İstanbul Üniversitesi Çeviribilim Dergisi
Istanbul University Journal of Translation Studies,
JTS 2025, (23): 94–114

https://doi.org/10.26650/iujts.2025.1719141
Submitted 14.06.2025
Revision Requested 02.09.2025
Last Revision Received 23.09.2025
Accepted 03.11.2025

İstanbul Üniversitesi Çeviribilim Dergisi
Istanbul University Journal of Translation Studies

Research Article

Open Access

Readability Transfer Capabilities of Neural Machine Translation Services

Tuğrul Güngör [1]

[1] Istanbul Gelisim University, Istanbul Gelisim Vocational School/Foreign Languages and Cultures Department, Istanbul, Türkiye

Abstract Neural Machine Translation (NMT) services demonstrate high semantic accuracy, but their ability to convey the readability of the source text is understudied. This study, therefore, provides a comprehensive evaluation of readability transfer in four leading NMT services: Amazon Translate, Azure Translator, DeepL, and Google Cloud Translation, across the English-German, English-Turkish, and German-Turkish language pairs. For this analysis, translations from various genres were assessed using a combination of language-specific readability formulas and textual metrics. Results revealed a significant directional asymmetry: readability decreased when translating from English to German or Turkish, but increased from German or Turkish to English. Statistically insignificant but consistent differences were found among the four NMT services in readability scores, with target language and source text properties having a greater influence. The findings reveal that readability is not inherently preserved in NMT and is significantly influenced by the characteristics of the target language and the nature of the source text. This highlights the critical importance of considering readability metrics alongside semantic accuracy when evaluating machine translation, especially for applications that require high accessibility or target a specific level of accessibility, suggesting potential requirements for readability-focused post-editing.

Keywords Cross-lingual Readability · Language Typology · Neural Machine Translation (NMT) · Readability, Readability Transfer

Citation: Güngör, T. (2025). Readability transfer capabilities of neural machine translation services. *İstanbul Üniversitesi Çeviribilim Dergisi–Istanbul University Journal of Translation Studies*, (23), 94-114. https://doi.org/10.26650/iujts.2025.1719141

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

© 2025. Güngör, T.

Corresponding author: Tuğrul Güngör tgungor@gelisim.edu.tr

Introduction

Translation has historically depended on various methods, from human-driven approaches to rule-based systems. Recently, the rise of artificial intelligence has revolutionized the field, enabling machine-generated output to often rival human translation. This shift has not only changed professional practices but also sparked new debates in translation studies, focusing on quality, accessibility, and usability.

Numerous qualitative and quantitative empirical studies have proven that the capabilities of machine translation (MT) services have improved significantly over the past decade (Intento, 2021; Işım & Balcıoğlu, 2023; Jiao et al., 2023; Stasimioti et al., 2020; Son & Kim, 2023; Toral & Way, 2018; Wang et al., 2022; Wu et al., 2016). This advancement has reached a point where the output of MT can be indistinguishable from that of human translators (Nygård, 2024; Yirmibeşoğlu et al., 2023). Essentially, Neural Machine Translation (NMT) services imitate the behavior of human translators by encoding the message in the source language and decoding it in the target language, all while maintaining contextual awareness (Forcada, 2017; Pérez-Ortiz et al., 2022). Therefore, it can be said that, similar to professional human translators, NMT services also undergo a comprehension stage to maximize translation quality. Although AI-based translation services can provide excellent quality in terms of meaning and style transfer, the quality of readability transfer remains understudied.

The readability of a text measures how easily that text can be processed and understood by its intended audience. For example, a text designed for sixth-grade students should use appropriate vocabulary and sentence structure. Therefore, when translating a text, it is crucial not only to ensure semantic accuracy but also to maintain its readability, unless simplification is not a secondary goal.

From a translation studies perspective, this study is framed within two complementary approaches. First, a *typological perspective* highlights how structural asymmetries between inflectional languages such as English and German, and an agglutinative language such as Turkish, influence the readability of translated texts. Second, a *reader-oriented perspective* emphasizes that translation quality should not only be measured by semantic accuracy but also by how accessible and comprehensible the text is for its intended audience. By combining these perspectives, the study investigates readability transfer in NMT as both a linguistic and functional phenomenon.

Previous research on machine translation has primarily focused on semantic accuracy, domain-specific performance, and fluency, with limited focus on readability. While modern NMT services have achieved significant accuracy, they have not been systematically evaluated for how well they maintain or alter the readability of the source text. This is particularly important in fields such as education, healthcare, and public information, where readability plays a critical role. Although differences in readability inevitably reflect language typology, this study specifically examines whether current NMT systems maintain or alter these differences, thus addressing readability transfer as a separate aspect of machine translation quality.

This study assesses the readability transfer performance of four leading NMT services: *Amazon Translate*, *Azure Translator*, *DeepL*, and *Google Cloud Translation*. The primary objectives of this study are to: (1) quantify the changes in readability scores and textual metrics when translating texts between English, German, and Turkish; (2) determine the influence of translation direction and target language characteristics on readability transfer; and (3) evaluate whether significant differences exist among the selected NMT services regarding their impact on translation readability.

Readability Formulas

Readability formulas are quantitative tools designed to assess how easily a particular text can be read and comprehended by its intended audience. These formulas typically focus on surface-level linguistic features such as word length, sentence length, and syllable count. They provide standardized scores that facilitate comparisons between texts. In this study, we employed language-specific readability formulas to investigate whether and how readability is impacted in the context of machine translation.

Compatibility and Comparability

Due to the nature of languages, readability formulas are highly language-dependent. In other words, a formula that works for one language will likely not perform well for another. For example, Turkish is an agglutinative language, whereas English is primarily an inflectional language. Because of potentially significant differences in word syllable and sentence lengths, formulas developed for English may not yield accurate results for Turkish. For instance, the average length of words in syllables in the first chapter of *Anna Karenina* was 1.41 in English, 1.72 in German, and 2.76 in Turkish. The average word lengths per sentence for these same texts were 23.05 in English, 23.47 in German, and 10.16 in Turkish. Therefore, applying the same formula for all these texts is likely to produce unexpected results, especially for typologically different languages, whereas the application of language-specific formulas is more likely to yield similar results. To exemplify, applying language-specific reading ease formulas results in 64.49 for English, 55.99 for German, and 61.49 for Turkish, whereas applying the original Flesch formula for English results in 56.84 for German and 75.25 for Turkish. Depending on text-specific features, the gap can increase dramatically. For instance, for *Universal Declaration of Human Rights*, the RE scores for language-specific formulas are 35.62 for English, 39.63 for German, and 44.91 for Turkish, whereas the English RE formula results in 50.29 for German and 64.57 for Turkish. This indicates that an undefined range of deviation could occur if the formulas used are not tailored for the target languages.

On the other hand, formula adaptations typically borrow the scoring system of the original formula. For example, adaptations like Amstad's (1978) for German and Ateşman's (1997) for Turkish maintain the same 0–100 scale as the original Flesch (1948) formula. This consistency allows for a certain level of comparability across different languages. As a result, these formulas are widely used in multilingual readability research, even though they mainly focus on surface-level linguistic features.

Adaptations of Formulas

English is one of the first languages to have readability measurement formulas. Over time, other languages, following in the footsteps of English, developed their adaptations or equivalents of these formulas. Each readability formula uses different metrics to measure readability in various aspects: *(1) reading ease*, indicating text readability, and *(2) reading grade level*, reflecting the academic grade required to read the text.

Reading Ease

The Flesch Reading Ease formula was initially developed for English by Flesch (1948) and later adapted for German by Amstad (1978) and for Turkish by Ateşman (1997). All language variants of the formula perform the same operations using precalculated constants specific to each language. The expected score range is from 0 to 100, though higher or lower values are also possible. Refer to Table 1 for the interpretation of reading ease scores.

Table 1
Interpretation of reading ease scores.

Reading Ease Score	Text Readability
90-100	Very easy
80-90	Easy
70-80	Fairly easy
60-70	Standard
50-60	Fairly difficult
30-50	Difficult
0-30	Very difficult

Reading Grade

Kincaid et al. (1975) proposed the Flesch-Kincaid Reading Grade to measure the reading grade of English texts. Unlike the constant value adaptation in the Flesch Reading Ease, the equivalents of this formula require different parameters and procedures. Bezirci and Yılmaz (2010) developed a formula tailored for the Turkish texts, and Bamberger and Vanecek (1984) proposed the Wiener Sachtextformel (WSTF) formula for the German texts.

Unlike the scoring system in *reading ease* formulas, *reading grade* varies by country. Therefore, if precise identification of school-year qualifications is required, interpreting these scores should be approached differently, particularly for values lower than 12. However, grade levels above 12 indicate a college-level education, and grade levels above 16 suggest academic difficulty. There is no upper limit, and scores above 16 should be understood as representing a threshold of general difficulty without associating them with any specific school year.

Concerns for the Turkish Formulas

Kalyoncu and Memiş (2024) compared the consistencies of Turkish formulas and concluded that the formulas disputed not only each other but also the experts' opinions. Both Ateşman and Bezirci-Yılmaz did not fully align with the experts' views. Moreover, Bezirci-Yılmaz reported that the texts were more difficult than what Ateşman's formula and human evaluators indicated. However, although there are concerns, the limitations in the quantity and variability of test data, along with the number and demographics of participants, may not be sufficient to necessarily deprecate these formulas. Still, this pattern also appears in the findings of this study: while Turkish source texts often show RE scores similar to those of English and German counterparts, their RG levels are significantly higher. These differences are not caused by text complexity but are due to the Bezirci–Yılmaz formula's tendency to overestimate difficulty, a concern already highlighted in existing literature.

Studies on Readability

Existing studies primarily examine the readability of texts written for native speakers or those prepared for non-native speakers across various domains, including education (Erol, 2015; Çıplak & Balcı, 2022; Yılmaz & Temiz, 2014), finance (Gosselin et al., 2021; Loughran & McDonald, 2014), and medicine (Ay & Duranoğlu, 2022; Oliffe et al., 2019). The main objective of these studies is to determine whether the materials are accessible to their intended target audiences. Moreover, it is often unclear whether the contents or translations were created by humans or machines.

A recent study by Momenaei et al. (2023) assessed the readability of responses produced by ChatGPT-4, a Large Language Model (LLM), and concluded that the outputs were generally suitable. In contrast, a study by Cohen et al. (2024) found that the outputs of ChatGPT and Google Search were more difficult to read than those generated by humans. Further studies investigating the readability of LLMs also yield contradictory findings, either resulting in higher readability levels than anticipated (Hanci et al., 2024; Ozduran et al., 2025; Strzalkowski, 2024) or achieving the desired level suitable for the target audience (Gondode et al., 2024; Meyer et al., 2024). Consequently, it is difficult to claim that machine-generated content maintains a consistent level of readability at the time of creation. Therefore, the ability of machine translators to convey readability is highly questionable.

Another study by Zhou et al. (2017) concluded that readability is measured more reliably with longer texts, and the structure of the text (such as the use of acronyms and hyphenated words) significantly influences readability. Additionally, the surface-level measurement capability of these formulas was emphasized in earlier studies (DuBay, 2004; Collins-Thompson, 2014). Although the formula developers aimed to create systems that enable cross-linguistic comparisons (e.g., a 0-100 scale and academic grade level) and designed their formulas to suit the characteristics of the target languages, deviations within or between languages may occur depending on how the text is structured. At this point, how NMT services construct messages in the target language becomes critically important.

Methodology

This study aimed to achieve three main objectives: (1) to quantify changes in readability scores when translating between English, German, and Turkish; (2) to explore the impact of translation direction and target language characteristics on readability transfer; and (3) to examine differences among selected Neural Machine Translation (NMT) services regarding their effect on readability.

To meet these goals, we created a corpus of texts in the three languages and conducted bidirectional translations using four NMT services: Amazon Translate, Azure Translator, DeepL, and Google Cloud Translation. We assessed readability with language-specific formulas and measured textual statistics like sentence and word length using the *SmoothText*[1] Python library and *Stanza* (Qi et al., 2020), enabling a systematic evaluation of readability across languages and translation systems.

The Data

The data was obtained from texts across various genres. All texts were available in English, German, and Turkish, either through human or curated machine translation. None of the translations were labeled as "adapted," "graded," or "simplified." The selected texts were comparable in terms of readability across the three languages, which allowed the study to observe how NMT specifically altered readability levels.Table 2. *Information about the adopted test data.*

Table 2

Information about the adopted test data.

Source	Parts Included	Genre	Character Length		
			English	German	Turkish
Flores+*	All (devtest)	General	130.4	151.99	134.15
			35-368	34-408	44-386
Anna Karenina	Chapter 1	Fictional	5,405	5,714	4,504

[1] https://pypi.org/project/smoothtext/

Source	Parts Included	Genre	Character Length		
			English	German	Turkish
Google FAQ	All	Technical	6,530	7,652	6,979
The Little Prince	Chapter 1	Fictional	2,811	2,763	2,254
The Social Contract	Book 1 / Chapter 1	Philosophical	879	1,032	902
Universal Declaration of Human Rights	All except the preamble	Legal	8,597	9,749	8,281

Due to a limitation in *Amazon Translate*[2], each piece of text was required to be fewer than ten thousand characters. Therefore, texts exceeding this limit were discarded or truncated to maintain compatibility with other translation services. See Table 2 for a complete list of the data used.

Translations

For each language pair (English-German, English-Turkish, and German-Turkish), the respective texts were translated in both directions using *Amazon Translate*, *Azure Translator*, *DeepL*, and *Google Cloud Translation*. These services were accessed, and the translations were conducted through the default API provided by their respective companies. The Python client libraries published by the service-providing companies were utilized whenever possible. No fine-tuning was done, and when a service offered alternative translations, the first option was always selected.

Readability Formulas

To evaluate readability in English, German, and Turkish, this study used six established formulas, categorized into two groups: *reading ease* and *reading grade*.

For *reading ease*, the Flesch Reading Ease formula (Flesch, 1948) was selected for English as it is the most widely used metric in readability research. To ensure comparability across the three languages, its direct adaptations were used: Amstad's (1978) version for German and Ateşman's (1997) version for Turkish. These adaptations maintain the original structure of the Flesch formula while incorporating language-specific constants to account for differences in average word and sentence length.

For *reading grade* level, three widely utilized and well-established measures were employed. The Flesch–Kincaid Grade Level (Kincaid et al., 1975) was selected for English, as it has become a standard tool in educational and applied contexts, particularly for determining the appropriate school grade level for texts. For German, the Wiener Sachtextformel (Bamberger & Vanecek, 1984) was chosen due to its popularity and widespread adoption as a formula for estimating the readability of expository texts. For Turkish, the Bezirci–Yılmaz formula (2010) was utilized, which is widely recognized in Turkish readability research.

These six formulas were selected to ensure a balanced framework for cross-linguistic comparison. Each formula has been developed or adapted to reflect the specific structural characteristics of the language it represents. At the same time, all are based on comparable surface-level features such as sentence length, word length, and syllable count, and provide common scoring systems such as 0-100 or academic grade. Other formulas (e.g., Gunning Fog, SMOG) were excluded because they are primarily designed for English and depend on language-specific assumptions that limit their applicability in multilingual contexts. For instance, the Gunning Fog Index assumes that words with three or more syllables are inherently difficult,

[2] *Amazon Translate* did not support translations of texts longer than ten thousand characters directly between German and Turkish. Text Translation for this language pair was possible, but the API limited it to ten thousand characters, and Document Translation did not support the language pair.

which is misleading in agglutinative languages like Turkish, where words such as *ar-ka-daş-lar* ("friends") are morphologically simple. Similarly, SMOG depends on the frequency of polysyllabic words, which tends to overstate difficulty in German and Turkish.

In addition, despite the concerns, the Bezirci–Yılmaz formula was chosen because it remains the most widely used readability measure for Turkish and ensures comparability with the Flesch-based metrics used for English and German. However, its tendency to overestimate grade-level difficulty is acknowledged, and this limitation is further discussed in Section 2.3.

Calculating Readability Scores

The readability scores for both the source texts and their translations were calculated using *SmoothText*, which incorporates these language-specific formulas. This library offers validated implementations of the selected formulas within a unified framework, ensuring consistency across languages. Its integration with Stanza also enabled direct calculation of surface-level features. Although other implementations exist, it was the most practical and reliable option, and no modifications were made to the formulas. In addition, no other libraries offered a unified framework to work with all three languages.

The calculations addressed two aspects: *(1) reading ease,* using Flesch's (1948) formula for English, Amstad's (1978) for German, and Ateşman's (1997) for Turkish; and *(2) reading grade*, applying Flesch-Kincaid Grade Level (Kincaid et al., 1975) for English, the Bezirci-Yılmaz (Bezirci & Yılmaz, 2010) formula for Turkish, and the Wiener Sachtextformel (Bamberger & Vanecek, 1984) for German.

The following thresholds were established to evaluate the significance of readability differences:

- Differences in reading ease scores under 10 points were considered insignificant, while those of 10 points or more were significant.
- Differences below one grade level are insignificant, while those of one grade level or greater are significant.

Calculating Textual Statistics

The *SmoothText* library also provided an interface for using *Stanza,* which was used to count sentences, words, and syllables in the texts. Subsequently, the average word and sentence lengths were calculated. These textual statistics complemented the formula-based scores by showing how surface-level linguistic structures (e.g., average syllable length of words in Turkish versus English) influenced the observed differences in readability.

Findings

To assess the readability transfer performances of NMT services, five texts and one dataset with 1,012 entries were translated between English-German, English-Turkish, and German-Turkish pairs using four popular translators: *Amazon Translate*, *Azure Translator*, *DeepL*, and *Google Cloud Translation*. Then, the readability and textual metrics of both the source and translated versions were calculated. The tables in this section present the source language (SL), target language (TL), translation model (TM) that performed the translation, reading ease (RE) score, reading grade (RG) level, sentence count (SC), sentence word length (SWL), and word syllable length (WSL) for each text. It should be noted that the metrics presented for the Flores+ dataset are average values computed over its 1,012 constituent entries.

English and German

This section presents the readability and textual metrics analysis for bidirectional translations between English and German.

English→German

The results for translations from English to German are presented in Table 3. RE scores generally decreased after translation into German, a consistent trend across the *Flores+* dataset (from 48.74 to 43.53-45.41), *Anna Karenina* (from 64.49 to 59.68-61.63), *Google FAQ* (from 48.36 to 39.05-43.84), *The Little Prince* (from 74.16 to 68.59-70.1), and *The Social Contract* (from 77.49 to 63.45-67.49). A notable exception was *Universal Declaration of Human Rights*, where the source text had an RE of 35.62, which subsequently increased in all German translations, ranging from 40.33 to 42.49.

RG level increases were observed for the *Flores+* dataset (from 11.17 to 11.47-11.9), *The Little Prince* (from 6.3 to 6.46-6.85), and *The Social Contract* (from 6.32 to 7.11-7.94). On the contrary, decreases were observed for *Anna Karenina* (from 9.68 to 8.08-8.53) and *Universal Declaration of Human Rights* (from 13.87 to 12.37-12.76). Furthermore, in the case of *Google FAQ*, the RG levels did not exhibit a consistent increase or decrease, as they varied depending on the model (e.g., decreasing to 11.38-11.47 and increasing to 11.64-12.13 from 11.51).

Examining the textual metrics, SC remained relatively stable across sources and translations, although minor variations were observed. SWL often showed a slight decrease in the German translations compared to the English source. Conversely, WSL consistently increased in all German translations, reflecting the morphological characteristics of the German language.

In summary, translation from English to German generally resulted in decreased RE and often increased RG levels. However, text-specific variations, particularly for the Universal Declaration of Human Rights, and subtle differences between translation models were also noticeable. Overall, no single translation model consistently produced the highest or lowest readability scores across all texts, with performance varying depending on the specific text.

Table 3

Readability and Textual Metrics of Source (English) and Translated (German) Versions.

Text	Metric	English→German				
		Source	Amazon Translate	Azure Translator	DeepL	Google Cloud Translation
Flores+	RE	48.74	44.09	45.17	45.41	43.53
	RG	11.17	11.7	11.48	11.47	11.9
	SC	1.13	1.19	1.18	1.17	1.21
	SWL	20.22	19.06	19.48	19.49	18.62
	WSL	1.63	2.0	1.97	1.97	2.01
Anna Karenina	RE	64.49	59.68	61.61	61.63	60.52
	RG	9.68	8.53	8.08	8.24	8.32
	SC	42.0	46.0	46.0	45.0	46.0
	SWL	23.05	19.96	20.24	20.11	19.85
	WSL	1.41	1.72	1.68	1.68	1.7
Google FAQ	RE	48.36	43.84	42.82	42.07	39.05
	RG	11.51	11.38	11.47	11.64	12.13
	SC	51.0	52.0	49.0	52.0	47.0

Text	Metric	English→German				
		Source	Amazon Translate	Azure Translator	DeepL	Google Cloud Translation
	SWL	21.35	21.08	22.47	20.83	22.74
	WSL	1.62	1.97	1.96	2.0	2.02
The Little Prince	RE	74.16	69.9	70.1	69.49	68.59
	RG	6.3	6.46	6.47	6.69	6.85
	SC	35.0	37.0	38.0	37.0	37.0
	SWL	14.89	13.16	13.11	13.14	12.7
	WSL	1.39	1.66	1.65	1.66	1.69
The Social Contract	RE	77.49	64.94	67.49	65.43	63.45
	RG	6.32	7.7	7.11	7.72	7.94
	SC	10.0	10.0	12.0	10.0	10.0
	SWL	16.8	16.5	13.83	16.6	16.0
	WSL	1.33	1.68	1.69	1.67	1.72
Universal Declaration of Human Rights	RE	35.62	41.53	42.49	41.61	40.33
	RG	13.87	12.72	12.37	12.76	12.69
	SC	60.0	64.0	64.0	64.0	64.0
	SWL	23.72	20.72	21.33	21.05	20.97
	WSL	1.74	2.01	1.99	2.01	2.03

German→English

Turning to the translations from German to English, Table 4 provides the detailed results. For this translation direction, RE scores generally increased following translation into English. This trend was observed for the *Flores+* dataset (from 43.91 to 49.32-49.91), *Anna Karenina* (from 55.99 to 60.02-63.15), *Google FAQ* (from 42.6 to 47.56-49.02), *The Little Prince* (from 67.98 to 73.53-76.51), and *The Social Contract* (from 69.1 to 79.83-83.16). The only exception was *Universal Declaration of Human Rights*, where the source RE of 39.63 decreased to a range of 35.11-37.19.

The changes regarding RG levels were mixed compared to the German source. RG levels decreased for the *Flores+* dataset (from 11.78 to 10.96-11.05), *Google FAQ* (from 11.66 to 11.27-11.58), *The Little Prince* (from 7.23 to 5.02-5.57), and *The Social Contract* (from 7.01 to 5.38-5.9). Conversely, RG levels increased for *Anna Karenina* (from 9.37 to 10.24-11.44) and *Universal Declaration of Human Rights* (from 12.85 to 13.49-13.84).

Table 4

Readability and Textual Metrics of Source (German) and Translated (English) Versions.

Text	Metric	German→English				
		Source	Amazon Translate	Azure Translator	DeepL	Google Cloud Translation
Flores+	RE	43.91	49.63	49.91	49.62	49.32
	RG	11.78	10.96	10.99	10.99	11.05
	SC	1.22	1.18	1.17	1.17	1.18
	SWL	19.01	19.87	20.13	19.98	20.03
	WSL	2.0	1.62	1.61	1.62	1.62
Anna Karenina	RE	55.99	63.15	60.46	60.02	63.08

Text	Metric	German→English				
		Source	Amazon Translate	Azure Translator	DeepL	Google Cloud Translation
	RG	9.37	10.25	11.44	11.4	10.56
	SC	38.0	38.0	35.0	35.0	37.0
	SWL	23.47	23.95	27.06	26.69	25.81
	WSL	1.72	1.4	1.4	1.41	1.39
Google FAQ	RE	42.6	47.56	48.47	49.02	48.73
	RG	11.66	11.5	11.41	11.27	11.58
	SC	54.0	53.0	53.0	54.0	51.0
	SWL	20.19	20.85	21.0	20.76	21.84
	WSL	2.0	1.63	1.62	1.62	1.61
The Little Prince	RE	67.98	73.53	75.11	76.51	75.51
	RG	7.23	5.57	5.3	5.02	5.23
	SC	37.0	37.0	38.0	40.0	39.0
	SWL	11.68	11.57	11.39	11.03	11.33
	WSL	1.72	1.44	1.42	1.41	1.42
The Social Contract	RE	69.1	83.16	82.6	79.83	80.57
	RG	7.01	5.48	5.38	5.9	5.71
	SC	12.0	11.0	12.0	11.0	11.0
	SWL	13.75	16.64	15.92	16.45	16.09
	WSL	1.66	1.26	1.28	1.3	1.3
Universal Declaration of Human Rights	RE	39.63	37.19	35.64	35.11	35.84
	RG	12.85	13.49	13.84	13.83	13.82
	SC	63.0	61.0	60.0	60.0	60.0
	SWL	21.27	23.07	23.58	23.27	23.63
	WSL	2.04	1.73	1.74	1.75	1.74

Examining the textual metrics, SC was observed to remain relatively stable between the German source and English translations, with only minor variations and often slight decreases. SWL generally increased; however, a notable exception was *The Little Prince*, where SWL decreased from 11.68 to 11.39-11.57. In contrast, reflecting linguistic differences, WSL consistently decreased across all texts and models.

In summary, translating from German to English often resulted in increased RE, but had varied effects on RG depending on the text. These readability shifts were accompanied by a consistent decrease in WSL and a general increase in SWL. Similar to the English→German direction, no single translation model consistently yielded the most or least readable outputs across all texts based on these metrics.

English and Turkish

Following the analysis of the English↔German pair, this section details the readability transfer observed for translations between English and Turkish.

English→Turkish

Table 5 summarizes the calculated metrics for the translation from English to Turkish. A similar trend to the English→German was observed, as RE scores consistently decreased in the *Flores+* dataset (from 48.74 to 43.78-45.46), *Anna Karenina* (from 64.49 to 48.83-51.25), *Google FAQ* (from 48.36 to 32.84-33.38), *The Little*

Prince (from 74.16 to 66.18-68.85), and *The Social Contract* (from 77.49 to 64.91-70.58) but increased only for *Universal Declaration of Human Rights* (from 35.62 to 44.08-46.83).

Contrary to the general decrease in RE, RG levels showed a consistent increase: from 11.17 to 17.87-18.78 for the *Flores+* dataset, from 9.68 to 16.39-18.01 for *Anna Karenina*, from 11.51 to 25.03-27.2 for *Google FAQ*, from 6.3 to 10.84-12.24 for *The Little Prince*, from 6.32 to 10.75-13.19 for *The Social Contract*, and from 13.87 to 19.81-22.06 for *Universal Declaration of Human Rights*.

Examining the textual metrics, SC was observed to remain highly stable across texts and models (e.g., unchanged at 60 for *Universal Declaration of Human Rights*). SWL consistently decreased across all translations, while WSL consistently increased, reflecting the characteristics of the Turkish language.

In summary, translating from English to Turkish generally resulted in decreased RE and in consistently increased RG. The shifts in SC, SWL, and WSL were consistent across all translations. Overall, no translation model consistently produced the most or least distinctive values.

Table 5

Readability and Textual Metrics of Source (English) and Translated (Turkish) Versions.

Text	Metric	English→Turkish				
		Source	Amazon Translate	Azure Translator	DeepL	Google Cloud Translation
Flores+	RE	48.74	45.46	45.15	43.78	44.89
	RG	11.17	17.87	18.21	18.78	18.72
	SC	1.13	1.15	1.16	1.16	1.17
	SWL	20.22	15.42	15.81	15.88	16.09
	WSL	1.63	2.81	2.8	2.83	2.79
Anna Karenina	RE	64.49	51.25	50.27	50.66	48.83
	RG	9.68	16.43	16.64	16.39	18.01
	SC	42.0	41.0	41.0	41.0	40.0
	SWL	23.05	15.39	16.29	16.49	16.83
	WSL	1.41	2.67	2.64	2.62	2.64
Google FAQ	RE	48.36	33.37	33.38	32.84	33.31
	RG	11.51	25.22	25.03	25.61	27.2
	SC	51.0	49.0	48.0	48.0	47.0
	SWL	21.35	16.06	16.75	16.69	17.26
	WSL	1.62	3.07	3.03	3.05	3.0
The Little Prince	RE	74.16	67.56	68.85	67.79	66.18
	RG	6.3	10.86	10.84	11.61	12.24
	SC	35.0	33.0	33.0	33.0	32.0
	SWL	14.89	10.42	10.64	10.7	10.91
	WSL	1.39	2.59	2.54	2.57	2.59
The Social Contract	RE	77.49	70.58	66.89	64.91	68.49
	RG	6.32	11.49	13.19	11.57	10.75
	SC	10.0	11.0	10.0	10.0	10.0
	SWL	16.8	10.45	12.5	13.4	12.5
	WSL	1.33	2.51	2.47	2.46	2.43
	RE	35.62	45.46	44.28	44.08	46.83

Text	Metric	English→Turkish				
		Source	Amazon Translate	Azure Translator	DeepL	Google Cloud Translation
Universal Declaration of Human Rights	RG	13.87	21.48	21.45	22.06	19.81
	SC	60.0	60.0	60.0	60.0	60.0
	SWL	23.72	17.83	18.62	18.3	17.32
	WSL	1.74	2.66	2.64	2.66	2.66

Turkish→English

The outcomes of translating from Turkish to English are shown in Table 6. Contrary to the trend in English-to-Turkish translations, Turkish-to-English translations resulted in mostly increased RE. The *Flores+* dataset (from 44.76 to 48.86-50.21), *Anna Karenina* (from 61.48 to 76.11-78.89), *Google FAQ* (from 31.91 to 47.54-50.79), and *The Social Contract* (from 65.03 to 80.69-87.24) had an increase in their REs, whereas *Universal Declaration of Human Rights* (from 44.91 to 35.56-40.96) showed a decrease. *The Little Prince* also showed variance in RE scores, with increases for three models (76.36-78.18) and a decrease for one (*Google Cloud Translation*: 73.25), compared to the source (75.03).

Unlike the variance in REs, RG levels displayed a consistent decreasing trend: from 18.08 to 10.85-11.12 for the *Flores+* dataset, from 10.85 to 5.39-5.89 for *Anna Karenina*, from 24.99 to 10.99-11.52 for *Google FAQ*, from 7.7 to 5.02-5.72 for *The Little Prince*, from 12.62 to 4.38-6.01 for *The Social Contract*, and from 21.15 to 12.08-13.92 for *Universal Declaration of Human Rights*.

SC remained stable across all sources and translations, except for *Google Cloud Translation* of *Universal Declaration of Human Rights*, which showed a notably higher sentence count of 74 compared to 62-66 from the other models and the source. SWLs trended upward, while WSLs consistently decreased in the English translations compared to the Turkish source, reflecting morphological differences between the languages.

In summary, the REs mostly increased, while the RGs consistently decreased in Turkish-to-English translations. Overall, none of the translation models produced distinctly high or low values for readability or textual metrics, except for one notable case, as mentioned above, with *Google Cloud Translation*.

Table 6

Readability and Textual Metrics of Source (Turkish) and Translated (English) Versions.

Text	Metric	Turkish→English				
		Source	Amazon Translate	Azure Translator	DeepL	Google Cloud Translation
Flores+	RE	44.76	49.81	50.21	49.23	48.86
	RG	18.08	10.96	10.85	10.86	11.12
	SC	1.2	1.18	1.2	1.18	1.15
	SWL	15.54	19.97	19.73	19.24	20.05
	WSL	2.82	1.62	1.61	1.63	1.63
Anna Karenina	RE	61.48	77.47	76.11	78.89	78.39
	RG	10.85	5.72	5.89	5.39	5.73
	SC	58.0	62.0	63.0	64.0	63.0
	SWL	10.16	14.39	14.33	13.86	14.94
	WSL	2.76	1.36	1.37	1.35	1.34
Google FAQ	RE	31.91	47.54	49.82	49.31	50.79

Text	Metric	Turkish→English				
		Source	Amazon Translate	Azure Translator	DeepL	Google Cloud Translation
	RG	24.99	11.52	11.1	11.27	10.99
	SC	53.0	54.0	54.0	54.0	54.0
	SWL	15.83	20.93	20.54	20.91	20.63
	WSL	3.13	1.63	1.61	1.61	1.6
The Little Prince	RE	75.03	76.36	76.91	78.18	73.25
	RG	7.7	5.39	5.17	5.02	5.72
	SC	40.0	35.0	36.0	35.0	35.0
	SWL	7.7	12.43	11.86	11.97	12.03
	WSL	2.58	1.39	1.39	1.38	1.43
The Social Contract	RE	65.03	80.69	85.58	85.31	87.24
	RG	12.62	6.01	4.86	4.38	4.49
	SC	11.0	11.0	12.0	13.0	12.0
	SWL	11.36	17.36	15.5	13.38	14.92
	WSL	2.59	1.28	1.25	1.28	1.23
Universal Declaration of Human Rights	RE	44.91	36.88	37.79	35.56	40.86
	RG	21.15	13.68	13.31	13.92	12.08
	SC	62.0	63.0	66.0	62.0	74.0
	SWL	17.95	23.63	22.65	23.85	19.43
	WSL	2.66	1.73	1.73	1.74	1.73

German and Turkish

Finally, this section examines the outcomes of bidirectional translations between German and Turkish, regarding readability and textual metrics.

German→Turkish

The calculated metrics for German to Turkish translation are presented in Table 7. The translations resulted in varied RE trends: The *Flores+* dataset (from 43.91 to 45.23-46.6), *The Little Prince* (from 67.98 to 71.49-73.9), and *Universal Declaration of Human Rights* (from 39.63 to 43.4-45.06) had an increasing trend, whereas *Anna Karenina* (from 55.99 to 43.06-52.65), *Google FAQ* (from 42.6 to 33.29-35.52), and *The Social Contract* (from 69.1 to 62.34-67.49) had a decreasing trend.

Unlike the variance in trends of the REs of the translations, the RGs had a consistent increasing trend: The *Flores+* dataset increased from 11.78 to 17.45-18.34, *Anna Karenina* from 9.37 to 16.47-21.19, *Google FAQ* from 11.66 to 22.86-25.38, *The Little Prince* from 7.23 to 8.64-9.36, *The Social Contract* from 7.01 to 10.88-13.83, and *Universal Declaration of Human Rights* from 12.85 to 20.74-22.87.

Regarding the textual metrics, SCs were observed to be highly stable, with only minor variations. Similar to English-to-Turkish translations, SWLs decreased, while WSLs increased, in correlation with Turkish characteristics.

In summary, REs varied in direction; half of the records in the table indicated an increase, while the other half indicated a decrease. On the other hand, RGs consistently increased. Overall, none of the translation models displayed distinctive high or low outputs.

Table 7

Readability and Textual Metrics of Source (German) and Translated (Turkish) Versions.

Text	Metric	German→Turkish				
		Source	Amazon Translate	Azure Translator	DeepL	Google Cloud Translation
Flores+	RE	43.91	46.6	46.39	45.23	46.58
	RG	11.78	17.45	17.69	18.34	18.21
	SC	1.22	1.22	1.22	1.23	1.24
	SWL	19.01	14.92	15.22	15.21	15.5
	WSL	2.0	2.81	2.8	2.82	2.78
Anna Karenina	RE	55.99	52.65	43.59	43.06	43.99
	RG	9.37	16.47	19.72	20.9	21.19
	SC	38.0	37.0	33.0	33.0	34.0
	SWL	23.47	16.65	19.82	20.06	19.47
	WSL	1.72	2.56	2.58	2.57	2.59
Google FAQ	RE	42.6	35.18	35.52	34.33	33.29
	RG	11.66	22.86	22.83	23.43	25.38
	SC	54.0	51.0	52.0	52.0	49.0
	SWL	20.19	15.49	15.69	15.38	16.82
	WSL	2.0	3.07	3.05	3.1	3.03
The Little Prince	RE	67.98	72.42	73.9	72.31	71.49
	RG	7.23	9.36	8.64	8.94	9.25
	SC	37.0	36.0	38.0	37.0	38.0
	SWL	11.68	8.64	8.61	8.76	8.97
	WSL	1.72	2.59	2.55	2.58	2.59
The Social Contract	RE	69.1	62.34	67.49	64.88	63.83
	RG	7.01	13.18	10.88	12.91	13.83
	SC	12.0	10.0	11.0	11.0	11.0
	SWL	13.75	11.2	12.45	11.64	11.82
	WSL	1.66	2.67	2.46	2.58	2.59
Universal Declaration of Human Rights	RE	39.63	45.06	43.52	43.4	44.69
	RG	12.85	21.72	22.12	22.87	20.74
	SC	63.0	60.0	60.0	60.0	60.0
	SWL	21.27	17.78	18.62	18.08	17.75
	WSL	2.04	2.67	2.66	2.69	2.68

Turkish→German

Table 8 presents the readability and textual metrics for Turkish-to-German translations. Similar to the trend in German-to-Turkish translations, there is variance in RE trends: it increased for *Anna Karenina* (from 61.48 to 66.68-69.06), *Google FAQ* (from 24.99 to 10.96-11.53), and *The Social Contract* (from 65.03 to 67.69-73.06), decreased for *The Little Prince* (from 75.03 to 68.81-72.09) and *Universal Declaration of Human Rights* (from 44.91 to 40.51-42.59). However, the RE scores of translations of the *Flores+* dataset mostly increased from 44.76 to 44.89-45.83, whereas the calculated score was 43.98, indicating an inverse trend, only for *Google Cloud Translation*.

Similar to the consistent trends in the German to Turkish translations, the RGs in Turkish to German translations showed a consistent downward trend: decreasing from 18.08 to 11.37-11.81 for the *Flores+* dataset, from 10.85 to 7.06-7.35 for *Anna Karenina*, from 24.99 to 10.96-11.53 for *Google FAQ*, from 7.7 to 6.4-7.22 for *The Little Prince*, from 12.62 to 6.16-7.71 for *The Social Contract*, and from 21.15 to 12.28-12.62 for *Universal Declaration of Human Rights*.

The SCs did not vary significantly across translations. SWLs showed a consistent increase, while WSLs showed a consistent decrease across the German translations compared to the Turkish source texts.

In summary, translating from Turkish to German mostly increased RE scores and consistently decreased RG scores. None of the translation models produced distinctively high or low values.

Table 8

Readability and Textual Metrics of Source (Turkish) and Translated (German) Versions.

Text	Metric	Turkish→German				
		Source	Amazon Translate	Azure Translator	DeepL	Google Cloud Translation
Flores+	RE	44.76	44.89	45.83	45.52	43.98
	RG	18.08	11.58	11.37	11.49	11.81
	SC	1.2	1.23	1.23	1.21	1.27
	SWL	15.54	18.76	18.99	18.82	17.91
	WSL	2.82	1.99	1.97	1.98	2.02
Anna Karenina	RE	61.48	68.77	66.68	69.06	67.33
	RG	10.85	7.06	7.35	7.12	7.18
	SC	58.0	61.0	60.0	61.0	62.0
	SWL	10.16	14.38	14.78	14.23	14.29
	WSL	2.76	1.66	1.68	1.65	1.68
Google FAQ	RE	31.91	43.21	45.79	45.77	43.54
	RG	24.99	11.53	11.13	10.96	11.45
	SC	53.0	56.0	55.0	59.0	57.0
	SWL	15.83	20.25	20.35	19.05	19.14
	WSL	3.13	1.99	1.95	1.97	2.01
The Little Prince	RE	75.03	70.74	72.09	70.58	68.81
	RG	7.7	6.78	6.4	6.58	7.22
	SC	40.0	35.0	35.0	31.0	36.0
	SWL	7.7	12.69	12.43	13.29	11.53
	WSL	2.58	1.65	1.63	1.64	1.7
The Social Contract	RE	65.03	68.32	73.06	67.69	70.86
	RG	12.62	7.71	6.16	7.05	7.06
	SC	11.0	14.0	13.0	12.0	13.0
	SWL	11.36	13.0	13.54	16.67	14.0
	WSL	2.59	1.69	1.6	1.64	1.63
Universal Declaration of Human Rights	RE	44.91	40.51	42.59	41.84	41.7
	RG	21.15	12.62	12.28	12.46	12.48
	SC	62.0	65.0	66.0	66.0	67.0
	SWL	17.95	21.62	21.76	21.12	20.49

Text	Metric	Turkish→German				
		Source	Amazon Translate	Azure Translator	DeepL	Google Cloud Translation
	WSL	2.66	2.01	1.98	2.0	2.01

Discussion

This study examined the readability transfer of four leading NMT services for English, German, and Turkish language pairs using established readability formulas and textual metrics. Results show that readability is not consistently preserved during translation, being significantly affected by the typological and morphological characteristics of the source and target languages and, to a lesser extent, by the source text's properties and the translation model used.

Directional Asymmetry and Linguistic Influence

The most interesting finding was the consistent pattern observed when translating between English and morphologically richer languages, German and Turkish. English→German translations (Table 3) and English→Turkish translations (Table 5) generally resulted in decreased RE and often increased RG levels. On the contrary, German→English translations (Table 4) and Turkish→English translations (Table 6) consistently resulted in increased RE and generally decreased RG levels. This bidirectional asymmetry strongly suggests that the target language's intrinsic linguistic features significantly impact the translated text's readability, often overriding the readability of the source text.

The main reason behind this appears to be WSL. German, and even more so, Turkish, utilize longer and more polysyllabic words compared to English, due to compounding in German and agglutination in Turkish. As observed, WSL consistently increased when translating into German or Turkish and consistently decreased when translating into English. This aligns with the core components of the readability formulas used, as it directly influences the calculated scores. While SWL also varied (often decreasing into German or Turkish and increasing into English or German), its impact seemed less consistent and potentially secondary to WSL in affecting the observed main RE and RG trends. SC remained relatively stable, suggesting the models mostly preserved sentence boundaries.

German↔Turkish translations (Tables 7 and 8) presented a more varied result for RE but maintained consistent trends for RG and WSL. It is important to note that these two languages are both complex, differing in type: inflectional vs. agglutinative. Translating into Turkish consistently increased RG and WSL, while translating into German consistently decreased RG and WSL. This further underscores the dominant role of target language morphology, especially agglutinative Turkish's tendency towards higher syllable counts, which affected the calculation of the readability scores, even when the source language is also complex. The mixed RE results in these pairs might indicate limitations in the RE formulas in considering nuances between two non-English and complex languages, or genuine variability in how easily sentence structures are formed.

It is important to highlight that the source texts used in this study were available in human or curated machine translations in English, German, and Turkish, all of which showed similar levels of readability, excluding RG levels for certain Turkish texts due to the known limitation of the Bezirci-Yılmaz formula. The significant asymmetries reported in this study only became apparent after the application of NMT. This indicates that the differences cannot be attributed solely to the typological contrasts between these languages. Instead, they point to the limitations of current NMT systems in preserving readability when faced with structural differences across languages. This distinction emphasizes that readability transfer is not only

a linguistic concern but also an issue related to translation technology, revealing a specific limitation of NMT services. Similar observations have been reported in previous research, which showed that morphological complexity, such as compounding in German or agglutination in Turkish, directly affects formula-based readability scores and contributes to cross-linguistic asymmetries (Crossley et al., 2011; François, 2014).

Text-Specific Variations and Exceptions

While general trends were strong, exceptions highlight the relationship between the source text characteristics and the translation process. *Universal Declaration of Human Rights*, characterized by low source RE in English and German, often defied the trend. Its RE frequently increased when translated from English into German or Turkish, and decreased when translated into English. This might be due to a *floor effect*, where the source text was already near the bottom of the readability scale. Regardless of the target language morphology, this effect, combined with the possibility that translating complex legalistic sentence structures inherently leads to some degree of simplification or restructuring. Thereby, the calculated ease of the text increases. In addition, the mixed RG results for *Google FAQ* in some directions also suggest that technical content might interact differently with translation models or readability formulas.

Translation Model Performance

A key finding is that no single translation model consistently produced the most or least readable output across all texts, language pairs, and metrics. While subtle differences existed for specific text-model combinations, the dominant factors affecting readability scores were the source and target language pairings and the specific text itself, rather than the translator. This suggests that current NMT systems, while achieving high semantic accuracy, may not be optimized for preserving or targeting specific readability levels. These systems potentially prioritize meaning and fluency over matching source text complexity metrics. The occasional outliers (e.g., *Google Cloud Translator*'s higher SC for *Universal Declaration of Human Rights* in the Turkish→English direction) were rare and did not indicate any consistent patterns.

The observation that no MT system consistently outperforms the others suggests that NMT services are not yet optimized for maintaining readability. This highlights the need to treat readability transfer as a separate aspect of MT quality, in addition to semantic accuracy, which NMT continues to encounter considerable challenges in.

Relation to Existing Literature and Concerns

This study confirms the gap identified in the introduction: while MT quality in terms of meaning is well-studied, readability transfer remains less explored. The findings in this study quantify how readability, as measured by standard formulas, is significantly altered during translation, often predictably based on language typology. However, it is important to note that readability formulas are not part of the translation process; instead, they serve as benchmarking tools for the product. For this reason, rather than being a main cause of the incapability of NMTs in this matter, they serve as an indicator of this incapability.

The results also provide context for the concerns raised about Turkish readability formulas (Kalyoncu & Memiş, 2024). While the Bezirci-Yılmaz formula showed consistent trends (increasing English→Turkish and German→Turkish, decreasing in Turkish→English and Turkish→German), aligning directionally with linguistic expectations (changes such as SWL and WSL), the absolute values produced were often very high (e.g., >20 for *Google FAQ* for the English→Turkish pair), potentially reflecting the formula's tendency to report higher difficulty, as noted by Kalyoncu and Memiş. This emphasizes the need for caution when interpreting absolute scores, particularly for Turkish, even if the directional trends appear robust. Moreover, RG is expected to be in negative correlation with RE, as higher RG indicates higher difficulty, whereas higher RE indicates less

difficulty and vice versa. However, translations from and to Turkish, specifically for *Universal Declaration of Human Rights*, resulted in a positive correlation between these scores. For example, translations from Turkish to English or German consistently resulted in decreased RE and RG, while translations from English and German to Turkish consistently resulted in increased RE and RG. This indicates that the exact text became less difficult to read according to one metric, and at the same time, it became more difficult to read according to another metric. When this finding is combined with the concerns and conclusions stated by Kalyoncu and Memiş, it raises concerns about the validity and conformity of these formulas.

Limitations

This study has certain limitations. Its findings are based on five texts and one dataset, which, despite covering several genres, do not fully represent the diversity or difficulty range of possible text types. Therefore, further research is necessary to investigate broader conditions. Moreover, the analysis examined only four NMT services, and since these systems are continuously updated, the results should be viewed as a temporal snapshot rather than a permanent ranking.

Implications and Future Research

The findings have practical implications for users of NMT services. They highlight that raw MT output, particularly when translating into morphologically complex languages, may result in texts that are significantly more difficult to read than the source or intended target level. This suggests a potential need for post-editing that focuses specifically on enhancing readability, beyond semantic correction. Users should be aware of this systematic shift, especially when utilizing MT for materials that require higher accessibility, such as public communications, patient information, or educational materials for younger pupils.

Future research could expand on these findings by:

- Including a wider variety of genres and text lengths.
- Testing additional language pairs with different typological characteristics.
- Investigating the performance of newer LLM-based translation approaches alongside NMT services.
- Correlating formulaic scores with expert opinions.
- Fine-tuning or training NMT services to produce outputs at the target level of accessibility.
- Further investigating the validity and cross-lingual comparability of readability formulas, especially for less-resourced languages like Turkish.

Conclusion

This study examined readability transfer in NMT with three main objectives: (1) assessing changes in readability scores across English, German, and Turkish, (2) analyzing the impact of translation direction and target language, and (3) comparing the readability of four leading NMT services: Amazon Translate, Azure Translator, DeepL, and Google Cloud Translation. To achieve these objectives, we first collected texts available in English, German, and Turkish, provided through human translation or curated machine translation. Later, using the APIs of the chosen NMT services, we translated these texts bidirectionally into the other two languages. Next, we conducted readability and textual analysis using the SmoothText Python library with Stanza as its backend. Finally, we reported and examined the outputs regarding direction- and NMT-based changes in readability and textual metrics.

The findings revealed that readability scores varied systematically by language and direction, with Turkish translations often exhibiting lower readability due to longer word lengths and more complex syllable structures, while English translations tended to enhance readability. This outcome was largely driven by

differences in word syllable length, reflecting the compounding nature of German and the agglutinative structure of Turkish.

Among the employed NMT services, no service consistently outperformed the others, indicating that readability preservation is not a current design priority in these services. DeepL maintained relatively stable results for German↔English pairs, whereas Google Translate and Amazon Translate showed more variable results. Azure Translator generally lagged behind in sentence-level fluency. Reference translations of the same texts did not show these differences; the asymmetries reported here highlight a specific limitation of NMT systems in conveying readability and show that they do not effectively address typological effects on readability. This ineffectiveness ultimately leads to discrepancies between the measured difficulties of source texts and their translations.

Importantly, the analysis also pointed out limitations of readability formulas, especially the Bezirci–Yılmaz measure for Turkish, which tended to inflate grade-level difficulty scores. This aligns with concerns in the existing literature and shows that relying solely on textual metrics may not yield reliable readability scores. However, as discussed above, these formulaic limitations are not the main cause of NMT incapability, but rather indicators of it.

These findings suggest that raw NMT output might not meet accessibility standards, particularly when translating into morphologically complex languages. Therefore, post-editing should focus on both semantic accuracy and readability.

Overall, this study highlights readability transfer as a crucial yet underexplored aspect of translation quality, suggesting that future research should include a broader range of texts and systems to develop a more comprehensive understanding of this issue, while also correlating formula-based scores with human judgments.

Peer Review Externally peer-reviewed.
Conflict of Interest The author has no conflict of interest to declare.
Grant Support The author declared that this study has received no financial support.

Hakem Değerlendirmesi Dış bağımsız.
Çıkar Çatışması Yazar çıkar çatışması bildirmemiştir.
Finansal Destek Yazar bu çalışma için finansal destek almadığını beyan etmiştir.

Author Details / Yazar Bilgileri
Tuğrul Güngör (Lecturer)
[1] Istanbul Gelisim University, Istanbul Gelisim Vocational School/Foreign Languages and Cultures Department, Istanbul, Türkiye
0000-0002-6945-417X tgungor@gelisim.edu.tr

References

Amstad, T. (1978). Wie verständlich sind unsere Zeitungen? [How understandable are our newspapers?]. Studenten-Schreib-Service.

Ateşman, E. (1997). Türkçede okunabilirliğin ölçülmesi [Measuring readability in Turkish]. *Dil Dergisi*, *58*, 71–74.

Ay, İ. E., & Duranoğlu, Y. (2022). Göz damlası prospektüslerinin okunabilirlik düzeyinin değerlendirilmesi [An evaluation of the readability of package inserts of eye drops]. *Anadolu Kliniği Tıp Bilimleri Dergisi*, *27*(1), 55–59. https://doi.org/10.21673/ANADOLUKLIN.993863

Bamberger, R., & Vanecek, E. (1984). Lesen-Verstehen-Lernen-Schreiben: Die Schwierigkeitsstufen von Texten in deutscher Sprache [Reading–understanding–learning–spelling: The degrees of difficulty of texts in German language]. Jugend und Volk.

Bezirci, B., & Yılmaz, A. E. (2010). Metinlerin okunabilirliğinin ölçülmesi üzerine bir yazilim kütüphanesi ve Türkçe için yeni bir okunabilirlik ölçütü [A software library for measurement of readability of texts and a new readability metric for Turkish]. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, *12*(3), 49–62.

Çıplak, G., & Balcı, A. (2022). Türkçe ders kitaplarındaki metinlerin okunabilirlik özelliğinin incelenmesi [Examining the readability of texts in Turkish textbooks]. *RumeliDE Dil ve Edebiyat Araştırmaları Dergisi*, *30*, 170–187. https://doi.org/10.29000/RUMELIDE.1193033

Cohen, S. A., Brant, A., Fisher, A. C., Pershing, S., Do, D., & Pan, C. (2024). Dr. Google vs. Dr. ChatGPT: Exploring the use of artificial intelligence in ophthalmology by comparing the accuracy, safety, and readability of responses to frequently asked patient questions regarding cataracts and cataract surgery. *Seminars in Ophthalmology*, *39*(6), 472–479. https://doi.org/10.1080/08820538.2024.2326058

Collins-Thompson, K. (2014). Computational assessment of text readability. *ITL - International Journal of Applied Linguistics*, *165*(2), 97–135. https://doi.org/10.1075/itl.165.2.01col

Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language*, *23*(1), 84–101.

Dubay, W. H. (2004). *The Principles of Readability*.

Erol, H. F. (2015). Yabancı dil olarak Türkçe ders kitaplarında okunabilirlik [Readability in Turkish as a foreign language textbooks]. *Türk Dili ve Edebiyatı Dergisi*, *50*(50), 29–38. https://dergipark.org.tr/tr/pub/iutded/issue/17078/178709

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. https://doi.org/10.1037/H0057532

François, T. (2014). An analysis of a French as a Foreign language corpus for readability assessment. *Proceedings of the Third Workshop on NLP for Computer-Assisted Language Learning*, *107*, 13–32.

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

Gondode, P., Duggal, S., Garg, N., Lohakare, P., Jakhar, J., Bharti, S., & Dewangan, S. (2024). Comparative analysis of accuracy, readability, sentiment, and actionability: Artificial intelligence chatbots (ChatGPT and Google Gemini) versus traditional patient information leaflets for local anesthesia in eye surgery. *The British and Irish Orthoptic Journal*, *20*(1), 183–192. https://doi.org/10.22599/BIOJ.377

Gosselin, A. M., le Maux, J., & Smaili, N. (2021). Readability of accounting disclosures: A comprehensive review and research agenda. *Accounting Perspectives*, *20*(4), 543–581. https://doi.org/10.1111/1911-3838.12275

Hancı, V., Ergün, B., Gül, Ş., Uzun, Ö., Erdemir, İ., & Hancı, F. B. (2024). Assessment of readability, reliability, and quality of ChatGPT®, BARD®, Gemini®, Copilot®, Perplexity® responses on palliative care. *Medicine*, *103*(33), e39305. https://doi.org/10.1097/MD.0000000000039305

Intento. (2021). The State of Machine Translation 2021. Intento Inc. Retrieved from https://try.inten.to/machine-translation-report-2021/

Işım, Ç., & Balcıoğlu, Y. S. (2023). ChatGPT: Performance of translate. 3rd International ACHARAKA Congress on Humanities and Social Sciences Proceedings Book, 47–51.

Jiao, W., Wang, W., Huang, J., Wang, X., Shi, S., & Tu, Z. (2023). *Is ChatGPT a good translator? Yes with GPT-4 as the engine*. https://docs.google.com/document/d/1GeFh6I5OrMHMI-iMFUz4e2hXhljl3xJhncFHeoAJAkg/edit?tab=t.0&usp=embed_facebook

Kalyoncu, M. R., & Memiş, M. (2024). Türkçe için oluşturulmuş okunabilirlik formüllerinin karşılaştırılması ve tutarlılık sorgusu [Consistency query and comparison of readability formulas created for Turkish]. *Ana Dili Eğitimi Dergisi*, *12*(2), 417–436. www.anadiliegitimi.com

Kincaid, J. P., Fishburne Jr., R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for navy enlisted personnel. http://library.ucf.edu

Loughran, T., & Mcdonald, B. (2014). Measuring readability in financial disclosures. *The Journal of Finance*, *69*(4), 1643–1671. https://doi.org/10.1111/JOFI.12162

Meyer, M. K. R., Kandathil, C. K., Davis, S. J., Durairaj, K. K., Patel, P. N., Pepper, J.-P., Spataro, E. A., & Most, S. P. (2024). Evaluation of rhinoplasty information from ChatGPT, Gemini, and Claude for readability and accuracy. *Aesthetic Plastic Surgery*, *49*(7), 1868–1873. https://doi.org/10.1007/S00266-024-04343-0/TABLES/3

Momenaei, B., Wakabayashi, T., Shahlaee, A., Durrani, A. F., Pandit, S. A., Wang, K., Mansour, H. A., Abishek, R. M., Xu, D., Sridhar, J., Yonekawa, Y., & Kuriyan, A. E. (2023). Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. *Ophthalmology Retina*, *7*, 862–868. https://doi.org/10.1016/j.oret.2023.05.022

Nygård, A. (2024). Testing teachers' abilities to distinguish between translations produced by students, Google Translate and ChatGPT. https://doi.org/https://hdl.handle.net/11250/3148912

Oliffe, M., Thompson, E., Johnston, J., Freeman, D., Bagga, H., & Wong, P. K. K. (2019). Assessing the readability and patient comprehension of rheumatology medicine information sheets: a cross-sectional health literacy study. *BMJ Open*, *9*, :e024582. https://doi.org/10.1136/bmjopen-2018-024582

Ozduran, E., Hancı, V., Erkin, Y., Özbek, İ. C., & Abdulkerimov, V. (2025). Assessing the readability, quality and reliability of responses produced by ChatGPT, Gemini, and Perplexity regarding most frequently asked keywords about low back pain. *PeerJ*, *13*, e18847. https://doi.org/10.7717/peerj.18847

Pérez-Ortiz, J. A., Forcada, M. L., & Sánchez-Martínez, F. (2022). How neural machine translation works. In D. Kenny (Ed.), *Machine translation for everyone: Empowering users in the age of artificial intelligence* (pp. 141–164). Language Science Press. https://doi.org/10.5281/zenodo.6760020

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://spacy.io/

Son, J., & Kim, B. (2023). Translation Performance from the User's Perspective of Large Language Models and Neural Machine Translation Systems. *Information*, *14*(10), 574. https://doi.org/10.3390/INFO14100574

Stasimioti, M., Sosoni, V., Kermanidis, K. L., & Mouratidis, D. (2020). Machine translation quality: A comparative evaluation of SMT, NMT and tailored-NMT outputs. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 441–450. https://aclanthology.org/2020.eamt-1.47/

Strzalkowski, P., Strzalkowska, A., Chhablani, J., Pfau, K., Errera, M. H., Roth, M., Schaub, F., Bechrakis, N. E., Hoerauf, H., Reiter, C., Schuster, A. K., Geerling, G., & Guthoff, R. (2024). Evaluation of the accuracy and readability of ChatGPT-4 and Google Gemini in providing information on retinal detachment: a multicenter expert comparative study. *International Journal of Retina and Vitreous*, *10*(1), 1–11. https://doi.org/10.1186/S40942-024-00579-9/FIGURES/2

Toral, A., & Way, A. (2018). *What level of quality can neural machine translation attain on literary text?* 263–287. https://doi.org/10.1007/978-3-319-91241-7_12

Wang, H., Wu, H., He, Z., Huang, L., & Church, K. W. (2022). Progress in machine translation. *Engineering*, *18*, 143–153. https://doi.org/10.1016/J.ENG.2021.03.023

Wu, Y., Schuster, M., Chen, Z., Le, Q. v, Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). *Google's neural machine translation system: Bridging the gap between human and machine translation*.

Yirmibeşoğlu, Z., Dursun, O., Dallı, H., Şahin, M., Hodzik, E., Gürses, S., & Güngör, T. (2023). *Incorporating human translator style into English-Turkish literary machine translation*.

Yılmaz, F., & Temiz, Ç. (2014). Yabancılara Türkçe öğretiminde kullanılan ders kitaplarındaki metinlerin okunabilirlik durumları [Readability conditions of texts in textbooks used in teaching Turkish to foreigners]. *International Journal of Languages' Education and Teaching*, *2*(1), 81–91. https://dergipark.org.tr/tr/pub/ijlet/issue/82527/1417234

Zhou, S., Jeong, H., & Green, P. A. (2017). How consistent are the best-known readability equations in estimating the readability of design standards? *IEEE Transactions on Professional Communication*, *60*(1), 97–111. https://doi.org/10.1109/TPC.2016.2635720