guidelines.[32] The text production tool offers an interface that facilitates FALC formatting as well as addressing the vagueness of some of the Inclusion Europe guidelines. For instance, the guideline that stipulates "easy-to-understand" words can be verified with the help of a built-in corpus and simpler alternatives can be suggested. LIREC also attempts to solve some of the difficulties related to evaluating how well a document meets E2R criteria; the tool provides a score from 0 to 100 for each full text or section of text and recommends a score of no less than 80 in order for a document to be considered Easy to Read. The interface was designed for ease of use, with the idea that it could be utilized by employees of ESATs, French work centers that employ people with disabilities (Jacquet & Poitrenaud, 2019). A similar project, FALC Assistant, is underway in Switzerland, led by the *Fondation pour la Recherche en faveur des personnes Handicapées (FRH)* and supported by

[31] http://lirec.fr Last accessed: October 9, 2020
[32] http://51.91.138.70/simples/ Last accessed: October 9, 2020

researchers from the Faculty of Translation and Interpreting at the University of Geneva.[33] These programs are currently only available for French and are still in the testing phases, but if they can be further improved and localized for other languages they could prove to be alternative or complementary to the generic NMT solution explored in this thesis.

### 3.5.2 Exploratory study

Both of the areas of research described in *Section 3.5.1* target the three main topics that we are interested in. However, recall that our definition of linguistic accessibility does not only include "classic" accessibility – the multilingual component is missing from them. Thus, Kaplan et al.'s (2019) exploratory study, presented at the Klaara Conference on Easy-to-Read Language in Helsinki, Finland, is the first to our knowledge to explore the intersections of controlled language for accessibility purposes and the latest developments in interlingual machine translation. In the exploratory study, researchers selected a 7-page French-language FALC text on Swiss disability insurance reform to translate into English using three free and generic MT systems: Yandex.Translate (SMT), Google Translate (NMT), and DeepL Translator (NMT). In order to compare the performance of these three systems, they recruited three native-English speaking translators in training (graduate level) to perform manual error annotation and E2R guideline violation annotation on the three resulting translations. In addition to this human evaluation, they also performed an automatic readability analysis using Coh-Metrix 3.0 (see "Coh-Metrix tool," *Section 4.3.2.1*). This thesis implements a modified version of their methodology, which is described in *Chapter 4*. Several of the findings that came out of this study encouraged us to continue on a similar path and explore whether steps could be taken to improve the outlook of this type of technology to the point where it might be a viable solution for producing E2R text in multiple languages.

First, their general findings from the human evaluation led to our decision to narrow our focus to one NMT system. Kaplan et al. (2019) found that DeepL produced by far the best results (for the French-to-English language pair and direction, and the administrative text genre), turning out an average of roughly half as many translation errors and E2R violations as the worst-performing system, Yandex.Translate. With regard to E2R, DeepL performed only marginally better than the other NMT system they tested, Google Translate, but about 30% in terms of translation quality.

[33] http://www.falc-assistant.ch/index Last accessed: January 4, 2021

Results from the Coh-Metrix automatic evaluation of nine indices that could quantify readability and adherence to E2R guidelines (e.g. 2nd person pronoun incidence, negation density, average word length in syllables) put DeepL squarely in the middle of the other two systems. It never obtained the highest nor the lowest score for any of the nine indices. However, these findings lead to questions regarding which segments should and should not be included in the analysis. Take, for instance, this French sentence from our corpus and the English translations we obtained:

| Source | Non-Pre-Edited | Pre-Edited |
|---|---|---|
| Pour combien de temps la CMI invalidité est-elle donnée? | For how long**? For how long?** is the disability **IJC** given? | For how long is the Disability **MIF** given? |

*Table 3.1: Example of a source-target pair from our corpus*

Both present major terminology issues, i.e. mistranslation of the acronym, but the non-pre-edited version also features an addition and a major fluency error, i.e. punctuation in the middle of the sentence. The erroneous punctuation in the non-pre-edited version would positively influence the entire analysis' average sentence length. If these were the only two sentences in the analyses, the pre-edited version would receive a score of 8 and the non-pre-edited version would receive a score of 3.6. For this reason, and because the size of our corpus allows – which was not necessarily the case for the exploratory study – our readability analysis (see *Section 4.3.2.1*) excludes segments with accuracy and fluency errors but allows for other errors such as style.

Next, an analysis of the error types identified in their findings prompted us to investigate one specific E2R guideline in order to determine its impact and potential solutions. Considering the nature of French administrative texts, which tend to contain many references to country- or culture-specific policies, services, and concepts that do not necessarily have English equivalents, certain types of errors were unsurprising. Accuracy errors, which include mistranslation, addition/omission, and under-/over-translation, and terminology errors, are to be expected in these types of translations. However, their evaluation turned up nearly the same number of fluency errors, including grammar, spelling, and punctuation errors, which are generally less prevalent in neural machine translation than in other types of MT (Bojar et al., 2016). They also observed some unusual translations, including highly ungrammatical sentences such as: "To succeed in reform, all people must who deal with disability insurance are participating" (Kaplan et al., 2019). This prompted a desire to find out what in the source text could be responsible for such unusual errors, and a subsequent investigation into the impact of certain characteristics that are specific to Easy-to-Read language, which became the foundation for our research.

Finally, one of the avenues for future research recommended by Kaplan et al. (2019) was a post-editing effort study. We know that the presumed reading level and tolerance for "textual disturbances," their willingness, or in this case ability, to accept grammatical or stylistic issues or unusual syntax produced by MT (Roturier, 2006, p. 157), of the target audience, adults with intellectual disabilities, are lower than those of the average adult. Ideally, the final text would contain no translation errors and no E2R violations. Even DeepL, the highest performing MT system in the exploratory study, produced an average of roughly one translation error for every four sentences, and one E2R violation for every five sentences. From these findings, we concluded that raw machine translation, i.e. text that is not revised before distribution, was not a valid option for this particular application. As a result, we included a post-editing effort study in this thesis to both provide another assessment of translation quality and help determine whether the benefits of using machine translation are offset by the cost of the required post-editing step.

# Chapter 4: Methodology

This chapter describes the methods and tools employed to study the effects of pre-editing on the neural machine translation (NMT) of Easy-to-Read administrative texts written and designed for people with intellectual disabilities. It begins (*Section 4.1*) with an introduction that summarizes the research goals and questions that inspired the experiments carried out within the framework of this thesis. In *Section 4.2*, we describe our materials, the corpus of texts developed to support our investigation, and introduce statistics about the texts to demonstrate how this thesis fits into the broader research contexts of simplified language, machine translation, and text accessibility. *Section 4.3* presents our research design: the human-based evaluation metrics and data generation methods used to measure translation quality (**RQ1**) and accessibility (**RQ2**), and the automatic evaluation and data generation method that was performed as a way to measure the readability of the raw and post-edited English texts generated by our generic NMT system (**RQ3**). Finally, *Section 4.4* provides a brief overview of the four different methods used in this investigation.

## 4.1 Introduction

Based on the encouraging findings from Kaplan et al.'s (2019) exploratory study (*Section 3.5.2*), this thesis attempts to provide a more in-depth exploration of NMT as a potential tool for E2R text production with the introduction of a pre-editing step into the translation workflow.

### 4.1.1 Research goals

As suggested in *Section 2.2.3.2*, cost of production, lack of training and awareness about how to produce and distribute E2R texts, and a dearth of research about the controlled language (CL) itself are all likely factors that have prevented E2R from reaching its full potential as a format of accessible communication. With this research, **we aim to further investigate how NMT could be improved to the point where it becomes a suitable tool for generating English E2R ("linguistically accessible" text) from French FALC that could be used by the primary target user group of this type of language: adults with intellectual disabilities (ID)**.

More specifically, we examine the impact of one specific guideline, which requires long sentences to be separated at spots that a person would naturally pause in speech, more similar to subtitles than to traditional prose (Inclusion Europe, 2009). Most prose moves to a new line when the character limit has been reached, sometimes even hyphenating a word onto two lines, whereas subtitling best practices dictate that "linguistic units" or "linguistic wholes" (i.e. clauses, phrases)

should be kept together, sometimes leading to line breaks that appear well before the possible end of a line (*BBC Subtitle Guidelines v. 1.1.8*, 2019). Irregular line breaks appear to have a similar effect on machine translation output as the long-term dependencies that provide a challenge for neural networks in long sentences; the more distance there is between a subject and verb, for example, the more difficult it is for a system to produce the correct agreement (Tang et al., 2018). **Whether or not, and to what extent, pre-editing text impacted by this guideline influences NMT output was assessed from three different perspectives: translation quality, accessibility, and readability.**

### 4.1.2 Research questions

Our studies were designed to answer three main research questions that address the three overarching dependent variables defined in *Chapter 3* – translation quality, accessibility, and readability. For all three questions, the independent variable was one specific pre-editing condition, the removal of line breaks:

> **Research Question (RQ) 1**: *How does the removal of line breaks from Easy-to-Read French-language administrative documents during the pre-editing process influence the* ***translation quality*** *of English output produced by a generic NMT system?*
>
> **Research Question (RQ) 2**: *How does the removal of line breaks from Easy-to-Read French-language administrative documents during the pre-editing process influence the* ***accessibility*** *of English output produced by a generic NMT system?*
>
> **Research Question (RQ) 3**: *How does the removal of line breaks from Easy-to-Read French-language administrative documents during the pre-editing process influence the* ***readability*** *of English output produced by a generic NMT system?*

A corpus of published Easy-to-Read documents was compiled (*Section 4.2*), and four methods of evaluation (*Section 4.3*) in two broad categories, human and automatic, were used to test our hypotheses in an attempt to answer these research questions. The hypotheses that were introduced in *Chapter 1* are also listed in each corresponding section of this chapter and summarized in *Section 4.4*.

## 4.2 Materials

In broad terms, a corpus is "a collection of texts [...] that are the object of literary or linguistic study" (Bernardini & Kenny, 2019, p. 119). Laviosa (2010) identifies a classification of corpora based on modern linguistics theory that includes six sets of properties: (1) sample (finite) or monitor (open); (2) synchronic or diachronic; (3) general or specialized; (4) monolingual, bilingual or multilingual; (5) written, spoken, mixed or multi-modal; and (6) annotated or non-annotated. We explore this classification as it pertains to the corpus used in this study in *Section 4.2.2*.

### 4.2.1 Controlled language corpora

Controlled language corpora have been used to study readability, accessibility, and translatability, although not extensively. Since this thesis focuses on French and English, we will primarily discuss controlled language corpora for these two languages, acknowledging that they do exist for others as well.[34] Some of the research goals for which controlled language corpora have been studied include: determining what factors impact readability and text comprehension by people with disabilities and other target populations of controlled languages (Mild Intellectual Disability (MID): Feng et al., 2009; Austism Spectrum Disorder (ASD): Yaneva & Evans, 2015); coming up with a gold-standard for accessible simplified texts (Štajner et al., 2015; Yaneva, 2015); developing automatic readability assessment tools (Feng et al., 2009); developing or evaluating automatic text simplification (TS) tools (Štajner et al., 2015; Yaneva, 2015), although it has been argued that non-manually simplified corpora are not necessary for building reliable tools (Glavaš & Štajner, 2015); and determining whether or not existing readability formulas are able to assess output of automatic TS systems (Štajner et al., 2015).

Much of the English-language research on CLs and readability or text simplification thus far relies on either encyclopedic or news texts. Simple English Wikipedia, for instance, which is said to be "written primarily in basic English and learning English,"[35] is a hotly contested source of CL texts. Some research has found this crowd-sourced corpus to be a suitable resource, such as Coster and Kauchak (2011), who aligned simplified (Simple English Wikipedia) and non-simplified

[34] Non-exhaustive list of simplified corpora in languages other than French and English: nine Brazilian Portuguese simplified corpora were built for PorSimples project on text adaptation tools for Brazilian Portuguese (Aluísio & Gasperin, 2010); a parallel corpus of *Alltagssprache* ("everyday language") and *Leichte Sprache* ("simple language") from German websites was developed with the goal of training a statistical MT system to translate from German to Simple German (Klaper et al., 2013); Bott and Saggion developed Simplext, a manually simplified Spanish corpus of newspaper articles along with the DILES research group from the Universidad Autónoma de Madrid for their work on a text simplification system (2014; Saggion et al., 2011); the Wablieft corpus of easy-to-read Flemish newspaper articles was used for a study on linguistic proxies of readability (Vandeghinste & Bulte, 2019).

[35] https://en.wikipedia.org/wiki/Simple_English_Wikipedia Last accessed: July 27, 2020

(Wikipedia) articles for a parallel corpus to serve as the basis for their phrase-based text simplification system. But the findings of other studies – like Štajner et al. (2012, p. 18), who observed that non-simplified texts from a fiction corpus scored higher in all of the categories used to measure readability than SimpleWiki articles – "cast doubt on the assumption that SimpleWiki serves as a paradigm of accessibility." Certainly, one advantage to Simple English Wikipedia is the breadth of data that it provides publicly; at the time of writing, it featured over 165,250 articles.[36] Unfortunately, at this time it only exists in English, and the conflicting research makes it a somewhat unreliable source of data for readability analyses. In one prominent study on readability assessment, Feng et al. (2009) acknowledge that the corpora they relied on, which included news articles geared toward elementary students, are not ideal, partially because texts written for children often do not match the interests of adults with ID. A lack of interest in the topics people read about can negatively affect motivation, which in turn impacts perceived difficulty. However they still consider their corpora valuable research tools because so few simplified/non-simplified paired corpora, and corpora with specific levels of readability (graded readers, for instance) exist (Feng et al., 2009).

Due to the critical importance of health literacy in societies, the readability of medical information has been the focus of much research and has led to the development of several simplified corpora. Grabar and Cardon (2018) built the comparable French-language CLEAR corpus, composed of encyclopedia entries, one set written for children and one set from Wikipedia and for the general population, medication packet inserts for lay people and their technical counterparts intended for medical professionals, and technical and simplified Cochrane reviews. While the CLEAR corpus was developed to be used by researchers and is available to the public online, it did not include Easy-to-Read texts. However, it does highlight the need for plain health information for all non-professionals, including an Easy-to-Read form for adults with disabilities, which could be an avenue for future research. Rossetti's (2019) work also focuses on a corpus of Cochrane summaries, exploring simplified health content from readability, comprehensibility, and machine translatability perspectives.

Felici and Griebel's (2019) corpus is perhaps the most similar to what was necessary for this study. While their Swiss-based investigation is certainly related to this thesis in that it highlights the work that remains to be done to further promote plain, multilingual administrative language that is accessible to as many citizens as possible, their corpus could not be applied to our work because

[36] https://simple.wikipedia.org/wiki/Main_Page Last accessed: July 27, 2020

of the important distinctions between plain language and Easy-to-Read that were explored in *Chapter 2* (Felici & Griebel, 2019).

### 4.2.2 Compiling a sample of Easy-to-Read documents

A collection of published Easy-to-Read texts was needed to try and answer our research questions. Since the corpora described above were not suitable for our purposes and no public corpora of French Easy-to-Read documents were available at the time of writing, we compiled one. Several selection criteria were established: document genre, non-availability of published human English translations, adherence to Easy-to-Read writing guidelines as evidenced by the presence of E2R logo, and PDF publication format.

First, all of the source documents must be French-language French and Swiss administrative texts related to the rights and inclusion of adults with disabilities in their community. We chose to focus on this genre because of the legal and civic duties established in Article 21 of the United Nations Convention on the Rights of Persons with Disabilities, which require member states to provide their citizens with information via "the form of communication of their choice" (*Convention on the Rights of Persons with Disabilities (CRPD)*, 2006). Documents were collected from the websites of three sources: the National Solidarity Fund for Autonomy in France (*Caisse Nationale de la Solidarité pour l'Autonomie (CNSA)*),[37] the French Ministry of Health and Solidarity (*Ministère des solidarités et de la santé*),[38] and the Swiss Federal Bureau for the Equality of People with Disabilities (*Bureau fédéral de l'égalité pour les personnes handicapées (BFEH) – Département fédéral de l'intérieur*).[39] Topics addressed in the documents include the financial and social support systems put in place for people with disabilities and how to claim or use them, and the laws and policies regarding disability rights.

Next, since DeepL is a free and generic NMT service and therefore could have been trained with any publicly available bilingual data, we selected texts with no known published English version or translation, which excluded documents such as the UN Convention itself. It should be noted that not all of the texts in the sample are adaptations of non-E2R documents; some, for example

[37] https://www.unapei.org/article/de-nouvelles-fiches-en-facile-a-lire-et-a-comprendre-falc-realisees-par-la-cnsa/ Last accessed: July 27, 2020
[38] https://solidarites-sante.gouv.fr/archives/archives-presse/archives-dossiers-de-presse/archive-courante-des-dossiers-de-presse/annee-2013/article/facile-a-lire-et-a-comprendre-un-an-au-service-de-la-solidarite Last accessed: July 27, 2020
[39] https://www.edi.admin.ch/edi/fr/home/fachstellen/bfeh/informationen-in-leichter-sprache.html Last accessed: July 27, 2020

the fact sheets produced by the CNSA, are stand-alone documents that exist only in an E2R format.

Because we were interested in studying the impact of line breaks, a rule that is specific to E2R and that does not appear in most other similar easy read or plain language guidelines, it was important that all of the documents in the corpus be written in accordance with these standards. We did so primarily by choosing documents bearing the Inclusion Europe E2R logo (*Figure 4.1*). According to the Inclusion Europe website, six conditions must be met if one wants to put the Easy-to-Read logo on their publication, the first of which is, "People using the logo must follow the European standards for making information easy to read and understand."[40] Inclusion Europe reserves the right to revoke permission to use the logo if they do not believe these guidelines are respected. According to Yaneva (2015), English human-produced easy-read texts – including but not limited to specifically Easy-to-Read samples – generally do comply with the guidelines that their authors claim to have followed. A study on compliance of German-language E2R texts came to a similar conclusion for some rules, such as sentence length and complexity, finding that less than 2% of sentences in the corpus studied contained complex structures, but not for others, such as the use of the passive voice and negative constructions (Nietzio et al., 2012). Nietzio et al. (2012) also argue that a style checker could be beneficial for E2R authoring, especially for inexperienced writers, and that some rules should be refined or reevaluated. Furthermore, people with intellectual disabilities participated in the creation of the majority of texts in our collection (Association Aires Paris, 2017). That the texts in our sample mostly complied with E2R guidelines is therefore a reasonable assumption, however documents are not systematically checked by Inclusion Europe before publication and we did not perform any further analysis on the source texts to ensure that they complied with E2R guidelines (refer to *Chapter 6* for a discussion of this and other limitations of the study).

*Figure 4.6: Easy-to-Read logo*

Finally, all of the documents gathered for this study were published PDFs, excluding administrative information written in E2R and published as text on a webpage. We made the decision to exclude webpages because this study addresses linguistic accessibility rather than the broader topic of web accessibility. Working with web-based documents would require other

[40] https://easy-to-read.eu/european-logo/ Last accessed: July 27, 2020

factors to be taken into consideration, such as W3C's WCAG 2.1 guidelines for distinguishability (Guideline 1.4) and navigability (Guideline 2.4) (Kirkpatrick et al., 2018).

The corpus fits into Laviosa's (2010) previously mentioned classification as follows:

(1) **sample or finite**: it contains a determinate number of words and was not and will not continue to be updated during the course of or after the study. The published French texts collected for this study included 41 documents for a total of 49,214 words. After extracting only those segments exceeding one line and filtering out exact duplicates, the French corpus contained 1,583 segments and a total of 24,011 words. More statistics about the size of the data set are presented in *Section 4.2.3*;

(2) **synchronic**: this parameter deals with time. The texts included in the collection were produced and published within a span of seven years, from 2012 to 2019, and therefore the sample cannot be relied upon to show any particular evolution of language over time, nor was it designed for that purpose;

(3) **specialized**: the corpus can be considered specialized for two reasons. Firstly, it contains only documents bearing an Easy-to-Read logo, and therefore should only contain language that adheres to E2R guidelines. Secondly, all of the documents included share a genre (administrative), subject matter (disability rights and issues), and text type (content-focused, according to Katharina Reiss's (2000, p. 32) typology; while the form, in the general sense of the word – *how* an author conveys the information that they convey – is an essential part of Easy-to-Read text production, Reiss makes the important distinction that the "form" in form-focused texts "contribute a special artistic expression," so we can safely conclude that the primary purpose of these documents is to inform);

(4) **bilingual**: not only can the E2R corpus used in this study be considered bilingual, since it contains texts in both French and English, but it can be further defined as *parallel*. A parallel bilingual corpus consists of source segments in one language and translated segments in another language, which are often manually aligned at the sentence level for annotation and analysis purposes (Bernardini & Kenny, 2019);

(5) **written**: easy language documents often feature symbols or images for a second layer of accessibility and readability by people with varying degrees of intellectual disability, learning disability, and/or language skills, although their effectiveness has been disputed (Cardone, 1999; Hurtado et al., 2014; Poncelas & Murphy, 2007). Guidelines 34-39 of

Inclusion Europe's Easy-to-Read "Standards for Written Information" encourage authors to add "photographs, drawings, or symbols" to support the text, and the majority of the documents in our test set did feature images to some extent (2009, pp. 20–21). However, since the scope of this investigation was limited to a linguistic analysis, images were omitted from all experiments, leading our corpus to fall under the written rather than the multi-modal category;

(6) **non-annotated**: other than paring the corpus down to isolate the phenomenon we were interested in studying (i.e. the presence of manual line breaks, described in *Section 4.2.3*), we did not perform any type of linguistic analysis on the French texts before translation.

### 4.2.3 Statistics on the Easy-to-Read texts used in this study

For the reasons described in *Section 3.5.2* on Kaplan et al.'s 2019 exploratory study, documents that did not contain at least one segment with a manually inserted line break were omitted from the corpus. Here, a "segment" could be a full sentence, but it could also be a title, a section header, or an item in a bulleted list, a common occurrence in E2R documents. The published French texts collected and used for this study included 41 documents and a total of 49,214 words (4,225 segments). The total number of words exceeding one line and therefore containing at least one manual or automatic line break was 33,377, in 2,248 segments. Repetition is inherent in Easy-to-Read writing, encouraged by Guideline 20 in Section 1 ("General standards for easy to understand information") of Inclusion Europe standards, which states: "It is OK to repeat important information. It is OK to explain difficult words more than once" (2009). Before proceeding with the translation, duplicate segments containing the exact same combination of words in the same order, and with line breaks in the same places, were also removed. This process resulted in a total of 24,011 source words and a total of 1,583 segments.

Of those 1,583 segments, 849, or 53.6%, contained at least one manually inserted line break (14,434 source words or 60% of the total word count). These statistics show that 20.09% of all segments in our corpus of French *Facile à lire et à comprendre* (FALC) documents, which account for 29.33% of all words, are impacted by Guideline 19. This finding further enforces our motivations to determine whether or not pre-editing can improve the machine translation of these types of texts, as it represents the potential for a significant improvement in the overall quality. Once the French side of the corpus was built and narrowed down to target our research questions, the generic NMT system DeepL Translator, developed by the German tech company DeepL and introduced in

*Section 2.3.4.1*,[41] was used to translate the 849 segments one by one into English, once with their original formatting (one or more manual line breaks) and once with formatting removed (this constitutes our pre-editing process). The non-pre-edited English corpus contained 13,677 words and the pre-edited English corpus resulted in 13,375 words. These statistics are summarized in *Table 4.1*.

| | **Documents** | **Sentences** | **French Source Text (ST) Words** | **Proportion of Total ST** | **English Target Text (TT) Words** |
|---|---|---|---|---|---|
| **Full corpus** | 41 | 4,225 | 49,214 | 100% (segments)<br>100% (words) | |
| **Text with ≥ 1 line break** | 41 | 2,248 | 33,377 | 53.21% (segments)<br>67.82% (words) | |
| **AND 0 duplicates** | 41 | 1,583 | 24,011 | 37.47% (segments)<br>48.79% (words) | |
| **AND ≥ 1 manual line break** | 41 | 849 | 14,434 (non-PrE)<br>14,434 (PrE) | 20.09% (segments)<br>29.33% (words) | 13,677 (non-PrE)<br>13,375 (PrE) |

*Table 4.2: Statistics about the corpus of FALC documents studied*

## 4.3 Research design

The research for this thesis, focused on the three dependent variables introduced in *Chapter 3* that make up what we call "linguistic accessibility," comprises two phases split into four different steps:

- *Phase 1* (*Section 4.3.1*) deals with methods of human evaluation and consists of three steps:
    - *Step 1* (*Section 4.3.1.1.1*) explores the translation quality variable using DQF-MQM error annotation;
    - *Step 2* (*Section 4.3.1.1.2*) also deals with translation quality, this time via a post-editing experiment, another metric of human Translation Quality Assessment (TQA);
    - and *Step 3* (*Section 4.3.1.2.1*) investigates our accessibility variable by way of an E2R violation annotation process on raw pre-edited and non-pre-edited segments;
- *Phase 2* (*Section 4.3.2*) centers around methods of automatic evaluation and consists of one step:
    - *Step 4* (*Section 4.3.2.1*), which examines the readability variable using measurements provided by the Coh-Metrix 3.0 text analysis tool.

[41] No version history is available for this tool, however translation was performed on November 12, 2019 using the free version of the web-based app, before the company achieved a so-called "quantum leap in translation quality" with the latest update to their system in February 2020.

### 4.3.1 Phase 1: Human evaluation of neural machine translation

Three of the four steps undertaken for this thesis relied on manual evaluation, or evaluation performed by humans; these three steps constitute *Phase 1*. This section presents the two variables that were studied using the Easy-to-Read texts that were translated from French to English using DeepL – **translation quality** and **accessibility** – how these three human-based evaluation studies were designed, and the instruments that supported data collection.

#### 4.3.1.1 Evaluating translation quality

Since the emergence of Translation Studies as its own field of research in the 1970s, scholars have visited and revisited the question of how to define and measure translation quality. As discussed in *Section 3.2*, an overwhelming number of models and approaches, both quantitative and qualitative, have been proposed to address this sticky subject. In professional translation contexts, quality can often be distilled down to the agreement upon and management of expectations between the translation provider and the translation buyer or consumer, whereas in translation studies contexts it often serves as a way to compare translations or translation processes (Moorkens et al., 2018).

The metrics of TQA that have emerged in the past two decades have various advantages and drawbacks. Automatic metrics rely on the idea that human translation is the ideal to strive for, so the more similar a machine translation is to a reference created by a human the better (Moorkens et al., 2018; Papineni et al., 2001). The first and most obvious limitation is the need for some form of human translated reference, which, as we have already established in *Section 2.2.3.2*, are not readily available for Easy-to-Read language. Another limitation is that even when a reference is present, a "good" translation could still obtain a poor score due to the fact that a single sentence can be translated in a myriad of ways. Unlike manual metrics, and one of the reasons we chose not to take that route for our evaluation, is that while automatic assessment can provide a global view of translation quality quickly, inexpensively, and objectively it does not reveal anything about why a translation is deemed good or bad (Moorkens et al., 2018). The two human quality assessment studies described below offer a more fine-grained analysis of the aspects of the texts that were most affected in the NMT process.

##### 4.3.1.1.1 Step 1: Error annotation

Error annotation is a common form of manual TQA in the translation industry; one editor from our post-editing study (*Section 4.3.1.1.2*) noted that she has used it "for numerous translation

agencies over the years." This step of the study aimed to test the following hypotheses using the Dynamic Quality Framework and Multidimensional Quality Metrics (DQF-MQM) introduced in *Section 3.2.2.1*:

**H1.0**: *Removing forced line breaks from French Easy-to-Read texts will improve the quality of English NMT output.*

> **H1.1**: *The segments that were pre-edited to remove manual line breaks prior to translation with DeepL NMT will contain* ***fewer translation errors*** *than the segments that contain manual line breaks.*
>
> **H1.2**: *The segments that were pre-edited to remove manual line breaks prior to translation with DeepL NMT will contain* ***less serious translation errors*** *than the segments that contain manual line breaks.*
>
> **H1.3**: ***Fluency*** *and* ***style*** *will be the two categories most positively affected by this pre-editing process.*

**Error annotation experimental design**

There were three minor deviations from the exploratory study (see *Section 3.5.2*) that inspired this research. First, this study did not use the "Company Style" subcategory of "Style" to indicate Easy-to-Read violations. The decision to separate the notions of quality and accessibility could seem problematic, especially from a Skopos theory point of view, in which the function of the translation product, or target text, are all-important; after all, the texts in our corpus were written for a very specific purpose and audience (Vermeer, 1978, as cited in Nord, 2010). However, since our three-pronged approach focuses on quality and accessibility as two different variables, it was important to isolate the two as much as possible. Second, due to the quantity of work required for this experiment, the author was the only annotator. Although it was not the main focus of the task, the post-editing experiment described in *Section 4.3.1.1.2* did include an error annotation element, so the data gathered could be analyzed against the errors flagged by the author for a small sample. Finally, while it could introduce subjectivity to the evaluation, the pre-edited and non-pre-edited segments were not randomized for the annotation process. This decision was made for greater ease of analysis (i.e. judgements sometimes must be based on context: the other segments

surrounding the segment in question) and because the author performed the translation and was therefore already aware which segments were the result of pre-editing and which were not.

#### 4.3.1.1.2 Step 2: Post-editing

Another way that researchers in Translation Studies and working professionals in the field evaluate machine translation quality is by measuring post-editing effort. Krings defines three measures of post-editing effort (PEE): temporal, cognitive, and technical (2001). We relied on two of these three components, temporal and technical, both of which were measured within the post-editing tool, to evaluate **H1.4**. Our indicator for temporal PEE was MateCat's time-to-edit (TTE) feature, which measures the amount of time an editor spends working on a segment, including pauses. The performance indicator that we used to evaluate the technical side of post-editing effort, which includes the physical operations that are used to change, add, delete, or move around parts of the segment, was the tool's "post-editing effort" measurement: the percentage of the pre-translated segment to which changes were made. Although cognitive effort is an important component of post-editing effort that contributes to the temporal parameter, it is also more difficult to measure than the other two dimensions because cognitive processes – what happens in an editor's mind – are not visible (Krings, 2001, p. 182). Krings (2001, p. 179) suggests think-aloud protocols as possible ways to gain insight into this parameter, but we did not perform one within the framework of this thesis due to time and resource constraints.

In theory, machine translations of controlled language such as Easy to Read should be easier to post-edit than those of natural language; as Quah (2006, p. 48) reasons, "As a result of the restrictions imposed on the controlled language sentences, the risk of errors in translation is reduced, thus reducing the burden of post-editing." Therefore, the study introduced in this section was designed to test **Hypothesis 1.4**:

> **H1.4**: *The segments that were pre-edited (i.e. manual line breaks were removed) prior to translation will require* ***less post-editing effort*** *to achieve publishable quality than the segments that were not pre-edited, and that therefore contained manual line breaks, when translated with DeepL NMT.*

**MateCat post-editing tool**

MateCat is a free, open-source, web-based computer assisted translation (CAT) tool developed in 2014. Although it was developed primarily for commercial purposes, MateCat also has applications in translator education and academic research, including post-editing studies (Federico et al., 2014). Zaretskaya et al. (2016, p. 87) used this tool in their examination of post-editing effort for different

MT errors, finding that, "The observed correlation between PE time and PEE was only weak. This means that different indicators of post-editing effort are not necessarily related: some errors require more time to find the right solution but do not necessarily involve many editing operations." Due in part to this finding, it was important to use multiple indicators in this research, as described in the discussion about post-editing parameters. In addition to the research that has

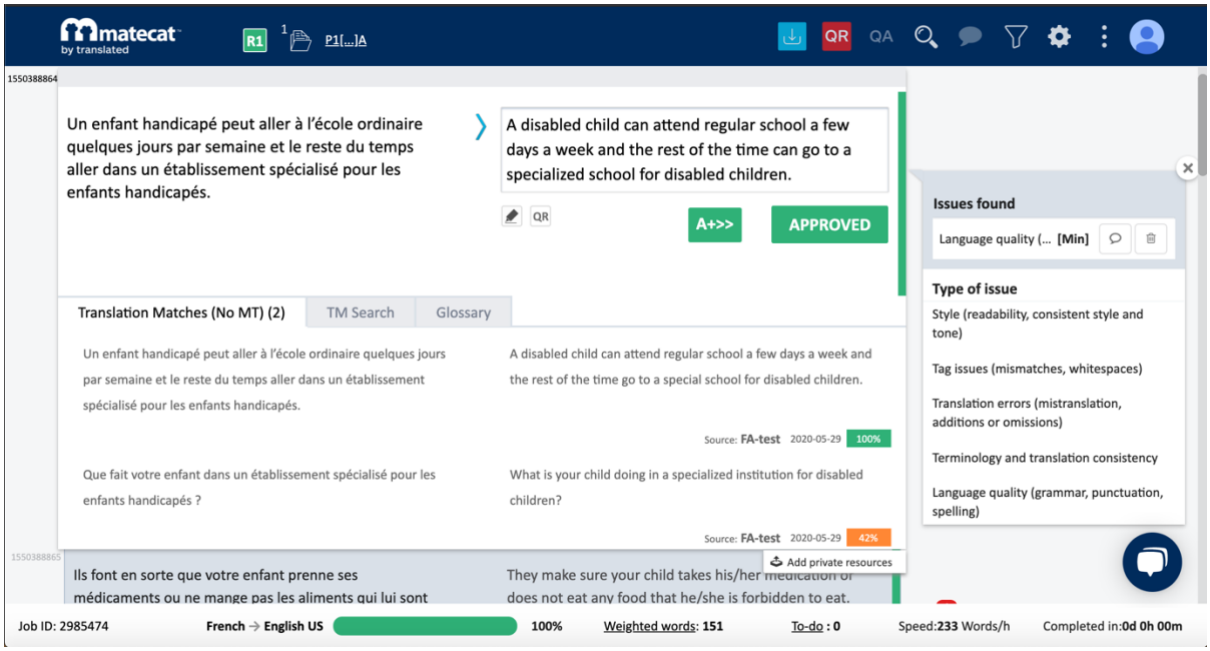



*Figure 4.7: MateCat post-editing environment*

already been performed using this tool, it was chosen due to its cost-effectiveness (free) and user-friendliness for both the editors and in terms of data analysis. We also considered it because it is introduced in the translation technology curriculum of the master's program from which the majority of participants were recruited, and we can assume that if participants were already familiar with the interface (shown in *Figure 4.2*), they would be less likely to skew results due to user error. MateCat version 2.0.0 was used for post-editing in June 2020.

*Figure 4.3* displays a sample entry from a MateCat post-editing log. The tool provides a comparative view of the source segment ("Source"), the target segment ("Suggestion" – input from our translation memory, half of our target segments were pre-edited and half were not), and the post-editor's revision ("Revision" – clicking the eye icon toggles a view of all of the additions and deletions performed) plus any errors that were flagged by the post-editor ("QA" – error types include style, tag issues, translation errors, terminology and translation consistency, and language quality, of major, minor, or neutral severity). Additionally, the tool measures post-editing time for each segment in total Time-to-Edit (TTE) as well as seconds per word, and Post-Editing Effort (PEE), the percentage of the suggested segment to which modifications were made. In the example

in *Figure 4.3*, we can see that the post-editor performed 4 substitutions, 3 deletions, and 1 addition, resulting in a PEE of 40% and a TTE of 1 minute and 24 seconds. In previous versions of MateCat, these reports could be exported as a .CSV file (Federico et al., 2014), however this feature has since been removed so the researcher compiled the data into Excel spreadsheets by hand.

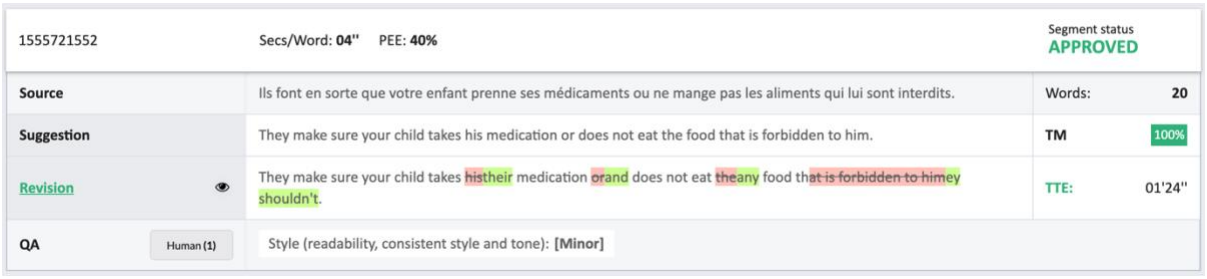



*Figure 4.8: Example of a MateCat 2.0.0 post-editing log*

### Post-editor profile

In order to evaluate **Hypothesis H1.4**, current (2019-2020) translation students and recent graduates of graduate-level translation schools in Switzerland, the US, and the UK were recruited to perform a post-editing task. Of the potential candidates identified and invited to participate via a call for participation email (*Appendix B*), six (N=6) volunteered to participate in the study, all of whom spoke English as a native language, had French as one of their working languages, and had completed at least one year of a master's program in translation (as self-reported on the background questionnaire and informed consent form, *Appendix C*). Although they were not required to have ever performed machine translation post-editing (MTPE) or error typology before, the call for participation stated that previous experience was preferred. All but one (N=5) participant had some degree of experience with MTPE, mainly within the framework of their education, and with translation quality assessment via error typology. Post-editors were given detailed instructions (*Appendix D*) on how to use the MateCat interface as well as a description of the quality standards that they were to achieve through the post-editing process (described in the next section, "Post-editing experimental design"). *Table 4.2* provides an overview of post-editor background and experience with MTPE and error annotation.

| Post-editor ID | English Dialect | MTPE Experience | Error Typology Experience | Education Level |
|---|---|---|---|---|
| PE1 | CA | Yes | Yes | 2nd-year MA |
| PE2 | UK | Yes | Yes | 2nd-year MA |
| PE3 | US | Yes | Yes | 3rd-year MA |
| PE4 | US | Don't know | Yes | 2nd-year MA |
| PE5 | US | Yes | No | MA graduate |
| PE6 | US | No | Yes | 3rd-year MA |

*Table 4.3: Post-editor background*

Since one post-editor reported that they had no prior experience with MTPE and one did not know if they had prior experience (however the experience they described points to yes), they were not asked to answer the perception questions. None of the participants who reported prior experience said that they performed MTPE often; all responded with a 1 or 2 on a 5-point Likert scale. Despite not performing this service frequently, all participants (N=4) were either neutral (score of 3) or comfortable (score of 4 or 5) editing to publishable quality and felt neutral about (score of 3) or slightly enjoyed (score of 4) the act of post-editing in general. *Table 4.3* displays a summary of post-editor perceptions of MTPE.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| I enjoy performing machine translation post-editing. | 0 | 0 | 2 | 2 | 0 |
| I prefer post-editing machine translation to editing human translation. | 0 | 1 | 2 | 1 | 0 |
| I feel comfortable post-editing to publishable (human-like) quality. | 0 | 0 | 1 | 1 | 2 |
| I feel comfortable post-editing to "good enough" quality. | 0 | 1 | 0 | 0 | 3 |
| Machine translation post-editing saves me time. | 1 | 0 | 2 | 0 | 1 |
| I perform machine translation post-editing often. | 1 | 3 | 0 | 0 | 0 |

*Table 4.4: Post-editor perceptions of MTPE, where 1 represents "Strongly disagree" and 5 represents "Strongly agree"*

**Post-editing experimental design**

Rather than ask evaluators to post-edit all of the segments from one of the 41 texts, 25 segments were chosen at random from the 849 source segments for which both pre-edited and non-pre-edited translations existed. This was done to obtain a more global evaluation of the translation quality in the data set. Dependent clauses that relied on other parts of the text (i.e. bullet points that appeared without a header segment) were omitted so that editors could work only within the interface, without having to refer to other segments for context. Sentences that resulted in identical

translations regardless of whether or not they were pre-edited were also omitted, since we can assume that the post-editing effort would be the same.

The post-editing tasks were performed remotely with MateCat, the free, web-based CAT tool described above. Editors received links to two files to post-edit and were instructed to complete each task in the order indicated, without consulting the other file. That said, we acknowledge that post-editing time for a segment could be reduced for that segment if an editor has already processed the source segment before (Gerlach, 2015). To reduce this bias, we provided two separate files so that evaluators could not directly use the changes they made to the first translation (pre-edited) to post-edit the second translation (non-pre-edited). The 25 segments in each file appeared in a randomized order to account for possible fatigue effects, with each source segment appearing only once per file. Half of the translators were instructed to complete File 1 first, and the other half were instructed to complete File 2 first. In other words, half of the time, an editor saw the pre-edited version of the target text first and the other half of the time they saw the non-pre-edited version first.

For reasons identified in *Section 3.5.2*, we were interested in achieving "publishable quality" translation, or "quality similar or equal to human translation." This is defined in the Translation Automation User Society (TAUS) MT Post-Editing Guidelines as translation that is "comprehensible…accurate…and stylistically fine," and that features correct grammar, punctuation, and syntax. Conversely, raw MT output or "good enough" (also referred to as "gist") translation, is comprehensible and accurate but not necessarily stylistically or grammatically flawless (TAUS & CNGL, 2010). The TAUS definition of "publishable quality" full post-editing, presented in part above, was provided in the task instructions so the editors would know what to edit for to achieve the desired result (see *Appendix D*). Editors had access to the French source text in order to evaluate translation accuracy, and a glossary (*Appendix E*) of acronyms and technical terms was provided. Since previous experience with MateCat was not a prerequisite for participating in this study, three warm-up segments were included at the beginning of the first file and dispersed randomly within the second file that each editor received; editors were not informed that they were practice, and the data for those six segments was excluded from our analysis. No time limit was imposed for either task, because our goal was not to replicate real post-editing conditions.

#### 4.3.1.2 Evaluating accessibility

Because this study addresses texts written with a specific set of guidelines in mind, achieving publishable translation quality does not equate to meeting the target audience's specific accessibility needs. Since this thesis focuses on a specific controlled language that was developed with the goal of making texts accessible to as many readers as possible, our accessibility indicator is directly related to the rules that it prescribes, rather than to other possible measures of accessibility, described in *Section 3.4*. So, we can reason that a text that complies with every E2R guideline will be more accessible than a text that includes many instances of long sentences and complex words, for instance (*Chapter 5* presents a discussion of the subjective nature of these metrics – "long," "complex," etc. – and therefore of E2R as a whole, one of the major limitations of this study). Thus, it was important to measure how well translated texts respond to the controlled language guidelines set forth by Inclusion Europe that must be met if a publisher wants to include the Easy-to-Read logo on their document.

##### 4.3.1.2.1 Step 3: Easy-to-Read violation annotation

With that in mind, an accessibility study was carried out to test **RQ2**, which can be measured quantitatively by **Hypotheses 2.1**:

**H2.0**: *Removing forced line breaks before performing NMT with DeepL will improve text accessibility.*

> **H2.1**: *The segments that were* ***pre-edited*** *to remove manual line breaks prior to translation with DeepL NMT will contain* ***fewer violations of E2R guidelines*** *than the segments that contain manual line breaks.*

To test these hypotheses, we performed the same assessment of adherence to Easy-to-Read guidelines that was carried out in the exploratory study that inspired this thesis work (Kaplan et al. 2019; *Section 3.5.2*) on the English NMT output (**H2.1**). In *Step 3*, the researcher annotated the 849 pre-edited segments and 849 non-pre-edited segments according to ad-hoc categories of general, word-level, sentence-level, and structural suggestions (*Appendix A*) laid out in the Inclusion Europe writing guide in order to measure the effect of pre-editing on the total number of E2R violations. A single round of annotations was performed using Excel.

### 4.3.2 Phase 2: Automatic evaluation of neural machine translation

The second phase of this thesis is based on automatic evaluation, or evaluation performed by machines. This section explains the third factor of linguistic accessibility that was studied using the

Easy-to-Read texts translated from French to English with NMT – **readability** – the tool used for data collection, and how the evaluation step was designed.

#### 4.3.2.1 Step 4: Automatic readability evaluation

The third variable of English neural machine translated texts that we aim to evaluate within the framework of this thesis is readability, through the testing of the following hypothesis:

**H3.0**: *Removing forced line breaks before NMT will* ***improve readability***.

*Section 3.3* presents a brief overview of the research in the field of readability that has been done up to this point as well as the different definitions and metrics that have been developed. Facing the challenge of pinning down one all-encompassing readability metric, we instead identified a combination of automatic measurements for linguistic features of readability, based on past work on readability for our target population (see Fajardo et al., 2013, 2014; Feng, 2009; Feng et al., 2009; Hurtado et al., 2014; Štajner et al., 2015, 2012; Yaneva, 2015; Yaneva et al., 2016, 2017; Yaneva & Evans, 2015). One study examined whether or not English Easy-Read texts on the internet complied with accessibility standards by performing an automatic analysis, using linguistic features of readability as proxies for a selection of writing rules (Yaneva, 2015). Drawing on Yaneva (2015), we adopt a slight adaptation of these proxies to test **H3.0** using the computational tool Coh-Metrix 3.0. For the reasons described in *Section 3.5.2*, we have excluded segments with accuracy and fluency errors from this analysis.

**Coh-Metrix tool**

The Coh-Metrix project at the University of Memphis began in 2002, and over the course of nine years researchers worked on refining and testing their system. The original goal was to develop an automatic measurement tool for text cohesion, but the final product was a broad, multilevel analysis tool including but not limited to lexical and syntactical components (McNamara et al., 2014). Crossley et al. (2007) validated the Coh-Metrix approach for predicting reading difficulty by comparing it to traditional readability formulas. But unlike traditional readability formulas, Coh-Metrix allows for an analysis that goes beyond shallow metrics such as word and sentence length and considers other factors of readability such as text cohesion. In practice, Rossetti (2019) used the tool to evaluate seven different variables of readability in her work on Cochrane plain language summaries, including referential cohesion and deep cohesion, which measure the overlap of ideas across a text and the use of connectives that tie ideas together, respectively. Because our data set came from many different sources and did not contain texts in their entirety cohesion metrics

would not have been accurately represented, however this could be an interesting possibility for a follow-up study.

### Automatic evaluation experimental design

In *Step 4* of the research, our corpus of pre-edited and non-pre-edited E2R segments previously translated into English by DeepL NMT was analyzed on the basis of over 100 indicators using Coh-Metrix 3.0. Although the focus of this step is on readability, due to the close relationship between readability and accessibility, certain indicators are more interesting than others for our analysis. We selected a sample of E2R writing guidelines that could be easily quantified and measured with Coh-Metrix, as presented by Yaneva (2015), to include in our discussion. *Figure 4.4* summarizes the indicator(s) associated with each guideline, described in more detail below.

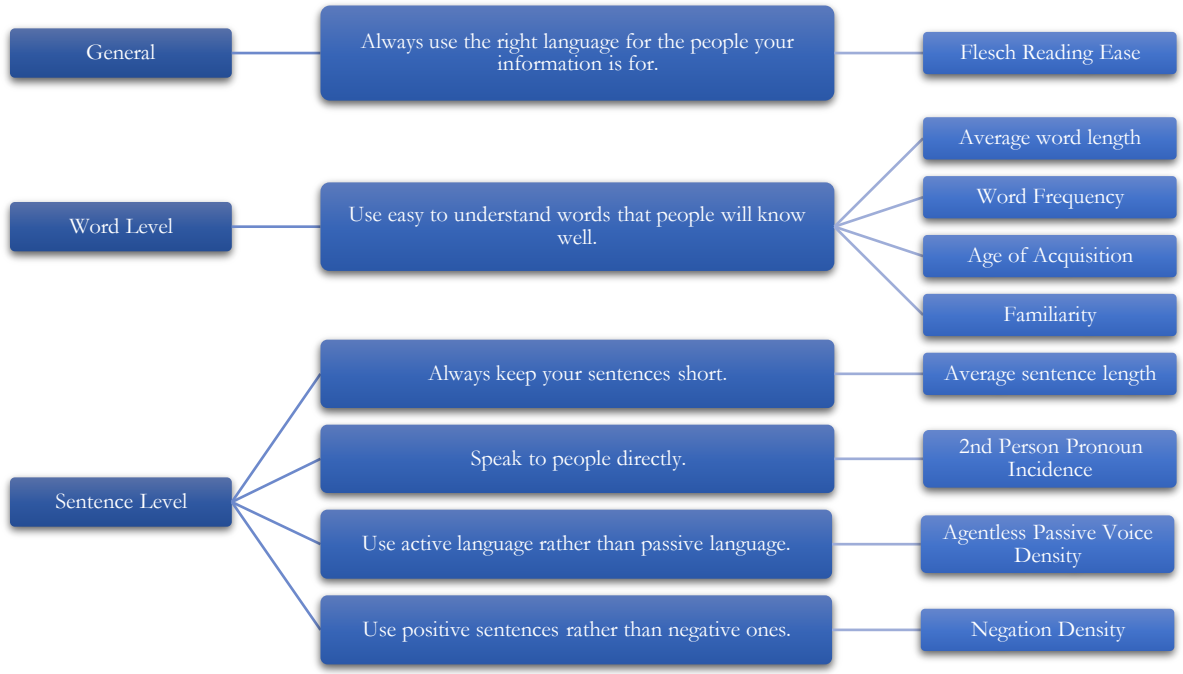



*Figure 4.4: Readability indicators measured in this study. From left to right: Easy-to-Read guideline high-level categories (ours), Easy-to-Read guidelines (Inclusion Europe), corresponding linguistic indicators (Coh-Metrix 3.0 metrics).*

**Flesch Reading Ease**: Traditional readability measure, introduced in our discussion of readability in *Section 3.3.1,* that provides a score of 0.0-100.0, with higher scores indicating a higher ease of understanding. A document that obtains a score of 30 is considered "very difficult," a score of 70 indicates that a text is "easy," and texts with a maximum score of 100 are considered readable by people who are "barely 'functionally literate,'" which, in 1949 around the time that the metric was developed, made up about 93% of American adults (Dubay, 2004, p. 21). Although the measurement is based on sentence length and syllables per word, a metric that Feng et al. (2009) argue may not be as useful in determining comprehension for adults with ID as they are for typical

readers, Štajner et al. (2012) focused on obtaining a Reading Ease score of 90 or higher, which corresponds to the equivalent of an American 5th grade reading level,[42] in their development of the FIRST project.

**Average word length**: Measures the mean number of syllables in a word. Shorter words are often considered easier to read (cf. Feng et al., 2009).

**Word Frequency**: Measured by Coh-Metrix using the CELEX database of 17.9 million English words. "Words that occur with a higher frequency are more familiar to the reader and are processed more quickly" (McNamara et al., 2014, p. 73). The guideline that supports the use of this indicator as well as the **Familiarity** indicator is: "Use easy to understand words *that people will know well*" (Inclusion Europe, 2009, emphasis mine), although its real utility for people with ID has been questioned (Feng et al., 2009). We use the Log frequency for our analysis because it has been shown that "word processing time tends to decrease linearly with the logarithm of word frequency rather than with raw word frequency […] because some words (such as the and is) have extremely high frequencies, with minimal incremental facilitation in reading time over words that are common but not nearly as frequent" (Graesser et al., 2004, p. 197).

**Age of Acquisition (AOA)**: Coh-Metrix relies on Gilhooley and Logey's ratings (1980). Higher AOA scores indicate more difficult words because they appear in children's vocabulary later (McNamara et al., 2014).

**Familiarity**: Content word (as opposed to function words, such as prepositions and conjunctions) familiarity, measured on a scale of 100-700 based on adult user ratings from the MRC Psycholinguistic Database. Higher scores indicate less familiar words (McNamara et al., 2014).

**Average sentence length**: Measures the mean number of words in a sentence. Shorter sentences are considered easier to read and understand because they rely less heavily on working memory, a cognitive feature that might affect people with ID more than neurotypical readers (Feng et al., 2009; Graesser et al., 2004).

**2nd Person Pronoun Incidence**: Number of instances of the pronoun "you" per 1,000 words (McNamara et al., 2014). The use of second person pronouns is explicitly instructed in our Easy-to-Read guidelines, which state that a way to speak directly to the reader is to "Use words like 'you'" (Inclusion Europe, 2009, p. 11).

[42] Corresponds to 10-11 years of age.

**Agentless Passive Voice Density**: Measures relative frequency of sentences featuring passive constructions, which are considered more difficult to process than active constructions (Just & Carpenter, 1987, as cited in McNamara et al., 2014, p. 72).

**Negation Density**: Measures incidence of negation, another linguistic feature that negatively impacts processing effort (Just & Carpenter, 1987, as cited in McNamara et al., 2014, p. 72).

We acknowledge the need for a user study measuring comprehensibility to confirm this variable, but publishable translation quality and accessibility, i.e. few to no violations of E2R guidelines, would need to be achieved before adults with ID could reasonably and ethically be asked to evaluate the readability of a supposedly Easy-to-Read text. Due to the significant number of Easy-to-Read violations and translation errors found throughout the previous three studies, the results of which are discussed in *Chapter 5*, more work would need to be done before that would be possible.

## 4.4 Summary of methods

*Table 4.4* summarizes the methodology by which each variable was tested, expounded upon in the previous sections of this chapter.

| Research Question | Dependent Variable | Hypothesis | Evaluation Type & Tool | Indicators |
|---|---|---|---|---|
| **RQ1**: How does the removal of line breaks from Easy-to-Read French-language administrative documents during the pre-editing process influence the **translation quality** of English output produced by a generic NMT system? | Quality | **H1.0**: Removing forced line breaks from French Easy-to-Read texts will **improve the quality** of English NMT output. | [**Phase 1; Step 1**] Human: DQF-MQM error typology annotation (*Section 4.3.1.1.1*) | • Error type<br>○ Accuracy<br>○ Fluency<br>○ Terminology<br>○ Style<br>○ Design<br>○ Locale convention<br>○ Verity<br>• Error severity<br>Critical, major, minor, neutral, kudos |
| | | | [**Phase 1; Step 2**] Human: MateCat post-editing effort measurement (*Section 4.3.1.1.2*) | • Time-to-edit (TTE)<br>• Post-editing effort[10] (PEE)<br>• Secondary indicators: Error type, severity |
| **RQ2**: How does the removal of line breaks from Easy-to-Read French-language administrative documents during the pre-editing process influence the **accessibility** of English output produced by a generic NMT system? | Accessibility | **H2.0**: Removing forced line breaks before performing NMT with DeepL will improve **text accessibility**. | [**Phase 1; Step 3**] Human: Easy-to-Read guideline adherence annotation (*Section 4.3.1.2.1*) | • Violation type<br>○ General<br>○ Word-level<br>○ Sentence-level<br>○ Structural |
| **RQ3**: How does the removal of line breaks from Easy-to-Read French-language administrative documents during the pre-editing process influence the **readability** of English output produced by a generic NMT system? | Readability | **H3.0**: Removing forced line breaks before NMT will **improve readability**. | [**Phase 2; Step 4**] Automatic: Coh-Metrix 3.0 (*Section 4.3.2.1*) | • General readability<br>○ Flesch Reading Ease<br>• Word-level readability<br>○ Average word length (syllables)<br>○ Word Frequency<br>○ Age of Acquisition<br>○ Familiarity<br>• Sentence-level readability<br>○ Average sentence length (words)<br>○ 2nd Person Pronoun Incidence<br>○ Agentless Passive Voice Density<br>○ Negation Density |

*Table 4.5: Summary of the research questions, dependent variables, primary hypotheses, evaluation methods, and indicators*

[43] Term used by MateCat, which corresponds to the "technical" dimension of post-editing effort in our definition, based on Krings's three-part description (2001); *Section 4.3.1.1.2.*

# Chapter 5: Findings and Discussion

## 5.1 Introduction

The three dependent variables that make up what we call linguistic accessibility – translation quality, accessibility, and readability – were tested over the course of four steps using the methodology described in *Chapter 4 (*see *Section 4.3).* After this introduction (*Section 5.1*), the rest of this chapter addresses the findings from our two-phase experiment on the linguistic accessibility of pre-edited and non-pre-edited segments that were translated using a generic neural machine translation (NMT) system. First, in *Section 5.2*, we discuss the results of *Step 1* and *Step 2*, which examine translation quality from two different angles: translation errors and post-editing effort. Next, in *Section 5.3*, we present the results of *Step 3*, a two-part process for evaluating our accessibility variable with Easy-to-Read (E2R) indicators. *Section 5.4* looks at the findings from our fourth and final step (*Step 4*), an automatic evaluation of various readability indicators. Finally, a brief summary of the findings and their implications is proposed in *Section 5.5*.

## 5.2 Translation quality

Translation quality was examined in two of the four steps of this research. The first step was an error annotation process performed by the researcher and the second was a post-editing study involving six participants, either translators in training or recent graduates from translation master's programs. These steps aimed to answer **Research Question 1** (**RQ1**) by testing **Hypothesis 1.0**:

> **H1.0**: *Removing forced line breaks from French Easy-to-Read texts will improve the quality of English NMT output.*

The findings from these steps are presented and discussed in *Sections 5.2.1* and *5.2.2*, respectively.

### 5.2.1 Step 1: Error annotation

Our first translation quality assessment step was carried out by the researcher using the Dynamic Quality Framework and Multidimensional Quality Metrics (DQF-MQM) error cataloging system described in *Section 3.2.2.1*. A short excerpt of the results from this step and from *Step 3,* which compares the translations and how they were annotated by the researcher, can be found in

*Appendix F*. It is worth noting that DeepL Translator produced identical translations, **or in other words pre-editing had no effect, for 195 of the 849 segments**.

#### 5.2.1.1 Error prevalence and severity

With **H1.1** and **H1.2**, we predicted that the number and severity of errors would be positively impacted by the pre-editing step:

> **H1.1**: *The segments that were pre-edited to remove manual line breaks prior to translation with DeepL NMT will contain* ***fewer translation errors*** *than the segments that contain manual line breaks.*
>
> **H1.2**: *The segments that were pre-edited to remove manual line breaks prior to translation with DeepL NMT will contain* ***less serious translation errors*** *than the segments that contain manual line breaks.*

Of the 849 segments evaluated, we found zero DQF-MQM errors in 358 of the pre-edited segments, compared to just 201 of the non-pre-edited segments. However, some segments contained multiple errors. Overall, 634 errors were flagged in the pre-edited test set, for a prevalence of 74.68% or nearly 3 in 4 segments. An average of more than one error per every non-pre-edited segment was found, a prevalence of 109.66% (931/849). This amounts to a **31.90% decrease in errors** that is directly related to the absence of manual line breaks, **results that would appear to support H1.1**. A one-tailed dependent t-test was chosen for statistical analysis based on the methodology of our study: within-group design, one independent variable, hypothesis that indicates direction (as a reminder, we predicted that pre-editing would *reduce* the number of errors in the translation) (Lazar, 2017). This significance test reveals a p-value of much less than 0.05, indeed providing strong evidence that we can reject the null hypothesis. These results are summarized in *Table 5.1*.

| | **Non-Pre-Edited** | **Pre-Edited** | ***t*-value** | ***p*-value** |
|---|---|---|---|---|
| **Errors per segment, Mean** | 1.096 | 0.746 | 14.60399356 | <0.001 |

*Table 5.6: Mean errors in non-pre-edited and pre-edited segments, t-value and p-value obtained from one-tailed dependent t-test*

No critical errors – errors that according to the TAUS Quality Dashboard (2016) "may carry *health, safety, legal or financial implications* … or which could be seen as offensive" – were found, and neutral errors – errors that reflect only the annotator's stylistic preferences – accounted for less than 2% of all errors for each of the test sets. All neutral errors had to do with using the person pronoun

"he" where a more inclusive form such as "he or she," "they," or simply a repetition or rephrasing of the noun, could have been employed instead. Curiously, although the French source text only ever featured "il" when referring to a general population that could be male or female, this word was translated as "he or she"/"his or her" 66.67% of the time (26 of 39 possible instances) in the pre-edited test set and 19.05% of the time (8 of 42 possible instances, including once erroneously when the text clearly referred to a specific male example) in the non-pre-edited test set. **The majority of the errors found in both test sets were of minor severity** (536 or 57.57% of all errors in the non-pre-edited test set and 382 or 60.25% of all errors in pre-edited test set), defined by the DQF-MQM framework as, "Errors that *don't lead to loss of meaning* and *wouldn't confuse or mislead* the user but would be noticed, would *decrease stylistic quality, fluency or clarity*, or would make the content less appealing" (TAUS, 2016). A lower percentage of errors were considered major in the pre-edited dataset than in the non-pre-edited dataset. We conducted another one-tailed dependent t-test to find out if these results could be considered significant. As displayed in *Table 5.2*, this test **does allow us to validate H1.2**. We suspect that the definitions of severity are fairly subjective, so these results would ideally need to be confirmed by a larger sample of annotators.

| | **Non-Pre-Edited** | **Pre-Edited** | ***t*-value** | ***p*-value** |
|---|---|---|---|---|
| **Major errors per segment, Mean** | 0.452 | 0.289 | 8.23975443 | <0.001 |

*Table 5.7: Mean major errors in non-pre-edited and pre-edited segments, t-value and p-value obtained from one-tailed dependent t-test*

*Table 5.3* provides a comparison of the number of errors found in each test set, as well as a breakdown by severity level.

| | **Total Errors** | | **Neutral** | | **Minor** | | **Major** | | **Critical** | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Non-Pre-Edited** | 931 | 100% | 12 | 1.29% | 536 | 57.57% | 383 | 41.14% | 0 | 0% |
| **Pre-Edited** | 634 | 100% | 7 | 1.10% | 382 | 60.25% | 245 | 38.64% | 0 | 0% |

*Table 5.8: Comparison of number and severity of errors, in both absolute value and percentage of the total, for pre-edited and non-pre-edited segments*

#### 5.2.1.2 Error type

We were also interested in the types of errors most impacted by the pre-editing process and formulated the following hypothesis, to be tested during the same error annotation task:

**H1.3**: ***Fluency*** *and* ***style*** *will be the two categories most positively affected by this pre-editing process.*

As shown in *Table 5.4*, the number of fluency errors was nearly three times as high for the non-pre-edited dataset as it was for the pre-edited dataset (156 vs. 56) and represented roughly twice as high of a proportion of total errors (16.76% vs. 8.83%).

| | Non-Pre-Edited | | | | Pre-Edited | | | |
|---|---|---|---|---|---|---|---|---|
| | **Minor** | **Major** | **Total** | **% of Total** | **Minor** | **Major** | **Total** | **% of Total** |
| **Accuracy** | 122 | 109 | 231 | 24.81% | 91 | 47 | 138 | 21.77% |
| **Fluency** | 103 | 53 | 156 | 16.76% | 52 | 4 | 56 | 8.83% |
| **Terminology** | 92 | 166 | 258 | 27.71% | 61 | 144 | 205 | 32.33% |
| **Style** | 218 | 56 | 286 | 30.72% | 178 | 50 | 235 | 37.07% |

*Table 5.4: Error type breakdown for non-pre-edited and pre-edited segments by absolute value and percentage of total errors for each test set*

Within the fluency category, grammar was the most frequent error subcategory; we reported 122 instances of ungrammatical segments in the non-pre-edited data and 46 in the pre-edited data, however these represent roughly the same proportion of the total respective fluency errors (78.21% for non-pre-edited and 82.14% for pre-edited). There was, however, a marked improvement in another fluency sub-category: punctuation. Punctuation errors made up 17.31% of all fluency errors in non-pre-edited segments, a number which dropped to 8.93% after pre-editing. We observed that DeepL struggles with sentence boundary disambiguation when a sentence is split onto multiple lines via forced line breaks (in our examples, line breaks are represented by forward slashes), sometimes adding punctuation in the middle of a segment, such as:

| Source | Non-Pre-Edited | Pre-Edited |
|---|---|---|
| Dès que la MDPH reçoit votre demande de recours elle vous envoie un accusé de réception / de votre demande. | As soon as the MDPH receives your appeal request, it sends you an acknowledgement of receipt**.** of your request. | As soon as the MDPH receives your appeal request, it sends you an acknowledgement of receipt of your request. |

*Table 5.5: Example that demonstrates how forced line breaks can produce unexpected results in translation, such as added punctuation*

Our findings regarding style errors, which only decreased by 17.83% after pre-editing, were particularly surprising. As Karreman et al. (2007) and Schmutz et al. (2019) reported, the style of E2R can sometimes be off-putting for readers without disabilities since is quite different from standard English. Without a follow-up investigation we cannot say with certainty whether a given style issue stemmed from the style of the FALC (*Facile à lire et à comprendre*) source text or was caused by forced line breaks.

Terminology was the only category to produce more major errors than minor errors; these errors were flagged as major 58% of the time for non-pre-edited segments and 70.24% of the time for pre-edited segments. Though more severe a majority of the time, the researcher observed that they were also the easiest to spot, particularly in texts of the administrative genre which tend to include many acronyms. In fact, we observed 93 acronym-related terminology errors in the non-pre-edited data versus 74 in the pre-edited data. *Table 5.6* shows an example of how terminology was sometimes the only thing that changed in the non-pre-edited and pre-edited translations:

| Source | Non-Pre-Edited | Pre-Edited |
|---|---|---|
| Le juge peut aussi prendre la même décision que la MDPH. | The judge can also make the same decision as the **MDPH**. | The judge may also take the same decision as the **CDM**. |

*Table 5.6: Example that demonstrates how line breaks could affect translation of terminology*

An avenue of future research could involve the new DeepL glossary feature, which was not yet available when this study was carried out, to determine whether it is an effective way of limiting terminology errors. Wang et al. (2017) raise the issue of terminological consistency in their research on cross-sentence context in NMT and propose a model that takes previous source sentences from the same document into consideration when determining the best output. DeepL does have a full document translation feature, though whether or not it incorporates technology similar to Wang et al.'s (2017) approach is unknown. With this in mind, it is possible that translating full documents rather than isolated sentences, as was the case in this thesis, could improve consistency, however this was not a logical option for our study since the test data was comprised of many different source documents.

Two other terminology issues particular to E2R are: the speed at which language surrounding disability changes, for instance the term "mental retardation," which has not only been replaced by "intellectual disability" within the last decade[44] but which has also quickly become derogatory;[45] as well as the differences in rate of change and adoption that seem to exist between English-speaking and French-speaking societies, such as the idea of people-first language. Performing a Google Books N-gram Viewer inquiry, displayed in *Figures 5.1 and 5.2,* showed that "people with disabilities" overtook "disabled people" in frequency in the 2019 version of the English corpus in 1990, but "*personnes en situation de handicap*" (which first appeared in 1987) did not become more

[44] https://www.federalregister.gov/documents/2013/08/01/2013-18552/change-in-terminology-mental-retardation-to-intellectual-disability Last accessed: October 8, 2020
[45] https://ncdj.org/style-guide/ Last accessed: January 4, 2021

prevalent than "*les handicapés*" in the 2019 version of the French corpus until 2014.[46] We observed that pre-editing also impacted this type of language: 42 instances of "handicapped" or "disabled," deemed inconsistent terminology by the researcher,[47] were found in non-pre-edited segments compared to just 29 in pre-edited segments. This is a surprising finding, since in theory, line breaks should have an impact on syntax, rather than terminology. This could be because line breaks are interpreted as end of sequence markers, and therefore less context is available from the rest of the sentence in non-pre-edited segments to help the technology select the best translation. This is one key disadvantage of neural machine translation in general; unlike rule-based machine translation, where rules and preferences can be hard coded, and statistical machine translation, where source segments that produce certain translations or errors can easily be found and removed from the corpus, neural machine translation is much more opaque and it is difficult to pinpoint and rectify problematic data introduced during training.

[46] https://books.google.com/ngrams/ Last accessed: October 8, 2020

[47] Not all people with disabilities prefer this terminology. No matter what terminological choices are made, however, consistency across communication is key.

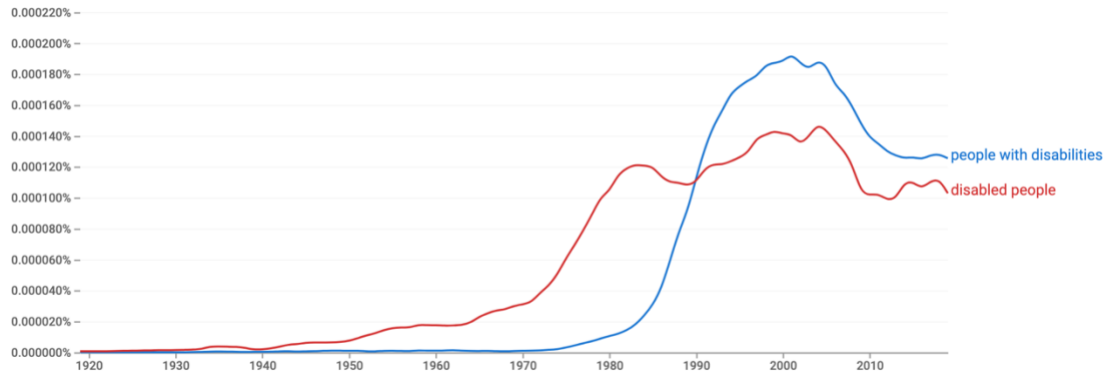



*Figure 5.9: Google Books N-gram analysis (1919-2019) of English terms in the disability lexicon*

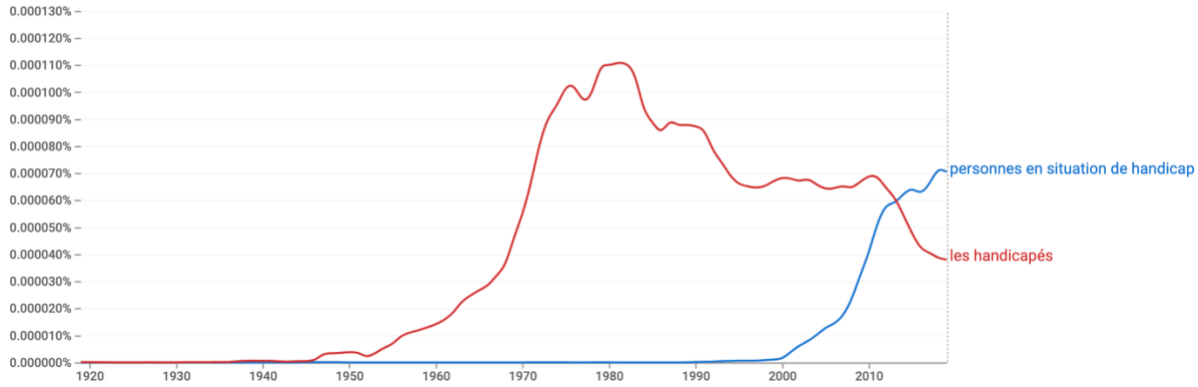



*Figure 5.10: Google Books N-gram analysis (1919-2019) of French terms in the disability lexicon*

No errors were found for either test set in the Design, Locale Convention, Verity, and Other categories, so they were not included in *Table 5.4*.

In conclusion, **Hypothesis H1.3 is only partially supported by these findings**. Fluency was the category most positively affected by pre-editing, experiencing a 64.10% decrease in total errors as well as a decrease in proportion of errors of roughly half (8.83%). The style category experienced a smaller decrease in total errors, of just 17.83%, and was the only category negatively impacted in terms of the proportion of total errors that it represented, increasing from 30.72% to 37.07% after pre-editing.

#### 5.2.1.3 Discussion and limitations of Step 1

Although we observed a marked decrease in errors, the number of errors found even when our pre-editing step was performed is still unacceptable for publishing purposes. This confirms a need for better pre-editing and likely also machine translation post-editing (see *Section 5.2.2*) if NMT with a generic system such as DeepL is to be a suitable method for E2R translation and text production.

Collecting data from only one annotator (the researcher) is a limitation of this study, particularly since the fairly low inter-rater reliability between the three participants in the exploratory study cast doubt on the objectivity of the DQF-MQM framework of evaluation. Inter-annotator agreement (IAA) is notoriously difficult to obtain due to the complexity of language; as Lommel et al. (2014, p. 36) conclude:

> Human annotators' meta-understanding of language is quite variable, even when working with professional translators. Even with an analytic framework and guidelines there is significant, and perhaps unavoidable, disagreement between annotators. To a large extent this disagreement reflects the variability of human language.

For instance, the lack of an official termbase, as was the case in this study, could lead some evaluators to classify an error as a mistranslation rather than a terminology error and vice versa, or to not classify it as an error at all. Additionally, there was the issue of regional differences; not having a clear idea of the target audience (primarily UK, primarily US, or primarily international) led to a bias toward US spelling, dialect-specific vocabulary, and date conventions. Presumably, the new DeepL regional English selector, which had not yet been released when translation was carried out for this study, could resolve some of this uncertainty. As mentioned in *Section 4.3.1.1.1 Error annotation experimental design*, another limitation of this study is that the annotator was the same person who performed the translation with DeepL, and because sentences were not randomized for ease of analysis; due to the layout of E2R, i.e. numerous bulleted lists and short sentences, context from surrounding segments was often needed to determine whether an error is present or not. The large corpus size and a lack of financial resources made it unfeasible to recruit other participants for such a time-consuming task, however we attempted to combat this potential bias by leaving sufficient time between translation and annotation and masking identifying information in the file names. Another way that we tried to counterbalance these limitations was by adding a second study to assess translation quality with multiple participants, the findings from which are presented and discussed in the next section (*Section 5.2.2*).

### 5.2.2 Step 2: Post-editing

In an attempt to obtain a more thorough answer for our first research question (**RQ1**), we conducted a second **translation quality** assessment step consisting of post-editing using the web-based translator workbench MateCat. *Step 2* aimed to test **Hypothesis 1.4**:

*The segments that were pre-edited (i.e. manual line breaks were removed) prior to translation will require* ***less post-editing effort*** *to achieve publishable quality than the segments that were not pre-edited, and that therefore contained manual line breaks, when translated with DeepL NMT.*

We focused our attention on three main aspects of the data collected with MateCat: Post-Editing Effort (PEE), Time-to-Edit (TTE), the final output ("Revision") produced by the post-editors. Results based on the first two types of data will be explored in this section, as they help us measure translation quality. The third type of data will be analyzed in *Section 4.3.1.2* to help us measure accessibility. Because post-editing time and effort can vary so much based on an editor's personal experience, we were interested in comparing results within each post-editor's individual dataset in addition to comparing results across the board.

#### 5.2.2.1 Temporal measurement: Post-editing time

The first indicator of translation quality is post-editing time, or TTE. *Table 5.7* shows an example of a non-pre-edited (again, line break represented by forward slash) and a pre-edited segment and the NMT output that was produced for each, and the corresponding TTE for our six post-editors. DeepL seems to have interpreted the line break as an end-of-sentence marker; in absence of the rest of the sentence, the first clause ("Demander une conciliation à la MDPH") is the typical structure of a header or title, and therefore the first verb could be translated in the present simple ("request," as it was translated in the non-pre-edited segment) or the present continuous ("asking," as it was translated in the pre-edited segment). However, in presence of the rest of the sentence following the line break, only the present continuous is acceptable. The TTE results for some segments, including the one shown in *Table 5.7*, were surprising. Clearly, there is a grammatical issue to resolve in the non-post-edited segment, and post-editors were also required to consult the glossary provided to determine how to deal with the acronym, which was incorrect in the non-pre-edited segment. Therefore, we would expect editors to spend less time working on the pre-edited segment. Yet this was only the case for two of the six post-editors. This result raises the question of how a lack of post-editing experience or subject matter experience could factor into TTE.

| | Non-Pre-Edited | Pre-Edited |
|---|---|---|
| **Source (FR)** | Demander une conciliation à la MDPH / vous permet d'avoir un rendez-vous avec une personne. | Demander une conciliation à la MDPH vous permet d'avoir un rendez-vous avec une personne. |
| **Target (EN)** | **Request** conciliation **from** the **CDM** allows you to have an appointment with a person. | **Asking for** conciliation **at** the **MDPH** allows you to have an appointment with a person. |
| **TTE (PE1/2/3/4/5/6)** | 53s / 30s / 177s / 33s / 10s / 546s | 68s / 0s / 43s / 64s / 17s / 1053s |

*Table 5.9: Example of NMT output of non-pre-edited and pre-edited segments with corresponding TTE*

*Figure 5.3* compares the mean TTE in seconds for all 25 non-pre-edited segments and all 25 pre-edited segments for each post-editor as well as for the whole group, including the standard deviation for each set of segments. Clearly, the results were extremely varied from editor to editor; the mean editing time for pre-edited segments ranged from 13.52s (0s for "perfect" segments that required no post-editing) to 261.76s, and from 24.12s (again, 0s for untouched segments) to 275.20s for non-pre-edited segments. Two-thirds (N=4) of post-editors spent a lower average time working on pre-edited segments than they did on non-pre-edited segments. For the post-editors who, on average, spent longer working non-pre-edited segments than pre-edited segments (PE1 and PE3), the difference in mean post-editing time for the two sets of segments was minimal: less than 2 seconds. The mean TTE for the group was 94.2s/segment for non-pre-edited segments and 81.59s/segment for pre-edited segments. Overall, editors **saved an average of 12.61 seconds** per segment when editing segments in which line breaks were removed before translation; in other words, the pre-editing process resulted in a **post-editing time gain of 13.38%**.

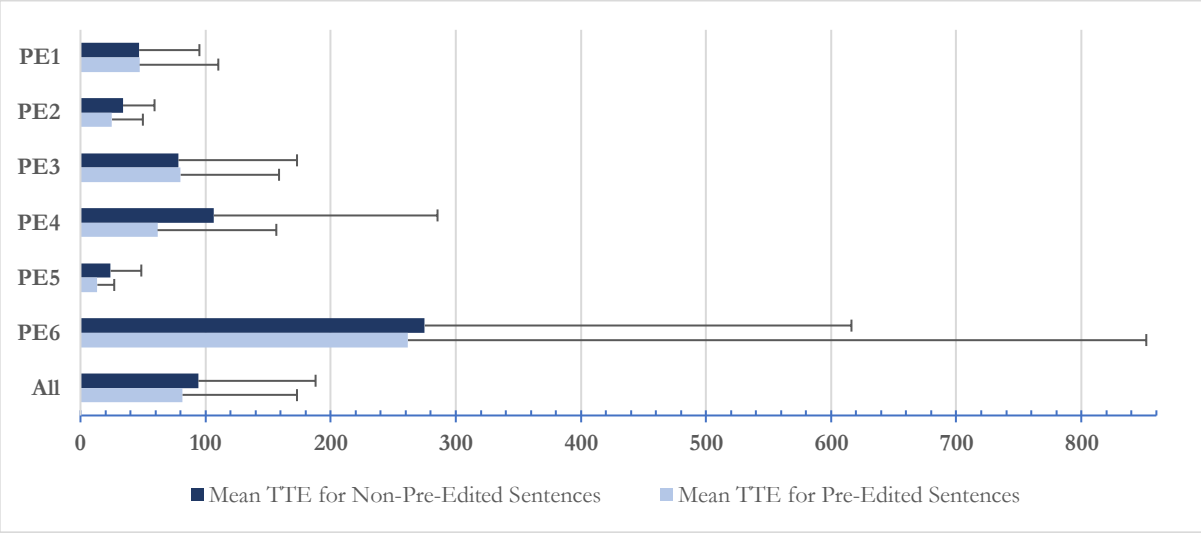



*Figure 5.11: Post-editor comparison of mean Time-to-Edit (TTE) in seconds for non-pre-edited vs. pre-edited segments*

Using a standard outlier calculation (Q1 - 1.5*IQR and Q3 + 1.5*IQR), we did observe several outliers in the data from one post-editor in particular, PE6. This person reported 20 instances (out of the 33 total outliers in 2*150 segments) of a TTE exceeding the upper bounds of 142.5 seconds for pre-edited segments and 182.38 seconds for non-pre-edited segments – 4 of which exceeded 15 minutes and one of which exceeded 45 minutes. We did instruct post-editors to close the webpage if they needed to take a break, to ensure that they clock was only running when they were actively working on the segments; PE6 may not have followed these instructions, however there is no way to be sure whether they were actually working on the segments (for instance, researching the topic) with exceedingly long TTE or not. PE6's total time spent on the first file they processed was 175.55 minutes, whereas the second file only took them 48.18 minutes or over 3.5 times shorter, which supports our decisions to split up the pre-edited and non-pre-edited segments evenly within each file and to give half of the participants one file first, and the other half the other file first. This post-editor also flagged a higher number of errors than any of the other editors, which could have contributed to their longer editing times. PE6 flagged 33 errors for pre-edited segments and 39 for non-pre-edited segments, compared to the group average of 23 for pre-edited and 28.33 for non-pre-edited. Excluding this person's data, the overall mean for post-editors 1 through 5 drops to 58s for non-pre-edited segments and 45.55s for pre-edited segments, a **time gain of 12.45 seconds per segment or 21.47%**.

*Table 5.8* shows a breakdown of post-editing time results per segment and per editor. Theoretically, a very long TTE for one non-pre-edited segment could skew the mean in favor of pre-edited segments, however this breakdown of results shows that that was probably not the case. All but two of the six editors spent less time on the pre-edited version than on the non-pre-edited version of the same segment for a majority of segments. One editor (PE3) spent less time on the non-pre-edited version more often, and the results were split for another editor (PE1). Overall, editors **spent less time on the pre-edited version 52.67% of the time** versus less time on the non-pre-edited version 38.67% of the time. Almost 1/10 (8.67%) of the time, they spent an equal amount of time on each version; in almost all cases, these were segments that post-editors spent 0 seconds revising.

| | # of Pre-Edited Segments with Shorter TTE than Non-Pre-Edited Counterpart (of 25) | # of Non-Pre-Edited Segments with Shorter TTE than Pre-Edited Counterpart (of 25) | # of Pre-Edited and Non-Pre-Edited Segments with Equal TTE (of 25) |
|---|---|---|---|
| **PE1** | 11 | 11 | 3 |
| **PE2** | **15** | 5 | 5 |
| **PE3** | 9 | **15** | 1 |
| **PE4** | **14** | 11 | 0 |
| **PE5** | **15** | 6 | 4 |
| **PE6** | **15** | 10 | 0 |
| **All** | **79** | 58 | 13 |

*Table 5.10: Distribution of segments with shorter Time-to-Edit (TTE), which represents higher post-editing productivity, per post-editor and in total (higher values marked in bold)*

Due to the extreme TTE values in our data, it is more useful to compare median values than mean values, particularly for determining the statistical significance of this experiment. Doing so revealed a median TTE of 37.0 seconds for non-pre-edited segments and 30.5 seconds for pre-edited segments, as shown in *Figure 5.4*. The data points shown are the outliers, with an indication that one extreme outlier lies beyond the bounds of the plot for clarity purposes.

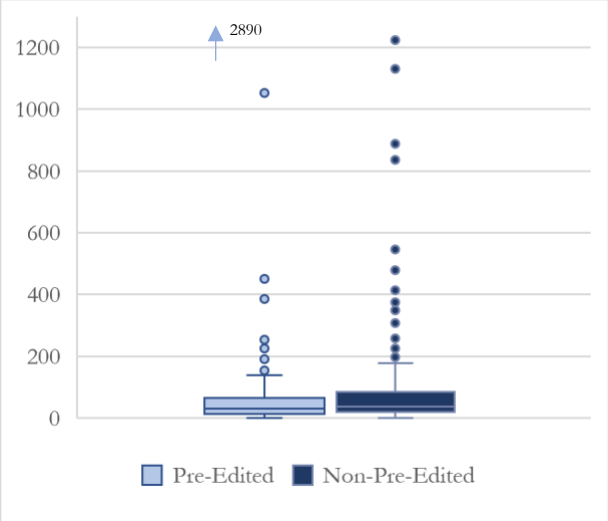



*Figure 5.12: Median Time-to-Edit (TTE) for pre-edited and non-pre-edited segments in seconds*

The p-value of 0.13, obtained using a Mann-Whitney U test (Lazar, 2017; Mann & Whitney, 1947) with a type I error of 5% (*Table 5.9*) **does not provide strong enough statistical significance to allow us to reject the null hypothesis** that pre-edited and non-pre-edited segments require the same amount of effort to post-edit when measured in TTE.

| | Non-Pre-Edited | Pre-Edited | P-value |
|---|---|---|---|
| **Time-to-Edit (TTE) in seconds, Median (Q1-Q3)** | 37.0 (19.2-84.5) | 30.5 (15.0-66.0) | 0.13 |

*Table 5.11: Median TTE for non-pre-edited and pre-edited segments for all participants, and p-value obtained from Mann-Whitney U test*

#### 5.2.2.2 Technical measurement: Post-editing effort

We anticipated the factors that could make post-editing time a less reliable measure of translation quality, such as lack of subject matter as well as post-editing experience and lack of familiarity with the post-editing environment, and therefore also analyzed post-editing effort from a technical point of view. *Table 5.10* shows the PEE for the same example segment used in *Table 5.7*. Although only one-third (N=2) of post-editors spent less time editing this particular pre-edited segment, regardless of the time spent thinking about the segment or performing research for it, five of the six participants ended up making fewer changes to the pre-edited segment, indicating higher translation quality.

| | Non-Pre-Edited | Pre-Edited |
|---|---|---|
| **Source (FR)** | Demander une conciliation à la MDPH / vous permet d'avoir un rendez-vous avec une personne. | Demander une conciliation à la MDPH vous permet d'avoir un rendez-vous avec une personne. |
| **Target (EN)** | **Request conciliation from** the **CDM** allows you to have an appointment with a person. | **Asking for conciliation at** the **MDPH** allows you to have an appointment with a person. |
| **PEE (PE1/2/3/4/5/6)** | 37% / 16% / 76% / 30% / 26% / 59% | 35% / 0% / 82% / 13% / 10% / 44% |

*Table 5.12: Example of NMT output of non-pre-edited and pre-edited segments with corresponding PEE*

*Figure 5.5* compares mean PEE in percentage of a raw segment that was amended for all 25 non-pre-edited segments and all 25 pre-edited segments for each post-editor as well as for the group, including the standard deviation for each set of segments. For this indicator, **all six post-editors experienced higher average productivity (i.e. made fewer changes) for pre-edited segments than for non-pre-edited segments**. The mean proportion of segments that was changed in the post-editing process was 20.02% for non-pre-edited segments and 15.63% for pre-edited segments.

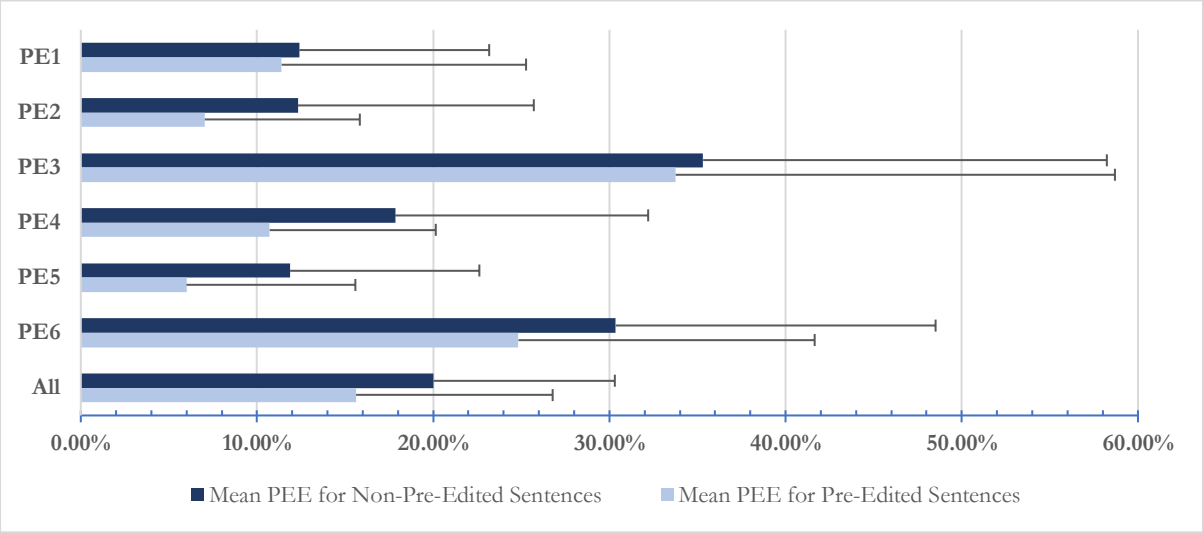



*Figure 5.13: Post-editor by post-editor comparison of mean Post-Editing Effort (PEE) in percentage of segment modified for non-pre-edited vs. pre-edited segments*

Once again, the breakdown by number of segments in *Table 5.11* supports these findings; 91, or **60.67%** of, pre-edited segments were modified less than the non-pre-edited version of the same segment, compared to 32 (21.33%) segments with the opposite findings, and 27 (18%) segment pairs with equal PEE.

| | # of Pre-Edited Segments with Lower PEE than Non-Pre-Edited Counterpart (of 25) | # of Non-Pre-Edited Segments with Lower PEE than Pre-Edited Counterpart (of 25) | # of Pre-Edited and Non-Pre-Edited Segments with Equal PEE (of 25) |
|---|---|---|---|
| **PE1** | **13** | 5 | 7 |
| **PE2** | **13** | 3 | 9 |
| **PE3** | **14** | 9 | 2 |
| **PE4** | **18** | 5 | 2 |
| **PE5** | **14** | 4 | 7 |
| **PE6** | **19** | 6 | 0 |
| **All** | **91** | 32 | 27 |

*Table 5.13: Distribution of segments with lower Post-Editing Effort (PEE), which represents higher post-editing productivity, per post-editor and in total (higher values marked in bold)*

We obtained a median PEE of 14% for non-pre-edited segments and 9% for pre-edited segments, as displayed in *Figure 5.6*. Once again, the data points shown are outliers, found using the same standard outlier calculation that was used for TTE.

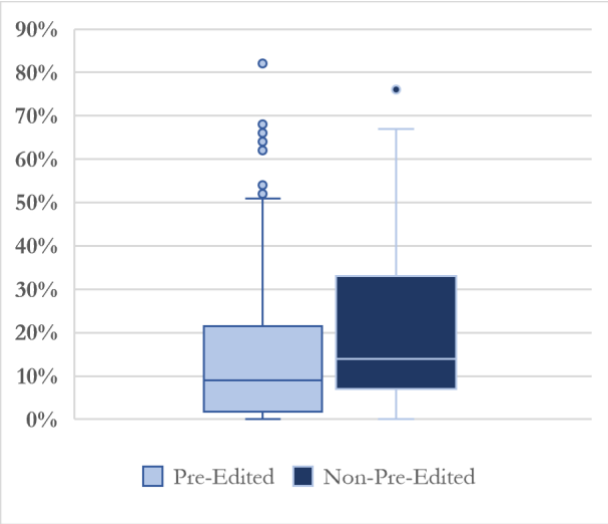



*Figure 5.14: Median Post-Editing Effort (PEE) for pre-edited and non-pre-edited segments, in percentage changed*

We used a Mann-Whitney U test (Mann & Whitney, 1947) with a type I error of 5% to determine the statistical significance of our PEE experiment as well (*Table 5.12*). The highly significant p-value of 0.007 obtained for this test **does allow us to reject the null hypothesis** that pre-edited and non-pre-edited segments require the same amount of effort to post-edit when measured in PEE.

| | **Non-Pre-Edited** | **Pre-Edited** | ***p*-value** |
|---|---|---|---|
| **Post-Editing Effort (PEE) in percentage, Median (Q1-Q3)** | 14.0 (7.0-32.8) | 9.0 (2.0-21.0) | 0.007 |

*Table 5.14: Median PEE for non-pre-edited and pre-edited segments for all participants, and p-value obtained from Mann-Whitney U test*

**The findings from *Step 2* partially support H1.4**: *The segments that were pre-edited (i.e. manual line breaks were removed) prior to translation will require* ***less post-editing effort*** *to achieve publishable quality than the segments that were not pre-edited, and that therefore contained manual line breaks, when translated with DeepL NMT.*

**Translation quality, when measured as post-editing effort, was significantly improved (p-value < 0.05) in terms of the percentage of changes made** when segments were pre-edited to exclude the manual line breaks instructed by E2R guidelines, **but not necessarily in terms of time spent editing**.

#### 5.2.2.3 Discussion and limitations of Step 2

Results for post-editing time could have been skewed by a few factors. First, familiarity with the topic. Since our corpus was built with texts from the administrative genre, explaining different structures and welfare benefit systems that concern people with disabilities, it featured quite a few terms, acronyms, and concepts that the average English speaker – and even many French speakers – would have never heard and that do not have an official English translation yet. Post-editors were allowed to use external resources, though the majority of this field-specific terminology was provided in a glossary. Despite this, several post-editors wrote comments about their doubts on terms that were not given to them. Additionally, the glossary was provided to post-editors as a text file. Had we taken advantage of MateCat's integrated glossary feature, it is possible that post-editors would have spent less time on segments with specific terms and edited them more reliably; for instance, we observed that editors did not correctly identify all instances of terminology errors that were present in the raw translations. Post-editors' screens were not recorded, and feedback was not systematically collected regarding the particular difficulties faced, so we are unable to know how each participant performed the task. Future work could provide insight into what resources editors consulted by carrying out this study in a more controlled environment.

We recognize that our post-editing study was not conducted under the same circumstances that normal post-editing would be when performed by professional translators. For one, post-editors did not know that they were editing Easy-to-Read language, simply because training the editors to produce E2R would have been too time-consuming and limiting our candidate pool to people who were already experienced at producing this type of writing would have been too restrictive. Second, post-editors did not have entire documents for context. Segments from different documents and related to different topics were mixed within the sample provided, and several editors commented that they could not be sure whether a structure was correct or not without knowing what came before it. We also did not account for the time it would take to insert new line breaks at the appropriate spots; we did not ask post-editors to perform this task because it would generally be done during the desktop publishing (formatting) phase of text production.

## 5.3 Accessibility

The findings reported in this section aim to answer our second Research Question (**RQ2**), regarding the text accessibility branch of our definition of linguistic accessibility. As explained in *Chapters 2* and *3*, the definition of accessibility that we have chosen to base our quantitative analysis on is directly tied to Inclusion Europe's set of E2R standards.

### 5.3.1 Step 3: Easy-to-Read violation annotation

The presence or absence of E2R guideline violations, which we measured in *Step 3* of our investigation, gives us one idea of the level of accessibility of the English translation; theoretically, if a text complies with these standards, meaning that it contains E2R few violations, it should be accessible to a majority of people with intellectual disabilities. We formulated a general hypothesis regarding the impact of pre-editing on accessibility, **H2.0**: *Removing forced line breaks before performing NMT with DeepL will improve text accessibility.*

We acknowledge that people with ID are a heterogenous group of readers and that a reception study would be needed to confirm these findings.

#### 5.3.1.1 Guideline violation prevalence

Of the 849 segments annotated in *Step 3*, we found zero Easy-to-Read violations in 288 of the pre-edited segments and in 212 of the non-pre-edited segments. Similarly to the error annotation in *Step 1*, when an Easy-to-Read violation annotation task was carried out, certain segments were found to have multiple violations. Overall, 961 violations were found in the non-pre-edited test set, which **decreased by 24.04%** after pre-editing, to just 730 violations. The large decrease in E2R violations that occurred when manual line breaks were removed before translation would seem to support **H2.1**:

> **H2.1**: *The segments that were **pre-edited** to remove manual line breaks prior to translation with DeepL NMT will contain **fewer violations of E2R guidelines** than the segments that contain manual line breaks.*

To test the statistical significance of our results, we once again conducted a one-tailed dependent t-test. The high t-value and very low p-value obtained (*Table 5.13*) do indeed allow us to reject the null hypothesis; **when pre-editing is performed to remove forced line breaks in the French source text, the English target text has significantly fewer Easy-to-Read violations**.

| | **Non-Pre-Edited** | **Pre-Edited** | ***t*-value** | ***p*-value** |
|---|---|---|---|---|
| **Violations per segment, Mean** | 1.132 | 0.860 | 10.76234967 | >0.001 |

*Table 5.15: Mean Easy-to-Read violations in non-pre-edited and pre-edited segments, p-value obtained from one-tailed dependent t-test*

Violations decreased almost across the board: by 52.94% in the General category, 14% in the Word Level category, and 16.10% in the Sentence Level category. The only category in which they stayed constant was the Structural category. These results are summarized in the table below (*Table 5.14*).

| | Total Errors | General | Word Level | Sentence Level | Structural |
|---|---|---|---|---|---|
| **Non-Pre-Edited** | 961 | 238 | 514 | 205 | 4 |
| **Pre-Edited** | 730 | 112 | 442 | 172 | 4 |

*Table 5.16: Comparison of number and type of Easy-to-Read violations for pre-edited and non-pre-edited segments*

#### 5.3.1.2 Guideline violation type

We will now highlight some of the most notable findings for the three categories that were improved by the pre-editing step.

**General**

Our General category includes two broad recommendations that concern that document as a whole: use appropriate language for your audience and explain the topic clearly. One of the most common violations that we observed of the first guideline in this category – "Always use the right language for the people your information is for" (Inclusion Europe, 2009, p. 9) – had to do with the disability-related language issues brought up in *Section 5.2.1.2*, which correspond to Terminology errors in the DQF-MQM framework. Next, since E2R assumes that authors will not produce so-called "textual disturbances," such as grammatical errors and extraneous words that would not appear in a typical published document, there are no guidelines that specifically account for these errors that were introduced during NMT. Since these disturbances disrupt clarity and make sentences more difficult to comprehend, as shown in the example in *Table 5.15* they were flagged as violations of the general E2R guideline that stipulates: "*Make sure you explain the subject clearly* and also explain any difficult words to do with that subject" (Inclusion Europe, 2009, p. 9). This was the guideline that experienced the most positive change as a result of pre-editing, falling from 174 errors in the non-pre-edited segments to just 56. These instances were generally classified as grammatical errors in *Step 1*.

| Source | Non-Pre-Edited | Pre-Edited |
|---|---|---|
| Un tribunal est un endroit où travaillent des professionnels / de la justice. | A court is a place where people **work of legal professionals**. | A court is a place where legal professionals work. |

*Table 5.17: Example of a violation of the clarity E2R guideline that can be mapped as a consequence of the line break rule*

**Word level**

In the Word Level category, we found 514 violations before removing line breaks versus 442 after, a positive change, but one that was smaller than expected. In fact, in several instances, the non-pre-edited segment actually resulted in a more E2R-compliant translation with respect to the words chosen; non-pre-edited segments featured 15 *fewer* occurrences of words that violated the guideline that stipulates the use "easy to understand words that people will know well" (Inclusion Europe, 2009, p. 10). One such example is shown in *Table 5.16*. This could be because, due to the line breaks, less context was available to the system to signal typical administrative language. Both are adequate and fluent translations of the source segment, but the pre-edited version reflects the more "sophisticated" (and subsequently less accessible) register that is often preferred in legal and official contexts, as shown by the bolded words. At the same time, this lack of register signaling also likely contributed to the prevalence of contracted forms, underlined in *Table 5.16*, which are an E2R violation; 60 violations of this guideline were flagged in the non-pre-edited data compared to just 13 in the pre-edited segments.

| Source | Non-Pre-Edited | Pre-Edited |
|---|---|---|
| Vous devez joindre une photocopie du deuxième courrier de la MDPH qui dit que vous n'avez pas le droit à l'aide. | You must attach a photocopy of the second letter from the MDPH **that says** you don't **have the right** to **help**. | You must attach a photocopy of the second letter from the MDPH **stating** that you are not **entitled** to **assistance**. |

*Table 5.18: Example of a surprising positive effect of line breaks on administrative language*

**Sentence level**

A majority of E2R violations in our Sentence Level category concerned segments that were too long: 90 of 205 sentence-level violations in the non-pre-edited data and 72 of 172 sentence-level violations in the pre-edited data. The Inclusion Europe instructions are once again vague regarding this aspect – and perhaps intentionally so, since they were designed to be adapted into many languages and some languages need more words than others to express the same ideas. Adding to the difficulty is the fact that we also observed sentences that would not necessarily have exceeded such a word or character limit, but which featured extraneous words that could be omitted without changing the meaning or which could be reformulated in a shorter, more concise manner, such as the example in *Table 5.17*. The translations are 12 and 13 words, respectively, but could be shorter and still appropriately reflect the source: "The MDPH chooses the ESAT where you do your internship."

| Source | Non-Pre-Edited | Pre-Edited |
|---|---|---|
| C'est la MDPH qui choisit l'ESAT où vous faites votre stage. | It is the MDPH that chooses ESAT where you do your internship. | It is the MDPH that chooses the ESAT where you do your internship. |

*Table 5.19: Example of relatively short translations which could be written in an even shorter and more succinct way*

Some violations of the short sentence guideline were already present in the French source text, where a sentence could have been split into two distinct ideas. One such example is shown in *Table 5.18*, where the conjunction "et" signals a second thought, and which produced 31-word English translations. This underpins the need for adequate validation in the source language prior to translation.

| Source | Non-Pre-Edited | Pre-Edited |
|---|---|---|
| Le 7 février 2013, Marie-Arlette Carlotti a visité le Foyer d'Accueil Médicalisé (FAM) Jean Favéris de l'association « Les jours heureux » et dit qu'il y a de plus en plus de personnes handicapées âgées. | On February 7, 2013, Marie-Arlette Carlotti visited the Foyer d'Accueil Médicalisé (FAM) Jean Favéris of the association "Les jours heureux" and says that there are more and more elderly disabled people. | On February 7, 2013, Marie-Arlette Carlotti visited the Foyer d'Accueil Médicalisé (FAM) Jean Favéris of the association "Les jours heureux" and said that there are more and more elderly disabled people. |

*Table 5.20: Example of a long sentence in the French source which was transferred to the English target texts*

The breakdown of errors per category and guideline are summarized in *Table 5.18*.

| Category | Easy-to-Read Guideline | Non-Pre-Edited | Pre-Edited |
|---|---|---|---|
| General | Always use the right language for the people your information is for. | **64** | 56 |
| | Make sure you explain the subject clearly and also explain any difficult words to do with that subject. | **174** | 56 |
| Word Level | Use easy to understand words that people will know well. | 185 | **200** |
| | Do not use difficult words. If you need to use difficult words, make sure you always explain them clearly. | 15 | 15 |
| | Use the same word to describe the same thing throughout your document throughout the document. | **215** | 180 |
| | Do not use words from other languages unless they are very well known. | 1 | 1 |
| | Avoid using initials. If you have to use initials, explain them. | 4 | **6** |
| | Try not to use percentages and big numbers. | 5 | **7** |
| | Be careful when you use pronouns. Make sure it is always clear who or what the pronoun is talking about. | **21** | 9 |
| | Do not use numbers with ordinal indicators or suffixes. | 2 | **3** |
| | Write numbers as digits, not as words. | 6 | **8** |
| | Avoid contractions. | **60** | 13 |
| Sentence Level | Speak to people directly. Use words like "you" to do this. | 8 | **10** |
| | Use positive sentences rather than negative ones where possible. | 4 | 4 |
| | Use active language rather than passive language where possible. | 59 | **72** |
| | Keep the punctuation simple. | **27** | 1 |
| | Where possible, use the present tense rather than the past tense. | **17** | 13 |
| | Always keep your sentences short. | **90** | 72 |
| Structural | Start new sentences on a new line. | **1** | 0 |
| | Use bullet points to list things. | 3 | **4** |

*Table 5.21: Summary of the results of Step 3, the Easy-to-Read violation annotation task (higher values highlighted in bold)*

#### 5.3.1.3 Discussion and limitations of Step 3

Two notable challenges were (i) interpreting the E2R guidelines, for instance determining what constituted a "difficult" word and a "long" sentence, and (ii) annotating violations in a consistent way. The researcher ultimately relied on her own judgement to decide whether (i) a word was too difficult or a sentence too long, and (ii) if there was an easier or shorter way to phrase the same idea. Employing a text analysis tool, such as the LIREC tool described in *Section 3.5.1* or the one

based on the COCA corpus developed by linguist Mark Davies,[48] which can highlight uncommon words and propose a list of alternatives in order of frequency, could reduce the subjectivity of this task. Even with the aid of this analysis, it would be necessary to define a methodology for selection or a cut-off number, which is one reason why we did not use this approach. Consider the word "receive," which has a frequency ranking of 497th, making it quite common in the English language. Yet an even simpler synonym exists: "get," which ranks just 39th.[49] This raises several interesting questions. Does opting for an extremely common word rather than a very common word make any difference to the target audience? If so, approximately what percentage of readers would benefit from the swap? Is the author be doing a disservice to readers in limiting the richness of a language? In any case, this obstacle highlighted the true need for including readers with disabilities in the process and showed that E2R authors would do well with more concrete guidance.

Consistency is another limitation of using a manual system to flag E2R violations. Once a decision has been made regarding the use of "receive" versus "get," for instance, we observed that it was difficult to ensure that each and every instance was evaluated in the same way, especially since the task was too long to complete in one sitting. We attempted to limit inconsistencies by using the search feature in Excel after the initial round of annotations were performed to verify that segments with the same issues were treated the same way, however this is an ungraceful solution to a problem which could be more easily resolved with an automatic text-checking system like LIREC.

## 5.4 Readability

In addition to the three human evaluations described in Sections 5.2 and 5.3, we also conducted an automatic evaluation of our English translations of FALC administrative documents obtained using DeepL NMT. This section addresses the third and final feature of linguistic accessibility that we have identified and measured in this thesis: **readability**.

### 5.4.1 Step 4: Automatic readability evaluation

We formed the following hypothesis to test using the Coh-Metrix web tool and the methodology described in *Section 4.3.2.1*:

> **H3.0**: *Removing forced line breaks before NMT will* ***improve readability****.*

[48] https://www.wordandphrase.info/analyzeText.asp Last accessed: December 5, 2020
[49] Ibid.

Because of the ethical and practical reasons laid out in our methodology, only the E2R violation-free subsets of the pre-edited (288 of 849) and non-pre-edited (212 of 849) translations were evaluated for readability. A more direct comparison, and therefore perhaps a more accurate automatic evaluation, could have been obtained by including the exact same sentences from the non-pre-edited and pre-edited data sets in the Coh-Metrix analysis. However, this decision was made in order to maximize the amount of data available for analysis. We will now address a selection of results of the automatic Coh-Metrix analyses, presented in *Table 5.20*.

First, we studied indicators that deal with the shallow measures of readability discussed in *Section 3.3.1*: **sentence length, word length, Flesh Reading Ease, and Flesh-Kincaid Grade Level**. As we established in *Section 5.3.1*, the Inclusion Europe guidelines do not give a precise number to aim for in terms of sentence length, only that authors should, "Always keep your sentences short" (2009, p. 11). However, as Nietzio et al. (2012) point out, this ambiguity means that sentence length can be difficult for authors and evaluators of E2R to judge. Their analysis of roughly 3,000 German Easy-to-Read sentences using the grammar checker LanguageTool found an average sentence length of eight words, most of which did not exceed 13 (Nietzio et al., 2012). Yaneva's (2015) study of easy-read English (not necessarily E2R) on the web discovered an even lower average, just 6.3 words per sentence ($SD$ = 2.17). Compared to these two studies, our translated segments were much longer, at 14.42 ($SD$ = 4.065) words/sentence for our non-pre-edited segments a very slight improvement after pre-editing, of 14.177 ($SD$ = 4.473) words/sentence. Although not ideal, these results are below Dubay's (2004) estimated average length of around 20 words per sentence for standard English texts and meet the recommendations of many modern readability guides for internet writing which advise no more than 20-25 words per sentence.

| Indicator | Non-Pre-Edited | Pre-Edited |
|---|---|---|
| Average sentence length (words) | 14.42 | **14.177** |
| Average word length (syllables) | **1.478** | 1.479 |
| Agentless passive voice density | **2.29** | 2.939 |
| Negation density | **6.542** | 9.307 |
| 2nd-person pronoun incidence | **85.378** | 83.272 |
| Average Log word frequency | **3.205** | 3.19 |
| Average age of acquisition | 333.39 | **329.126** |
| Average familiarity | 583.75 | **584.179** |
| Flesch Reading Ease | 67.16 | **67.322** |
| Flesch-Kincaid Grade Level | 7.474 | **7.391** |

*Table 5.22: A summary of results from the Coh-Metrix 3.0 analysis of the segment of our translated English corpus featuring no Easy-to-Read violations (the better of the two scores for each category is highlighted in bold text)*

The word length for both of our test sets was almost identical (non-pre-edited $M$ = 1.478, $SD$ = 0.961; pre-edited $M$ = 1.479, $SD$ = 0.957) and was on par with Yaneva's (2015) findings ($M$ = 1.44, $SD$ = 0.12), however with greater variance. The traditional formulas of readability that are calculated by Coh-Metrix are the Flesh Reading Ease and the Flesh-Kincaid Grade Level measurements. Since these are based on the previous two indicators we discussed, it comes as no surprise that there is little difference between the non-pre-edited (67.16; 7.474) and pre-edited scores (67.322; 7.391). We are also able to see the direct impact that the sentence length had on these measurements when they are compared to the scores Yaneva (2015) obtained of 78.84 ($SD$ = 10.9) and 3.83 ($SD$ = 1.75). Štajner et al. (2012) set an even higher Reading Ease score of 90 as the minimum for the FIRST language technology developed for the conversion of texts intended for people with autism spectrum disorders (ASD). Although there is no official score for E2R, Flesch (1979) sets the minimum threshold for Plain English (PE) at 60 his scale of 0-100, and we know that E2R was designed to be even easier than PE. Recall that WCAG 2.1 guideline 3.1.5 – Reading Level advises providing easy-language version of any text that requires reading knowledge that surpasses the lower secondary level. In the United States, the system the Flesch scores are based on, lower secondary education corresponds to 7th through 9th grade, or, once again, no less

than 60 on Flesch's Reading Ease scale. Therefore, though our scores are lower than the other published research, they do fall within, albeit at the high end of, the acceptable range.

We also studied indicators that correspond to the Word Level category of E2R guidelines: **word frequency, age of acquisition, and familiarity**. Again, there was little difference between the non-pre-edited and pre-edited scores. The largest difference that was observed was in the average Age of Acquisition (AOA) indicator, for which the pre-edited data scored approximately 4 points higher; however, relative to the AOA scale of 0-700, this difference is minor. When compared to the data from Yaneva (2015), our texts received somewhat better scores for word frequency (2.494 for non-pre-edited and 2.509 for pre-edited versus 2.43) and familiarity (583.75 for non-pre-edited and 584.179 for pre-edited versus 580.8), but somewhat worse scores for age of acquisition (333.39 for non-pre-edited and 329.126 for pre-edited versus 317.4; lower scores indicate that a word is more likely to be learned earlier on in life). Then again, Fajardo et al. (2014) found no significant correlation between word frequency (or length, for that matter) and comprehension in young adults with intellectual disabilities during their reception study of Easy-to-Read (IFLA guidelines), so these indicators may not be the best predictors of whether a text is easy to read and understand by the primary target audience or not.

Finally, we examined the Coh-Metrix data for three more indicators that we can associate with the Sentence Level E2R guidelines: **agentless passive voice density** ("Use active language rather than passive language where possible"), **negation density** ("Use positive sentences rather than negative ones where possible"), and **second-person pronoun incidence** ("Speak to people directly. Use words like "you" to do this") (Inclusion Europe, 2009, p. 11). The non-pre-edited test set outperformed the pre-edited test set in all three categories, but generally by a small margin. Some of the most surprising results were the scores obtained for negation density, which counts occurrences of expressions such as "no" and "not," and which are to be avoided in E2R writing. The pre-edited data featured a negation density of 9.307, in contrast to a score of 6.542 for the non-pre-edited data. As previously mentioned, a more direct sentence-by-sentence comparison would be needed to confirm these findings, since we observed that most instances of negative sentences in the target texts were already present in the source text. We hypothesize that a human translator would make the switch from a negative construction to a positive one more easily than a neural system since it requires complex semantic analysis.

#### 5.4.1.1 Discussion and limitations of Step 4

Although the pre-edited test set scored very slightly better in the comprehensive Flesch readability indices of the automatic evaluation, **these results do not provide conclusive data to support our hypothesis** that removing line breaks during pre-editing improves readability, at least insofar as it is defined by purely quantitative measurements. In fact, of the 10 indicators examined, the non-pre-edited test set obtained better results exactly half of the time, notably for all three indicators related to sentence construction (negative vs. positive constructions, active vs. passive voice, and direct speech through the use of second-person pronouns).

These scores help us paint a picture of the overall level of the translations resulting from generic neural machine translation before and after line breaks were removed, but they do not allow us to capture the entire readability landscape. There were several limiting factors to this step of our study. The smaller sample size of data in the non-pre-edited set (3057 words and 212 sentences compared to 4058 words and 288 sentences in the pre-edited set), due to the omission of non-E2R-compliant segments, could have made that data less reliable. Additionally, due to the limited amount of published research on Easy-to-Read, there was little empirical evidence to compare our data to, besides the work we have cited (Nietzio et al., 2012; Yaneva, 2015). Ideally, these measurements would be compared to the scores for the French source texts, which were published with the FALC logo and therefore are considered reliably accessible, as discussed in *Section 4.2.2*. However, presently, the Coh-Metrix tool does not support the French language, and there is no other comparable tool that provides the same measurements. It would also be interesting to compare the results to an English-language "gold standard," human translated and verified by members of the target readership, though none currently exist for the texts studied.

## 5.5 Summary of the results

In this chapter, we examined the linguistic accessibility of English translations of FALC administrative documents from three perspectives over the course of four empirical steps, in order to determine whether a specific pre-editing process could improve the prospects of generic NMT technology to handle this type of controlled language. In each step, we compared results from translations that featured line breaks in accordance with Inclusion Europe guidelines and those from translations in which line breaks were removed.

*Steps 1* and *2* aimed to answer our first Research Question (**RQ1**), which *Step 1*, a human evaluation of translation quality via error annotation produced findings that **support hypotheses H1.1 and H1.2**; translation quality was improved when manual line breaks were removed, both in terms of

number of errors and severity. Hypothesis **H1.3 was partially supported** by the outcome of this step; fluency was the error category most positively impacted by pre-editing, but style was the category least positively impacted. In *Step 2*, a group of graduate-level translation students evaluated post-editing effort. Results from this step **partially support hypothesis H1.4**; editing effort decreased significantly after pre-editing, and editing time was also reduced, though the results were not statistically significant.

*Step 3*, designed to answer our second Research Question (**RQ2**), involved human annotation of the same pre-edited and non-pre-edited translations, this time in search of violations of the aforementioned Easy-to-Read guidelines. Our results, a significant decrease in violations for the segments that had been edited to remove line breaks prior to translation, **support hypothesis H2.1** (and therefore the overarching H2.0), demonstrating that text accessibility is improved when this particular pre-editing step is performed.

Finally, in *Step 4* of our experiment, an automatic assessment of readability using the Coh-Metrix 3.0 web tool intended to answer our third Research Question (**RQ3**). The split data obtained in this step **does not support hypothesis H3.0**.

Overall, it would appear that pre-editing to remove the manual line breaks stipulated by Easy-to-Read guidelines prior to neural machine translation with DeepL improves translation quality and accessibility (E2R compliance), but not readability. Although some factors of linguistic accessibility were improved as a result of this pre-editing step, as evidenced by our findings from *Steps 1* and *3*, the raw translations obtained were still far from publishable, and the time and effort saved through the use of free machine translation may be offset by high post-editing costs, as shown in *Step 2*.

# Chapter 6: Conclusions

This chapter opens (*Section 6.1*) with an overview of the research conducted for this thesis and a summary of the key takeaways from each step of the experiment. It concludes (*Section 6.2*) with a brief discussion of the study's drawbacks and of its successes, both of which provide ample opportunity for further investigations into the topic of barrier-free communication.

## 6.1 Summary of the findings

Up to this point, the literature combining controlled language (CL), machine learning and translation, and barrier-free communication is quite limited. Several studies in recent years have combined two of the three areas of research – in particular the relationship between CL and neural machine translation (NMT) – exploring issues such as the neural machine translatability of Cochrane Plain Language Summaries (Rossetti, 2019) or that of controlled technical communication (Marzouk and Hansen-Schirra, 2019). A few projects and areas of research have even touched on all three. Automatic text simplification using machine learning, such as the work by T. Wang et al. (2016) and Chen et al. (2017), comes to mind, as does the two-part SIMPLES project, whose aim is to produce an automatic text summarization tool and an authoring tool specifically for the French version of Easy-to-Read (*Français facile à lire et à comprendre, FALC*) (Jacquet & Poitrenaud, 2019; Chehab et al., 2019). Nonetheless, the exploratory study that inspired the present thesis is the only other work to our knowledge that has dealt specifically with Easy-to-Read (E2R) language and interlingual neural machine translation with the goal of linguistic accessibility.

This study aimed to investigate the degree of linguistic accessibility of controlled language generated via NMT. The experiments carried out over the course of two distinct phases were designed to accomplish the research goal of determining whether a free, generic NMT system like DeepL could be a suitable method of producing Easy-to-Read English translations of French source texts.

In *Phase 1* (*Section 4.3.1*) of the study, linguistic accessibility was assessed through human evaluation by means of three different methodological perspectives in three corresponding research steps. *Step 1* (*Section 4.3.1.1.1*) the researcher completed a DQF-MQM error annotation as one way to assess the translation quality variable. The quantitative data collected in *Step 1* of our study suggests that the quality of DeepL neural machine translation is significantly higher when a pre-editing step, which removes the manually inserted line breaks that are characteristic of E2R, is performed. As

expected, quality was significantly improved both in terms of the number and the gravity of errors present, validating two of our three hypotheses (**H1.1** and **H1.2**) about our first research question (**RQ1**). Perhaps the most encouraging finding of this step is the total number of translation errors, which decreased by nearly 32% as a result of pre-editing. Unexpected results regarding the categories of errors that would experience the most positive change due to pre-editing do not allow us to validate our third hypothesis (**H1.3**) for **RQ1**. In *Step 2* (*Section 4.3.1.1.2*) of the study, translation quality was evaluated from another angle, by way of a post-editing experiment with six volunteer participants. Participants edited 25 different translations twice for each source sentence, once for the version that had been pre-edited and once for the version that had not. We predicted that the post-editing effort necessary to achieve a publishable translation would be lessened when pre-editing was performed prior to translation (**H1.4**). Findings were split for this hypothesis; participants did expend less post-editing effort when it came to the number of changes they made (Post-Editing Effort or PEE), but the difference in time they spent editing (Time-to-Edit or TTE) between the two sets of translations was not statistically significant. Our second research question (**RQ2**) focuses on accessibility, which, as the reader will recall, was based on adherence to E2R guidelines. As such, *Step 3* (*Section 4.3.1.2.1*) utilized an E2R violation annotation to evaluate this component. Roughly a 24% decrease in violations was observed, allowing us to validate hypothesis **H2.1**, which predicted that the level of accessibility would improve as a result of pre-editing.

*Phase 2* (*Section 4.3.2*) of the study concerned the readability component of linguistic accessibility. It was designed based on an automatic evaluation using the Coh-Metrix 3.0 text analysis tool in order to answer our third research question (**RQ3**). In the fourth and final step of the study, *Step 4* (*Section 4.3.2.1*), we examined a selection of ten readability indicators provided by Coh-Metrix for translations with and without pre-editing. While we had predicted that readability would improve with pre-editing (**H3.0**), results for this step were inconclusive. In fact, the non-pre-edited translations obtained better scores than their pre-edited counterparts for exactly half of the indicators, including all of the indicators related to sentence structure.

## 6.2 Next steps

With this study, we aimed to contribute to the body of knowledge surrounding alternative methods of producing accessible multilingual communication. Through our experimentation, we were indeed able to answer the research questions posed at the outset of our investigation. Manually formatting texts to introduce line breaks at natural pauses in speech helps make information more accessible to a target audience of people with intellectual disabilities or other conditions that affect

reading skills. However, it is clear that this authoring recommendation was developed with humans in mind, not machines. We discovered that this particular formatting standard poses a major problem to the DeepL system's handling of French Easy-to-Read administrative texts, but that pre-editing the source texts to address the issue has a positive impact on the subsequent translation quality and accessibility (though not necessarily the readability) of English target texts. Our research helped fill some noticeable gaps in the literature regarding the weaknesses of neural machine translation in dealing with controlled language such as E2R, and how it can be improved upon as a tool for producing linguistically accessible text. In doing so, it also shed light on questions that still need answering as governments and other societal actors work toward what Felici and Griebel (2019) believe must be one of their primary goals: speaking the language of their citizens.

First, different language combinations and text types should be considered in future work on the machine translation of E2R. This was a notable limitation of our study; the specificities of the natural languages and the text genre selected do not allow us to generalize our results. Since this research was performed, the company has rolled out a glossary feature and a regional English dialect option, in addition to several new language pairs and updated neural networks. Due to the constant and rapid evolution of this type of technology, it is likely that the quality of neural machine translation will continue to improve, perhaps even to the point of learning how to correctly handle the sentence boundary disambiguation that caused problems in the texts in our study. From a technological perspective, sentence boundary disambiguation could be explored further, since it has implications in other types of translation, such as subtitles. Until then, however, other E2R guidelines that might prove problematic could be investigated and solutions could be tested in the pre-editing process, as we have done with the line break guideline.

A second substantial limitation of this research is the absence of collaboration with members of the disability community. Christiane Maaß (2019) describes the need for accessible communication research using the analogy of a pair of beach sandals. The left flip-flop in her analogy represents text studies, such as the one we have spent the last five chapters describing. But what is a flip-flop without its mate? (Simply a "flip"…) A pair of sandals without both feet is far less helpful than a complete set, just as a text study without a corresponding reception study gives us only part of the insight we need to draw useful conclusions. Moreover, the Inclusion Europe guidelines on E2R themselves mandate the inclusion of members of the target readership in all aspects of text production. Although the study carried out for this thesis thoroughly evaluated E2R translations from four unique perspectives, we can only make presumptions about how the texts would actually

be received. An important follow-up study, ideally involving English-speaking readers with intellectual disabilities, would be necessary to validate these initial findings.

Finally, although aspects of linguistic accessibility were markedly improved thanks to the pre-editing step that was introduced, it would be imprudent to overlook the raw data; the truth is that even the pre-edited translations we obtained were far from usable in terms of quality and accessibility. This realization underscores the real possibility that generic NMT is not the most practical solution to the obstacles that are currently preventing multilingual E2R from becoming widespread. Two avenues of research related to this eventuality immediately come to mind. First, an analysis of whether the use of neural machine translation actually makes the process of E2R text production faster, easier, and more cost-effective than human translation from scratch, after the cost of pre- and post-editing services (ideally by native English-speaking E2R experts) are factored in. And second, further empirical exploration into E2R authoring tools such as LIREC and FALC Assistant as promising alternatives or complements to the partial solution we have presented in this research.

Much work remains to be done to ensure that written communication implemented in societies optimally serves *all* citizens, including those who have particular challenges due to language skills, reading skills, or both. However, with this research we hope to have shown that barrier-free communication via multilingual Easy-to-Read language is not only a worthwhile objective to strive for, but also that there exist creative technological solutions to achieving it.

# Works Cited


Aikawa, T., Schwartz, L., King, R., … Lozano, C. (2007). Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. *Proceedings of Machine Translation Summit XI*, 1–7. http://research.microsoft.com/apps/pubs/default.aspx?id=69483

Aizpurua, A., Harper, S., & Vigo, M. (2016). Exploring the relationship between web accessibility and user experience. *International Journal of Human-Computer Studies*, *91*, 13–23.

Aluísio, S. M., & Gasperin, C. (2010). Fostering Digital Inclusion and Accessibility: The PorSimples Project for Simplification of Portuguese Texts. *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, 46–53. http://dl.acm.org/citation.cfm?id=1868701.1868708

Arnold, D., Balkan, L., Meijer, S., … Sadler, L. (1994). *Machine translation: An Introductory Guide*. NCC Blackwell.

Association Aires Paris. (2017, June). Accès à l'information. *AIRES*, *3*.

Bahdanau, D., Cho, K., & Bengio, Y. (2015, May 7). *Neural Machine Translation by Jointly Learning to Align and Translate*. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA. arXiv:1409.0473.

Ballard, M. (1992). *De Cicéron à Benjamin: Traducteurs, traductions, réflexions*. Presses universitaires de Lille.

Barthe, K., Juaneda, C., Leseigneur, D., … Vayrette, A. (1999). GIFAS Rationalized French: A Controlled Language for Aerospace Documentation in French. *Technical Communication*, *46*(2), 220–229.

Bawa Mason, S. (2019). Joss Moorkens, Sheila Castilho, Federico Gaspari, Stephen Doherty (eds): Translation quality assessment: From principles to practice. *Machine Translation*, *33*(3), 269–277. https://doi.org/10.1007/s10590-019-09241-w

*BBC Subtitle Guidelines v. 1.1.8*. (2019). BBC. https://bbc.github.io/subtitle-guidelines/

Becker, L. (2019). 'Immigrants' as recipients of Easy-to-Read in Spain. *Journal of Multilingual and Multicultural Development*, 1–13. https://doi.org/10.1080/01434632.2019.1621874

Bernardini, S., & Kenny, D. (2019). Corpora. In M. Baker & G. Saldanha (Eds.), *Routledge Encyclopedia of Translation Studies* (3rd ed., pp. 110–115). Routledge.

Bojar, O., Chatterjee, R., Federmann, C., … Zampieri, M. (2016). Findings of the 2016 Conference on Machine Translation. *Proceedings of the First Conference on Machine Translation: Shared Task Papers*, *2*, 131–198. https://doi.org/10.18653/v1/W16-2301

Bott, S., & Saggion, H. (2014). Text simplification resources for Spanish. *Language Resources and Evaluation*, *48*(1), 93–120. https://doi.org/10.1007/s10579-014-9265-4

Bouillon, P. (2017, November 23). *Traduction Automatique I - Systèmes de TA Statistiques* [Course].

CARDET. (2014). *Adults with a Learning Disability – Observatory of Best Practices, D12: Resource Database*. Lifelong Learning Programme - European Commission.

Cardone, D. (1999). Exploring the Use of Question Methods: Pictures Do Not Always Help People with Learning Disabilities. *The British Journal of Development Disabilities*, *45*(89), 93–98. https://doi.org/10.1179/096979599799155894

Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). Approaches to Human and Machine Translation Quality Assessment. In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), *Translation Quality*

*Assessment* (Vol. 1, pp. 9–38). Springer International Publishing. https://doi.org/10.1007/978-3-319-91241-7_2

Castilho, S., Moorkens, J., Gaspari, F., … Way, A. (2017). Is Neural Machine Translation the New State of the Art? *The Prague Bulletin of Mathematical Linguistics*, *108*(1). https://content.sciendo.com/view/journals/pralin/108/1/article-p109.xml

CHANGE. (2016). *How to Make Information Accessible: A guide to producing easy read documents*. https://www.changepeople.org/Change/media/Change-Media-Library/Free%20Resources/How-to-make-info-accessible-guide-2016-Final.pdf

Chehab, N., Holken, H., & Malgrange, M. (2019). *Étude Recueil des besoins FALC*. Holken Consultants & Partners.

Chen, P., Rochford, J., Kennedy, D. N., … Scott, W. (2017). Automatic Text Simplification for People with Intellectual Disabilities. In H. Yang (Ed.), *Artificial Intelligence Science and Technology* (pp. 725–731). World Scientific. https://www.worldscientific.com/doi/pdf/10.1142/9789813206823_0091

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). *On the properties of neural machine translation: Encoder-decoder approaches*. Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8).

Cortier, C. (2006). De quelques enjeux et usages historiques du Français fondamental. *Documents pour l'histoire du français langue étrangère ou seconde*, *36*, 9–12.

COSLA. (2008). *Guide pratique de la rédaction administrative*. Comité d'Orientation pour la Simplification du Langage Administratif (COSLA). Ministère de la fonction publique et de la réforme de l'Etat, République Française.

Coster, W., & Kauchak, D. (2011). Simple English Wikipedia: A New Text Simplification Task. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 665–669. https://www.aclweb.org/anthology/P11-2117

Crego, J., Kim, J., Klein, G., … Zoldan, P. (2016). *SYSTRAN's Pure Neural Machine Translation Systems*. arXiv:1610.05540.

Crossley, S. A., McCarthy, P. M., Dufty, D. F., & McNamara, D. S. (2007). Toward a New Readability: A Mixed Model Approach. *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, 6.

Dale, E., & Chall, J. S. (1948). A Formula for Predicting Readability. *Educational Research Bulletin*, *27*(1), 11–28. JSTOR.

Davison, A., & Kantor, R. N. (1982). On the Failure of Readability Formulas to Define Readable Texts: A Case Study from Adaptations. *Reading Research Quarterly*, *17*(2), 187–209. JSTOR. https://doi.org/10.2307/747483

de Landsheere, G. (1963). Pour une application des tests de lisibilité de Flesch à la langue française. *Le Travail Humain*, *26*(1/2), 141–154.

Doherty, S. (2012). *Investigating the effects of controlled language on the reading and comprehension of machine translated texts: A mixed-methods approach* [Dublin City University (DCU)]. http://doras.dcu.ie/16805/

Drndarevic, B., Štajner, S., & Saggion, H. (2012). Reporting Simply: A Lexical Simplification Strategy for Enhancing Text Accessibility. *Proceedings of the W3C Easy-to-Read on the Web Syposium*. http://www.w3.org/WAI/RD/2012/easy-to-read/paper7/

Dubay, W. H. (2004). The Principles of Readability. *Costa Mesa, CA: Impact Information*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.91.4042

European Commission. (2017). *Progress Report on the implementation of the European Disability Strategy (2010—2020)*. Brussels, Belgium.

Fajardo, I., Ávila, V., Ferrer, A., … Hernández, A. (2014). Easy-to-read Texts for Students with Intellectual Disability: Linguistic Factors Affecting Comprehension. *Journal of Applied Research in Intellectual Disabilities*, *27*(3), 212–225. https://doi.org/10.1111/jar.12065

Fajardo, I., Tavares, G., Ávila, V., & Ferrer, A. (2013). Towards text simplification for poor readers with intellectual disability: When do connectives enhance text cohesion? *Research in Developmental Disabilities*, *34*(4), 1267–1279. https://doi.org/10.1016/j.ridd.2013.01.006

Federico, M., Bertoldi, N., Negri, M., … Germann, U. (2014). The MateCat Tool. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, 129–132.

Felici, A., & Griebel, C. (2019). The challenge of multilingual 'plain language' in translation-mediated Swiss administrative communication: A preliminary comparative analysis of insurance leaflets. *Translation Spaces*, *8*(1), 167–191. https://doi.org/10.1075/ts.00017.fel

Feng, L. (2009). Automatic readability assessment for people with intellectual disabilities. *ACM SIGACCESS Accessibility and Computing*, *93*, 84–91. https://doi.org/10.1145/1531930.1531940

Feng, L., Elhadad, N., & Huenerfauth, M. (2009). Cognitively motivated features for readability assessment. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on - EACL '09*, 229–237. https://doi.org/10.3115/1609067.1609092

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. https://doi.org/10.1037/h0057532

Flesch, R. (1979). *How to Write Plain English: A Book for Lawyers and Consumers*. Harper & Row.

Forcada, M. L. (2017). Making sense of neural machine translation. *Translation Spaces*, *6*(2), 291–309. https://doi.org/10.1075/ts.6.2.06for

FRANET. (2016). *Monthly data collection on the current migration situation in the EU, Thematic focus: Disability*. European Union Agency for Fundamental Rights (FRA). https://fra.europa.eu/sites/default/files/fra_uploads/fra-august-2016-monthly-migration-disability-focus_en.pdf

Freyhoff, G., Hess, G., Kerr, L., … Van Der Veken, K. (1998). *Make it Simple. European Guidelines for the Production of Easy-to-Read Information for People with Learning Disability for authors, editors, information providers, translators and other interested persons*.

Gerlach, J. (2015). *Improving Statistical Machine Translation of Informal Language: A Rule-based Pre-editing Approach for French Forums*. University of Geneva.

Girletti, S., Bouillon, P., Bellodi, M., & Ursprung, P. (2019). Preferences of end-users for raw and post-edited NMT in a business environment. *Proceedings of the 41st Conference Translating and the Computer*, 47–59.

Glavaš, G., & Štajner, S. (2015). Simplifying Lexical Simplification: Do We Need Simplified Corpora? *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 63–68. https://doi.org/10.3115/v1/P15-2011

Goddard, C. (Ed.). (2018). *Minimal English for a Global World: Improved Communication Using Fewer Words*. Palgrave Macmillan.

Gowers, E. (2014). *Plain Words* (R. Gowers, Ed.; Revised). Penguin UK.

Grabar, N., & Cardon, R. (2018). CLEAR – Simple Corpus for Medical French. *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, 3–9. https://doi.org/10.18653/v1/W18-7002

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202. https://doi.org/10.3758/BF03195564

Grudniewicz, A., Bhattacharyya, O., McKibbon, K. A., & Straus, S. E. (2015). Redesigning printed educational materials for primary care physicians: Design improvements increase usability. *Implementation Science*, *10*(1), 156. https://doi.org/10.1186/s13012-015-0339-5

Gunning, R. (1952). *The Technique of Clear Writing*. McGraw-Hill; /z-wcorg/.

Harper, R., & Zimmerman, D. (2009). Exploring plain language guidelines. *2009 IEEE International Professional Communication Conference*, 1–6. https://doi.org/10.1109/IPCC.2009.5208669

Hassell, J. (2018, August 23). The importance of text accessibility: How IBM's Content Clarifier shows us what we've forgotten. *Hassell Inclusion*. https://www.hassellinclusion.com/blog/importance-text-accessibility/

Hearne, M., & Way, A. (2011). Statistical Machine Translation: A Guide for Linguists and Translators: SMT for Linguists and Translators. *Language and Linguistics Compass*, *5*(5), 205–226. https://doi.org/10.1111/j.1749-818X.2011.00274.x

Hennequin, L. (2019, May 12). Européennes: Les personnes handicapées mentales pourront voter grâce à ce langage simplifié. *HuffPost*. https://www.huffingtonpost.fr/entry/les-handicapes-mentaux-pourront-voter-aux-europeennes-grace-a-ce-langage-simplifie_fr_5cd58c16e4b0796a95dab0a9

Huijsen, W.-O. (1998). Controlled Language—An Introduction. In T. Mitamura (Ed.), *Proceedings of the Second International Workshop on Controlled Language Applications – CLAW 98* (pp. 1–15). Language Technologies Institute, Carnegie Mellon University.

Hurtado, B., Jones, L., & Burniston, F. (2014). Is Easy Read information really easier to read? *Journal of Intellectual Disability Research*, *58*(9), 822–829. https://doi.org/10.1111/jir.12097

Hutchins, J. (2005). Machine Translation: General Overview. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (1st ed., pp. 501–511). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199276349.001.0001

Hutchins, W. J., & Somers, H. L. (1992). *An introduction to machine translation*. Academic Press.

Ignacio Madrid, R., Ávila, V., Fajardo, I., & Ferrer, A. (2012). Writing Easy-to-Read Documents for People With Intellectual Disabilities. In M. Torrance, D. Alamargot, M. Castelló, … L. van Waes (Eds.), *Learning to Write Effectively: Current Trends in European Research*. BRILL. https://doi.org/10.1163/9781780529295_076

Inclusion Europe. (2009). *Information for all. European standards for making information easy to read and understand*. European Commission. http://www.inclusion-europe.org/etr/en/european-easy-to-read-standards

Institut National de la Statistique et des Études Économiques (INSEE). (2020). *Étrangers – Immigrés: Pays de naissance et nationalités détaillés*. INSEE. https://www.insee.fr/fr/statistiques/4510522?sommaire=4510556

ISO/IEC 40500:2012 Information Technology – W3C Web Content Accessibility Guidelines (WCAG) 2.0, (2012). https://www.iso.org/standard/58625.html

Jacquet, B., & Poitrenaud, S. (2019, December 11). *Présentation LIREC: Plateforme d'aide à la mise en FALC pour les ESAT et foyers de vie*. Workshop projet SIMPLES, Paris, France.

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.

Kaplan, A., Rodríguez Vázquez, S., & Bouillon, P. (2019). Measuring the Impact of Neural Machine Translation on Easy-to-Read Texts: An Exploratory Study. *Conference on Easy-to-Read Language Research (Klaara 2019)*. https://archive-ouverte.unige.ch/unige:123648

Karreman, J., van der Geest, T., & Buursink, E. (2007). Accessible Website Content Guidelines for Users with Intellectual Disabilities. *Journal of Applied Research in Intellectual Disabilities*, *20*(6), 510–518. https://doi.org/10.1111/j.1468-3148.2006.00353.x

Kirkpatrick, A., O Connor, J., Campbell, A., & Cooper, M. (Eds.). (2018). *Web Content Accessibility Guidelines (WCAG) 2.1*. W3C Recommendation. https://www.w3.org/TR/2018/REC-WCAG21-20180605/

Kittredge, R. I. (2003). Sublanguages and controlled languages. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (pp. 430–447).

Klaper, D., Ebling, S., & Volk, M. (2013). Building a German/Simple German Parallel Corpus for Automatic Text Simplification. *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, 19. https://doi.org/10.5167/uzh-78610

Klare, G. R. (1976). A Second Look at the Validity of Readability Formulas. *Journal of Reading Behavior*, *8*(2), 129–152. https://doi.org/10.1080/10862967609547171

Koehn, P. (2017). Draft of Chapter 13: Neural Machine Translation. In *Statistical Machine Translation* (2nd ed.). arXiv preprint arXiv:1709.07809.

Krings, H. P. (2001). *Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes* (G. Koby, Ed.; S. Litzer, G. Shreve, & K. Mischerikow, Trans.; Vol. 5). Kent State University Press.

Kuhn, T. (2014). A Survey and Classification of Controlled Natural Languages. *Computational Linguistics*, *40*(1), 121–170. https://doi.org/10.1162/COLI_a_00168

Lamotte, H., & Therwath, A. (2016). Orsay facile: Inclure les personnes déficientes intellectuelles dans l'élaboration de documents adaptés. In E. Nardi & C. Angelini (Eds.), *Best Practice 5: A tool to improve museum education internationally* (pp. 61–70). Edizioni Nuova Cultura.

Language, n. (2008). In *OED Online* (Third Edition). Oxford University Press. www.oed.com/view/Entry/105582

Laviosa, S. (2010). Corpora. In *Handbook of Translation Studies* (Vol. 1, pp. 80–86). John Benjamins.

Lazar, J. (2017). *Research Methods in Human-Computer Interaction.* (2nd ed.).

Lommel, A. (2018). Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies. In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), *Translation Quality Assessment: From Principles to Practice* (Issue volume 1). Springer; eBook Collection (EBSCOhost).

Lommel, A., Popovi, M., & Burchardt, A. (2014). Assessing Inter-Annotator Agreement for Translation Error Annotation. *LREC Workshop on Automatic and Manual Metrics for Operational Translation Evaluation*, 31–37.

López, J. S. (2006). Vocabulaires logiques, vocabulaires simplifiés et Français élémentaire. *Documents pour l'histoire du français langue étrangère ou seconde*, *36*, 97–118.

Luce, A. (2018). *Asylum Seekers and Refugees with Intellectual Disabilities in Europe*. Samuel Centre for Social Connectedness. https://www.socialconnectedness.org/wp-content/uploads/2019/10/Asylum-Seekers-and-Refugees-with-Intellectual-Disabilities-in-Europe-1-1.pdf

Maaß, C. (2019, September 19). Easy Language and beyond: How to maximize the accessibility of communication. *Invited Plenary Speech at the Klaara 2019 Conference on Easy-to-Read Language Research*. Klaara 2019 Conference on Easy-to-Read Language Research, Helsinki, Finland. https://hildok.bsz-bw.de/frontdoor/index/index/docId/996

Macketanz, V., Ai, R., Burchardt, A., & Uszkoreit, H. (2018). TQ-AutoTest – A Semi-Automatic Test Suite for (Machine) Translation Quality. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 886–892.

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, *18*(1), 50–60. http://dx.doi.org/10.1214/aoms/1177730491

Marzouk, S., & Hansen-Schirra, S. (2019). Evaluation of the impact of controlled language on neural machine translation compared to other MT architectures. *Machine Translation*, *33*(1–2), 179–203. https://doi.org/10.1007/s10590-019-09233-w

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press.

Mears, C., Rodgers, J., Townsley, R., … Thurman, S. (2004). *Information for All*. Norah Fry Research Centre, University of Bristol.

Mencap. (2000). *Am I making myself clear? Mencap's guidelines for accessible writing.* Mencap. http://www.accessibleinfo.co.uk/pdfs/Making-Myself-Clear.pdf

Microsoft Translator. (2016, November 15). Microsoft Translator launching Neural Network based translations for all its speech languages. *Microsoft Translator Blog*. https://www.microsoft.com/en-us/translator/blog/2016/11/15/microsoft-translator-launching-neural-network-based-translations-for-all-its-speech-languages/

Ministère de l'intérieur. (2018). *Élections des représentants au Parlement Européen du 26 mai 2019—Mémento à l'usage des candidats*. République Française. https://www.interieur.gouv.fr/Actualites/L-actu-du-Ministere/Document-a-l-attention-des-candidats-aux-elections-europeennes-2019

Mitamura, T. (1999). Controlled Language for Multilingual Machine Translation. *Proceedings of Machine Translation Summit VII*, 46–52.

Moorkens, J., Castilho, S., Gaspari, F., & Doherty, S. (2018). *Translation Quality Assessment: From Principles to Practice* (Issue volume 1). Springer; eBook Collection (EBSCOhost).

Neubig, G., Morishita, M., & Nakamura, S. (2015). *Neural Reranking Improves Subjective Quality of Machine Translation*. 35–41. https://aclweb.org/anthology/papers/W/W15/W15-5003/

NHS England. (2018). *Guide to making information accessible for people with a learning disability*. https://www.england.nhs.uk/ourwork/accessibleinfo/resources/

Nietzio, A., Naber, D., & Bühler, C. (2014). Towards Techniques for Easy-to-Read Web Content. *Procedia Computer Science*, *27*, 343–349. https://doi.org/10.1016/j.procs.2014.02.038

Nietzio, A., Scheer, B., & Bühler, C. (2012). How Long Is a Short Sentence? – A Linguistic Approach to Definition and Validation of Rules for Easy-to-Read Material. In K. Miesenberger, A. Karshmer, P. Penaz, & W. Zagler (Eds.), *Computers Helping People with Special Needs* (Vol. 7383, pp. 369–376). Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-31534-3_55

Nord, C. (2010). Functionalist approaches. In *Handbook of Translation Studies* (Vol. 1, pp. 120–128). John Benjamins.

Nyberg, E., Mitamura, T., & Huijsen, W.-O. (2003). Controlled language for authoring and translation. In H. Somers (Ed.), *Computers and Translation. A translator's guide.* (pp. 245–281). John Benjamins Publishing. https://benjamins.com/#catalog/books/btl.35.17nyb/details

O'Brien, S. (2010). Controlled language and readability. In G. M. Shreve & E. Angelone (Eds.), *Translation and Cognition* (pp. 143–165). John Benjamins Publishing Co.

O'Brien, S. (2004). Machine translatability and post-editing effort: How do they relate? *Translating and the Computer*, *26*.

Office fédérale de la statistique. (2020). *Langues principales en Suisse: Population résidante permanente de 15 ans et plus*. Confédération suisse. https://www.bfs.admin.ch/bfs/fr/home/statistiques/population/langues-religions/langues.assetdetail.11607335.html

Ogden, C. K. (1930). *Basic English: A general introduction with rules and grammar*. Paul Treber and Co., Ltd.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 311. https://doi.org/10.3115/1073083.1073135

People First New Zealand. (2017). *Make it Clear. A guide to making Easy Read information*. https://www.peoplefirst.org.nz/download/862/

Poncelas, A., & Murphy, G. (2007). Accessible Information for People with Intellectual Disabilities: Do Symbols Really Help? *Journal of Applied Research in Intellectual Disabilities*, *20*(5), 466–474. https://doi.org/10.1111/j.1468-3148.2006.00334.x

Quah, C. K. (2006). *Translation and technology*. New York: Palgrave Macmillan. http://site.ebrary.com/id/10487862

Redish, J. (2000). Readability Formulas Have Even More Limitations Than Klare Discusses. *ACM Journal of Computer Documentation (JCD)*, *24*(3), 132–137. https://doi.org/10.1145/344599.344637

Reiss, K. (2000). *Translation Criticism—The Potentials and Limitations: Categories and Criteria for Translation Quality Assessment* (E. F. Rhodes, Trans.). St. Jerome; New York: American Bible Society.

Reuther, U. (2003). Two in One – Can it work? Readability and Translatability by means of Controlled Language. *Proceedings of the Joint Conference Combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop*, 124–132.

Rodríguez Vázquez, S. (2013). Localizing accessibility of text alternatives for visual content in multilingual websites. *ACM SIGACCESS Accessibility and Computing Newsletter*, *105*, 34–37. https://doi.org/10.1145/2444800.2444807

Rodríguez Vázquez, S. (2016). *Assuring accessibility during web localisation: An empirical investigation on the achievement of appropriate text alternatives for images* [PhD]. University of Geneva.

Rossetti, A. (2019). *Simplifying, Reading, and Machine Translating Health Content: An Empirical Investigation of Usability*. [PhD] Dublin City University (DCU).

Roturier, J. (2006). *An investigation into the impact of controlled English rules on the comprehensibility, usefulness and acceptability of machine-translated technical documentation for French and German users* [Dublin City University (DCU)]. http://doras.dcu.ie/18190/

Saggion, H., Gómez Martínez, E., Etayo, E., … Bourg, L. (2011). Text Simplification in Simplext: Making Texts more Accessible. *Procesamiento Del Lenguaje Natural*, *47*, 341–342.

Savard, C., Québec (Province), Ministère du revenu, & Service des formulaires. (2003). *Pour qu'on vous lise--tout simplement: Techniques de rédaction en langue claire et simple*. Ministère du Revenu, Service des formulaires.

Schatz, T., Haberstroh, J., Bindel, K., … Knopf, M. (2017). Improving Comprehension in Written Medical Informed Consent Procedures. *GeroPsych*, *30*(3), 97–108. https://doi.org/10.1024/1662-9647/a000169

Schmutz, S., Sonderegger, A., & Sauer, J. (2019). Easy-to-read language in disability-friendly web sites: Effects on nondisabled users. *Applied Ergonomics*, *74*, 97–106. https://doi.org/10.1016/j.apergo.2018.08.013

Schwitter, R. (2002). English as a Formal Specification Language. *Proceedings of 13th International Conference on Database and Expert Systems Applications (DEXA 2002)*, 228–232. http://dx.doi.org/10.1109/DEXA.2002.1045903

Slocum, J. (1985). A Survey of Machine Translation: Its History, Current Status and Future Prospects. *Computational Linguistics*, *11*(1), 1–17.

Somers, H. (2005). Machine Translation: Latest Developments. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (1st ed., pp. 512–528). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199276349.001.0001

Štajner, S., Evans, R., Orăsan, C., & Mitkov, R. (2012). What Can Readability Measures Really Tell Us About Text Complexity? *Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) Workshop*, 14–21.

Štajner, S., Mitkov, R., & Corpas Pastor, G. (2015). Simple or Not Simple? A Readability Question. In N. Gala, R. Rapp, & G. Bel-Enguix (Eds.), *Language Production, Cognition, and the Lexicon* (Vol. 48, pp. 379–398). Springer International Publishing. https://doi.org/10.1007/978-3-319-08043-7_22

Stern, H. H. (1983). *Fundamental Concepts of Language Teaching*. Oxford University Press.

Sutherland, R. J., & Isherwood, T. (2016). The Evidence for Easy-Read for People With Intellectual Disabilities: A Systematic Literature Review: The Evidence for Easy-Read for People With Intellectual Disabilities. *Journal of Policy and Practice in Intellectual Disabilities*, *13*(4), 297–310. https://doi.org/10.1111/jppi.12201

Tang, G., Müller, M., Rios, A., & Sennrich, R. (2018). Why Self-Attention? A Targeted Evaluation of Neural Machine Translation Architectures. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4263–4272. https://doi.org/10.18653/v1/D18-1458

TAUS. (2016). *TAUS Quality Dashboard: From Quality Evaluation to Business Intelligence*. TAUS.

TAUS & CNGL. (2010). *MT Post-editing Guidelines*. https://www.taus.net/academy/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines

Trujillo, A. (1999). *Translation Engines: Techniques for Machine Translation*. Springer-Verlag London.

P028 – Directives de la Confédération pour l'aménagement de sites Internet facilement accessibles, 2.03 (2016). https://www.isb.admin.ch/isb/fr/home/ikt-vorgaben/prozesse-methoden/p028-richtlinien_bund_gestaltung_barrierefreie_internetangebote.html

*Convention on the Rights of Persons with Disabilities (CRPD)*, (2006) (testimony of United Nations). https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html

United Nations World Tourism Organization. (2020). *World Tourism Barometer*, *18*(5). https://www.e-unwto.org/doi/epdf/10.18111/wtobarometereng.2020.18.1.5

Vandeghinste, V., & Bulte, B. (2019). Linguistic Proxies of Readability: Comparing Easy-to-Read and regular newspaper Dutch. *Computational Linguistics in the Netherlands*, *9*, 81–100.

Volkart, L., Bouillon, P., & Girletti, S. (2018). Statistical vs. Neural Machine Translation: A Comparison of MTH and DeepL at Swiss Post's Language Service. In *Proceedings of the 40th Conference Translating and the Computer* (pp. 145–150).

Vollenwyder, B., Schneider, A., Krueger, E., … Mekler, E. D. (2018). How to Use Plain and Easy-to-Read Language for a Positive User Experience on Websites. In K. Miesenberger & G. Kouroupetroglou (Eds.), *Computers Helping People with Special Needs* (Vol. 10896, pp. 514–522). Springer International Publishing. https://doi.org/10.1007/978-3-319-94277-3_80

Wang, L., Tu, Z., Way, A., & Liu, Q. (2017). Exploiting Cross-Sentence Context for Neural Machine Translation. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2826–2831. https://doi.org/10.18653/v1/D17-1301

Wang, T., Chen, P., Rochford, J., & Qiang, J. (2016). Text Simplification Using Neural Machine Translation. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, 2.

Winkler, K., Kuhn, T., & Volk, M. (2014). Evaluating the Fully Automatic Multi-Language Translation of the Swiss Avalanche Bulletin. *Controlled Natural Language*, *8625*, 44–54. https://doi.org/10.1007/978-3-319-10223-8_5

World Health Organization & World Bank Group. (2011). *World Report on Disability*. World Health Organization Press. https://www.who.int/disabilities/world_report/2011/report.pdf

Wu, Y., Schuster, M., Chen, Z., … Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv*, *abs/1609.08144*. http://arxiv.org/abs/1609.08144

Wyner, A., Angelov, K., Barzdins, G., … Sowa, J. (2010). On Controlled Natural Languages: Properties and Prospects. In N. E. Fuchs (Ed.), *Controlled Natural Language* (pp. 281–290). Springer-Verlag.

Yaneva, V. (2015). Easy-read Documents as a Gold Standard for Evaluation of Text Simplification Output. *Proceedings of the Student Research Workshop Associated with RANLP 2015*, 30–36.

Yaneva, V., & Evans, R. (2015). Six Good Predictors of Autistic Text Comprehension. *Proceedings of the International Conference Recent Advances in Natural Language Processing*, 697–706. https://www.aclweb.org/anthology/R15-1089

Yaneva, V., Orasan, C., Evans, R., & Rohanian, O. (2017). Combining Multiple Corpora for Readability Assessment for People with Cognitive Disabilities. *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 121–132. https://doi.org/10.18653/v1/W17-5013

Yaneva, V., Temnikova, I., & Mitkov, R. (2016). Evaluating the Readability of Text Simplification Output for Readers with Cognitive Disabilities. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 7.

Zaretskaya, A., Vela, M., Pastor, G. C., & Seghiri, M. (2016). Measuring Post-editing Time and Effort for Different Types of Machine Translation Errors. *New Voices in Translation Studies*, *15*, 63–92. https://www.iatis.org/images/stories/publications/new-voices/Issue15-2016/Articles/Anna_Zaretskaya_FINAL.pdf

# Appendix A: Easy-to-Read guidelines used for annotation

General

Always use the right language for the people your information is for.

Make sure you explain the subject clearly and also explain any difficult words to do with that subject.

Word Level

Use easy to understand words that people will know well.

Do not use difficult words. If you need to use difficult words, make sure you always explain them clearly.

Use the same word to describe the same thing throughout your document throughout the document.

Do not use difficult ideas such as metaphors.

Do not use words from other languages unless they are very well known.

Avoid using initials. If you have to use initials, explain them.

Try not to use percentages and big numbers.

Be careful when you use pronouns. Make sure it is always clear who or what the pronoun is talking about.

Avoid all special characters where possible.

Do not use numbers with ordinal indicators or suffixes,

Write numbers as digits, not as words.

Avoid contractions.

Where possible, write dates out in full.

Sentence Level

Speak to people directly. Use words like 'you' to do this.

Use positive sentences rather than negative ones where possible.

Use active language rather than passive language where possible.

Keep the punctuation simple.

Where possible, use the present tense rather than the past tense.

Always keep your sentences short.

Structure

Start new sentences on a new line.

Never split 1 word over 2 lines.

Where possible, 1 sentence should fit on 1 line. If you have to write 1 sentence on 2 lines, cut the sentence where people would pause when reading out loud.

Use headings that are clear and easy to understand.

Use bullet points to list things.

## Appendix B: Post-editing study call for participation

*The following email was sent to all current students and recent graduates of the English section of UNIGE FTI Translation MA program, and a modified version was also sent to contacts at select US and UK universities featuring translation technology courses.*

Dear fellow students,

I hope you are all doing well and staying safe! I am reaching out because I am conducting a post-editing study for my master's thesis at the University of Geneva.

My project is about neural machine translation and text accessibility. I am looking for native English-speaking participants who have some previous experience with Machine Translation Post-Editing, which could include post-editing labs or exercises done within the framework of your MA studies. Those of you have taken Professor Bouillon's *Traduction Automatique 2* course will already be familiar with the CAT tool you will be using, but I will also provide detailed instructions, so the course is not a prerequisite.

Participation is voluntary and will take a maximum of 1 hour.

If you are interested in participating, please fill out the short consent form at the link below and I will contact you with more information.

Feel free to call, text, or email me if you have any questions that are not covered in the form. I look forward to hearing from many of you!

Best regards,
Abbe Kaplan
+33 07 77 77 45 60

# Appendix C: Post-editor background questionnaire and informed consent

*The following Google form included in the call for participation email; complete responses were used to reach out to participants individually to send instructions and assign post-editing tasks directly from MateCat.*

Thank you for your interest in the study and for taking the time to fill out this short form.

In the first section you will be asked some questions about yourself and your background. In the second section, you will be provided with an explanation of what participation in the study entails. In the third section you will be asked to confirm that you want to take part in it.

Email address * _____

Background Questionnaire

Are you a native speaker of English? *

Yes
No

If yes, what dialect?

US
UK
Canadian
Australian
Other: _____

Have you ever performed Machine Translation Post-Editing (MTPE) in any capacity? *

Yes
No
Don't know

If yes, please briefly describe your experience.
_____

Have you ever performed translation quality evaluation using error typology? *

Yes
No
Don't know

If yes, please briefly describe your experience.
_____

Which of the following best describes your education level? *

1st-year MA student
2nd-year MA student
3rd-year MA student
Recent graduate (MA degree obtained within the last 2 years)
Other: _____

What is your gender? *

Female
Male
Prefer not to say
Other: _____

Perceptions of MTPE

Since you have some prior experience with Machine Translation Post-Editing, please indicate how much you agree or disagree with the following statements.

I enjoy performing machine translation post-editing.

Strongly disagree 1 2 3 4 5 Strongly agree

I prefer post-editing machine translation to editing human translation.

Strongly disagree 1 2 3 4 5 Strongly agree

I feel comfortable post-editing to publishable (human-like) quality.

Strongly disagree 1 2 3 4 5 Strongly agree

I feel comfortable post-editing to "good enough" quality.

Strongly disagree 1 2 3 4 5 Strongly agree
Machine translation post-editing saves me time.

Strongly disagree 1 2 3 4 5 Strongly agree

I perform machine translation post-editing often.

Strongly disagree 1 2 3 4 5 Strongly agree

About the Study

The goal of this study is to compare the effects of pre-editing on post-editing effort for administrative documents developed for accessibility purposes. It is part of a research project for a master's thesis at the University of Geneva. Participants are invited to take part individually, using their personal computers, any time between May 22 and June 19, 2020.

You will be asked to perform post-editing on sentences that have been translated from French to English by a machine translation system, using MateCat, a free, web-based translation and revision tool.

The data collected will be treated confidentially and anonymously by the research team.

We anticipate no potential risks associated with the study. Participation in this research study is voluntary and you may withdraw your participation at any time without consequences.

Please confirm that you agree with the following statements.

I have read and understood the information provided about the study. *

Yes
No

I understand that participation involves Machine Translation Post-Editing. *

Yes
No

I understand that even if I agree to participate now, I can withdraw at any time or refuse to answer any question without consequences of any kind. *

Yes
No

I understand that I can withdraw permission to use data from my participation within one week of submitting it, in which case the material will be deleted. *

Yes
No

I understand that all information I provide for this study will be treated confidentially. *

Yes
No

I voluntarily agree to take part in this research study. *

Yes
No

To sign this form, please type your full name. *
_____

# Appendix D: Instructions for post-editing study participants

Thank you again for agreeing to contribute to my master's thesis research! Please read these guidelines carefully before beginning your post-editing tasks.

General Information and Instructions

- For this study, you will be working with MateCat, a free, web-based CAT (computer-assisted translation) tool to perform post-editing, a type of revision that is done on machine-translated text.

- You do not need to create a MateCat account or log into an existing account for this study.

- Please work on a desktop or laptop computer (MateCat supports Chrome and Safari browsers, so work with a recent version of one of these if possible). Do not use a tablet or mobile phone.

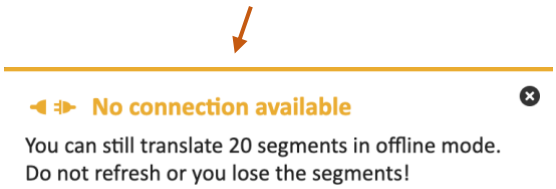
- There are two links in the email I sent you. Every link is unique, so do not share your links with other participants. Click the first link (labeled **First Task**) to work on the first task. Complete this task first, before proceeding to the second task in the email (labeled **Second Task**).

    - **Important**: Do not refer to the work you did in the first task while working on the second task.

- You may take as long as you need to complete the tasks provided that they are finished and submitted by June 19, 2020.

- Do your best to complete each task in one sitting. If you need to take a break for any reason, close the browser you are working in. You can come back to it at any time using the links in the email.

- Your work will be saved automatically as long as you have an internet connection. If you go offline, a yellow warning will show up at the bottom of the page. Before closing the page, make sure that this warning is not showing; your work will not be saved if it is.




- The sentences were chosen at random from several different documents. They do not have any bearing on one another. We are not interested in the structure of the text, only the individual segments.

- You may come across a source sentence that you have already seen. Edit the target sentence according to the guidelines below without referring to the work you have already done.

Linguistic Guidelines

- We are interested in obtaining "publishable quality" translation, which entails full post-editing. That means*:
    - This level of quality is generally defined as being **comprehensible** (i.e. an end user perfectly understands the content of the message), **accurate** (i.e. it communicates the same meaning as the source text), **stylistically fine**, though the style may not be as good as that achieved by a native-speaker human translator. **Syntax is normal, grammar and punctuation are correct.**
        - Aim for grammatically, syntactically and semantically correct translation.
        - Ensure that key terminology is correctly translated and that untranslated terms belong to the client's list of "Do Not Translate" terms."
        - Ensure that no information has been accidentally added or omitted.
        - Use as much of the raw MT output as possible.
        - Basic rules regarding spelling, punctuation and hyphenation apply.
        - Ensure that formatting is correct.

*TAUS MT Post-Editing Guidelines, 2010 (emphasis is mine)

- You may use any online or offline reference materials you need to help you with general vocabulary.

- Refer to the glossary text file you received via email for acronyms and other specialized terminology.

- Use whatever variety of English spelling and grammar you are most comfortable with (e.g. US, UK, etc.), but remain consistent throughout the two tasks.

Post-Editing Guidelines

- Before beginning to edit, check to make sure this icon is visible in the top menu bar:

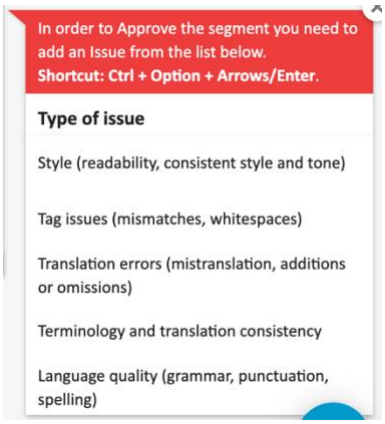

- If a translated segment does not require any editing, just click 

(or press CTRL+Enter) to go to the next segment.

- If a translated segment does require editing, make changes to the translation in the target window.
    - Once you finish editing the segment, select one or more categories and severity levels for the issue(s) you found in the translation in the box that pops up to the right side of the translation. If it does not pop up automatically, click the icon shown here to open it.

    - You may add comments to a segment in addition to identifying the error you corrected, but this is optional.

    - Then click APPROVED (or press CTRL+Enter) to go to the next segment.
    - **Important**: you will not be able to approve a segment that you have made edits to unless you select at least one error (see below) and severity level. You may add more than one error if appropriate.

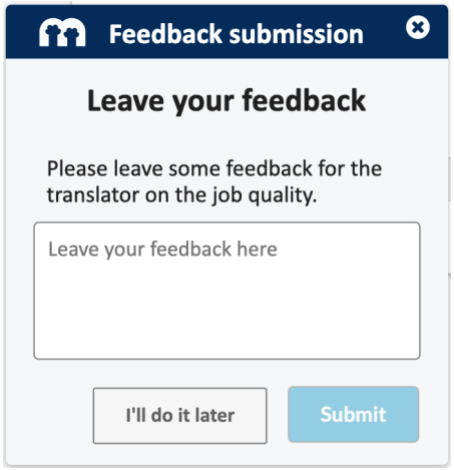



- The bar to the right of unapproved segments starts out **blue**. It turns **blue-and-white striped** when a segment has been modified but not approved. It turns **green** when a segment has been approved, regardless of whether it was modified. Before exiting the page when your work is complete, ensure that the bar is **solid green** for every segment.

- You may go back to a previous segment at any time, just be sure to approve it again if you make changes.
- Do not click any of the 3 tabs at the bottom of each segment ("Translation Matches," "TM Search" and "Glossary") or change any of the settings in the top menu bar.

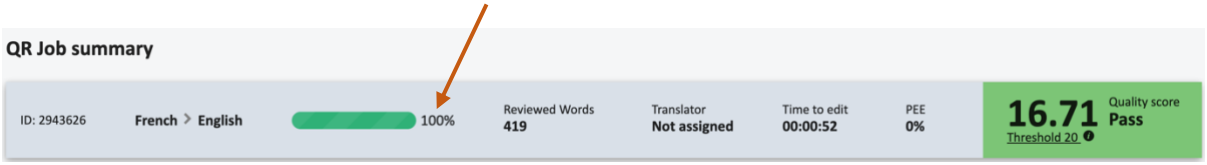
Deliverables

Once you have finished post-editing to the best of your ability and approved every segment, a feedback box should pop up in the bottom left-hand corner of your screen. Please type in your name and click **Submit** to let me know that you have completed the task:




To double check that you have approved every segment, click the QR icon in the top menu bar, and select "Open QR."


Ensure that the progress bar reads 100%.




When you are happy with your work and positive that both tasks have been saved, **send me an email to let me know that you are finished**. You do not need to download or send me any documents; I can access your work directly from my dashboard.

If you encounter any issues (technical or otherwise) or have questions/concerns about these instructions before or during the study, please contact me right away:

Abigail Kaplan
+33 07 77 77 45 60
Abigail.Kaplan@etu.unige.ch

## Appendix E: Glossary of terms provided to post-editors

AAH (Allocation aux adultes handicapés) -- AAH

AEEH (Allocation d'éducation de l'enfant handicapé) -- AEEH

AVS (Auxiliaire de Vie Scolaire) -- special education paraprofessional

CMI (carte mobilité inclusion) -- CMI card

ESAT (Établissements et services d'aide par le travail) -- ESAT

MDPH (Maison départementale des personnes handicapées) -- MDPH

PCH (prestation de compensation du handicap) -- PCH

Personne(s)/enfant(s)/travailleur(s)/etc. handicapées -- person/people/child/children/worker(s) with disabilities

Politique pour les personnes handicapées -- Disability Policy

## Appendix F: Excerpt of pre-edited and non-pre-edited segments, annotated based on the DQF-MQM and Easy-to-Read frameworks

| # | Source | Non-Pre-Edited Target | Pre-Edited Target | Translation Error Category | Translation Error Subcategory | Severity | Easy-to-Read Category | Easy-to-Read Guideline |
|---|---|---|---|---|---|---|---|---|
| 04 | Dans ce document, Marie-Arlette Carlotti explique ce qu'elle fait pour les personnes en situation de handicap et contre l'exclusion. | In this document, Marie-Arlette Carlotti explains what it does for people with disabilities and against exclusion. | In this document, Marie-Arlette Carlotti explains what she does for people with disabilities and against exclusion. | Accuracy \| - | Mistranslation \| - | Major \| - | Word \| - | Be careful with pronouns \| - |
| 11 | Un pays plus solidaire pour aider les personnes en grande difficulté | A country with more solidarity to help people in great difficulty | A more united country to help people in great difficulty | Style \| - | Unidiomatic \| - | Minor \| - | Word \| - | Use easy words \| - |
| 16 | Les autres personnes doivent faire attention aux personnes en situation de handicap et les aider quand elles ont besoin d'aide. | Other people must be careful people with disabilities and help them when they need help. | Other people should pay attention to people with disabilities and help them when they need help. | Fluency \| Accuracy | Grammar \| Mistranslation | Major \| Minor | General \| - | Use the right language \| - |
| 17 | Pour ne pas être exclues, les personnes en difficulté doivent demander des aides financières. | So as not to be excluded, people in difficulty must apply for financial assistance. | In order not to be excluded, people in difficulty must apply for financial assistance. | Style \| - | Unidiomatic \| - | Minor \| - | Sentence; Word \| Sentence | Positive sentences; Same word for same thing \| Positive sentences |
| 23 | Tous les ministres doivent penser aux personnes en situation de handicap quand ils font une loi. | All ministers must think to people with disabilities when they make a law. | All ministers must think about people with disabilities when making legislation. | Fluency \| - | Grammar \| - | Minor \| - | - \| Word | - \| Use easy words |
| 28 | Depuis le 16 octobre 2012, Marie-Arlette Carlotti veut que les Auxiliaires de Vie Scolaire se forment plus pour pouvoir aider tous les enfants en situation de handicap. | Since October 16, 2012, Marie-Arlette Carlotti wants to that the School Life Auxiliaries are formed more to be able to help all children with disabilities. | Since October 16, 2012, Marie-Arlette Carlotti wants the School Life Auxiliaries to be trained more so that they can help all children with disabilities. | Fluency; Terminology; Accuracy \| Terminology; Style | Grammar; General; Mistranslation \| Inconsistent with termbase; Unidiomatic | Major; Major; Major \| Major; Minor | General; Word \| General | Explain clearly; Use easy words\| Use the right language |