The State of Machine Translation for Websites

A Comparative Study of the Top Machine Translation Engines

WEGLOT Nimdzi

July 2022

INFORMATION CONTAINED IN THIS REPORT

Introduction

Purpose

The purpose of this study is to evaluate and compare a selection of leading machine translation technologies.

What was done

We completed a series of human reviews on five of the leading machine translation (MT) technologies across seven language pairs with a focus on usability, and evaluating performance with the goal of articulating the practical business impact that MT has on translating corporate website content.

Scope of work and methodology

Editors with user hats on

The resources who performed the sample reviews were professional linguists. The role of these linguists was primarily to represent the user who also is a language professional. We call these resources **editors**. The editors were not predisposed in any way for or against machine-translated content and were used to working with it. They have revised, edited, and commented on the target content segments where applicable. The editors were not asked to directly compare machine and human translation or to perform a service akin to LQA, but rather to evaluate the criteria of the **usability** of the machine translation as well as the accuracy and reliability. The main principle and perspective was the following:

If it's not broken, don't fix it. Focus on what's usable.

The sample

The website content sample contained extracted localizable content from several web pages of a service company's corporate website. It combined headlines, body text, calls to action, and alt texts. The size was 168 segments and 1061 source words. Here are some additional data about the sample:

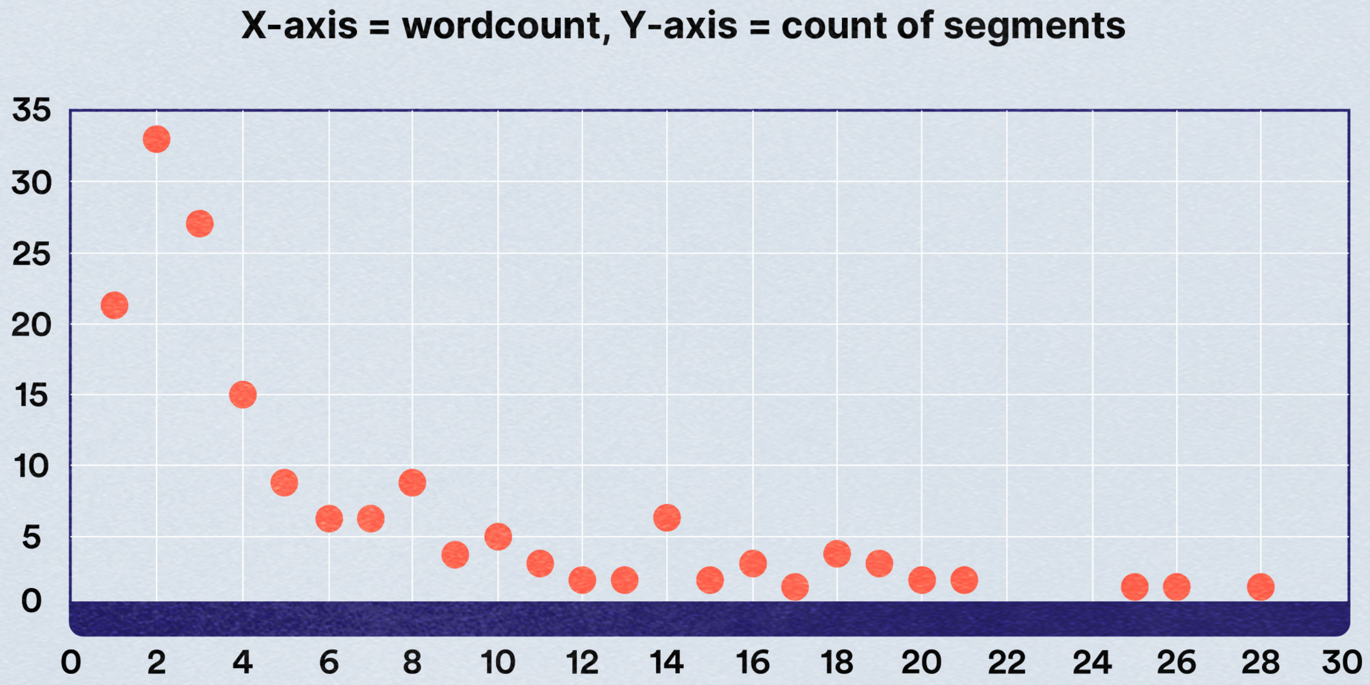


Figure 1: Chart showing the count of segments with a specific word count. The highest number of segments in the sample have two words each.

The scope of review: languages, engines, and editors

The review job consisted of 68 reviews broken down into the review of 5 engines across 7 language pairs with 2 editors per language pair. In other words, for each language there were 5 iterations of the same sample done by 2 different people in parallel.

- MT engines: Amazon Translate, DeepL, Google Cloud, Microsoft Translator, ModernMT[1]
- The source language: en-US
- The target languages: fr-FR, de-DE, es-ES, it-IT, zh-CN, ar-EG, pt-PT

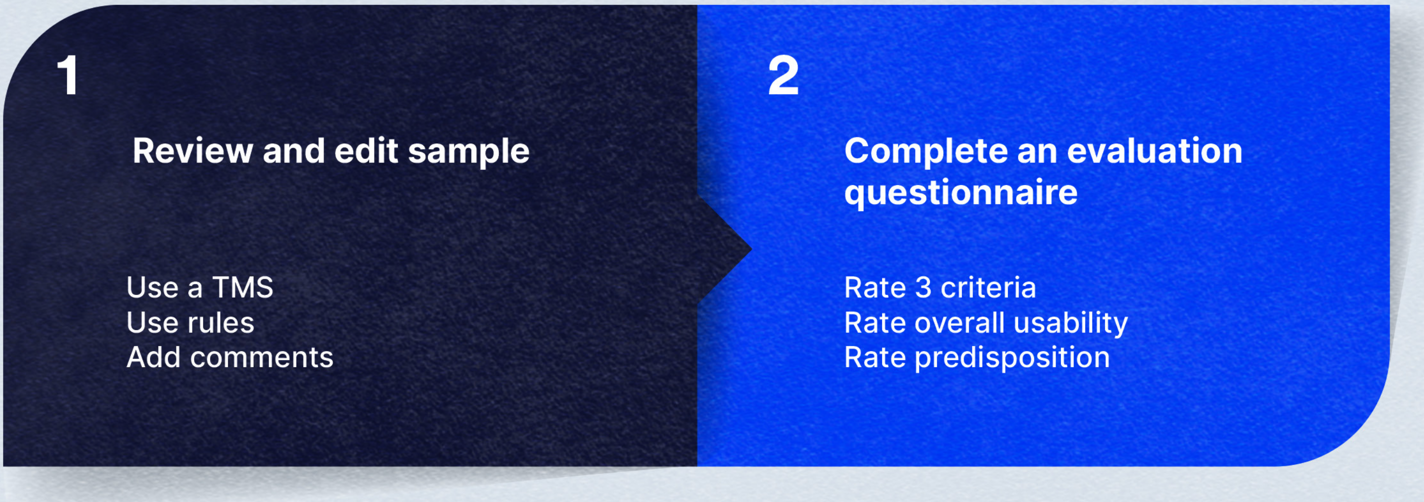
[1] Of the technologies listed, ModernMT is an adaptive machine translation system, but we have not used that capability to get comparable results. The sample was pre-translated using MT, which means the adaptive behavior is not utilized.

Row Labels	MT sample 1	MT sample 2	MT sample 3	MT sample 4	MT sample 5	Grand Total
Editor 1 ar-EG	1		1	1	1	4
Editor 1 de-DE	1	1	1	1	1	5
Editor 1 es-ES	1	1	1	1	1	5
Editor 1 fr-FR	1	1	1	1	1	5
Editor 1 it-IT	1	1	1	1	1	5
Editor 1 pt-PT	1	1	1	1	1	5
Editor 1 zh-CN	1	1	1	1	1	5
Editor 2 ar-EG	1		1	1	1	4
Editor 2 de-DE	1	1	1	1	1	5
Editor 2 es-ES	1	1	1	1	1	5
Editor 2 fr-FR	1	1	1	1	1	5
Editor 2 it-IT	1	1	1	1	1	5
Editor 2 pt-PT	1	1	1	1	1	5
Editor 2 zh-CN	1	1	1	1	1	5
Grand Total	14	12	14	14	14	68

Figure 2: The scope: Editors and engines were anonymized. One of the MT engines does not offer Arabic, so 2 samples were not translated.

Review assignment, process, method, and definitions

The review process consisted of two main steps, editing review and evaluation. In the project schedule, we allowed space between sample reviews so the editors were not influenced by the preceding task.



1

Review and edit sample

Use a TMS
Use rules
Add comments

2

Complete an evaluation questionnaire

Rate 3 criteria
Rate overall usability
Rate predisposition

Figure 3: The review process and high-level guidelines for sample editing

Step 1: Review and edit

The editors were asked to follow these rules to distinguish USABLE (no touch), TOUCHED, and REWORKED segments:

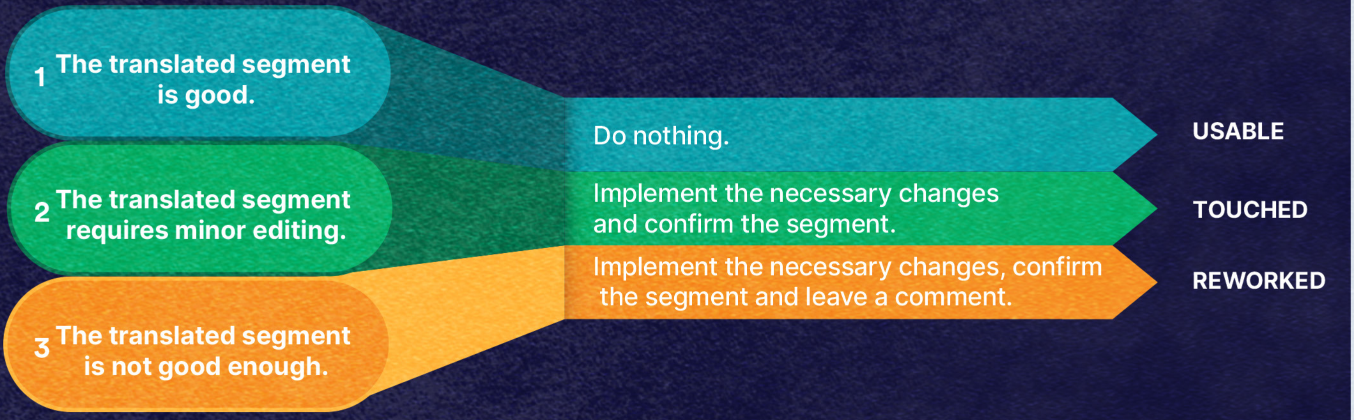

1 The translated segment is good.
2 The translated segment requires minor editing.
3 The translated segment is not good enough.
Do nothing.
Implement the necessary changes and confirm the segment.
Implement the necessary changes, confirm the segment and leave a comment.
USABLE
TOUCHED
REWORKED

Figure 4: Editing guidelines

Step 2: Evaluate

Questions

The editors answered the following questions for each sample they completed in a questionnaire:

1. Please rate the machine translation's **ACCURACY**.
2. Please rate the machine translation's **RELIABILITY**.
3. Was the target **CONTENT DESIGN** ok?
4. How would you evaluate the machine-translated sample (**OVERALL USABILITY)**?

We also asked each editor (only once during the whole project) the following question:

- Rate the extent to which you were positively surprised by the quality of the translation presented to you. This is explained and evaluated in the Predisposition section.

Concept definitions

For the sake of clarity, we have defined the terms as follows:

1. **ACCURACY** rating: How well the target conveys informational content of the source text.
2. **RELIABILITY** rating: Quality of translation when facing issues such as structural ambiguity, idiomatic expressions, syntactic multiplicity, and lexical ambiguity.
3. **CONTENT DESIGN**: Formatting, markup, locale standards, tags, etc.
4. **OVERALL USABILITY** rating: A question asked at the end of the survey, so editors give one overall rating to the sample quality from the perspective of how usable it is.

Rating scale

We used this rating scale for questions 1, 2, and 4:

- **Very good**: If the machine-translated sample was used as is (no editing), users would most likely use it (even if they may feel that this is MT).
- **Acceptable**: Some segments needed editing, but these corrections were mostly light adjustments.
- **Poor**: Most segments in the sample required editing and some had to be translated from scratch.
- **Very bad**: In most cases the machine translation was not useful. Deleting the MT output and rewriting the translation from scratch was more effective than editing.

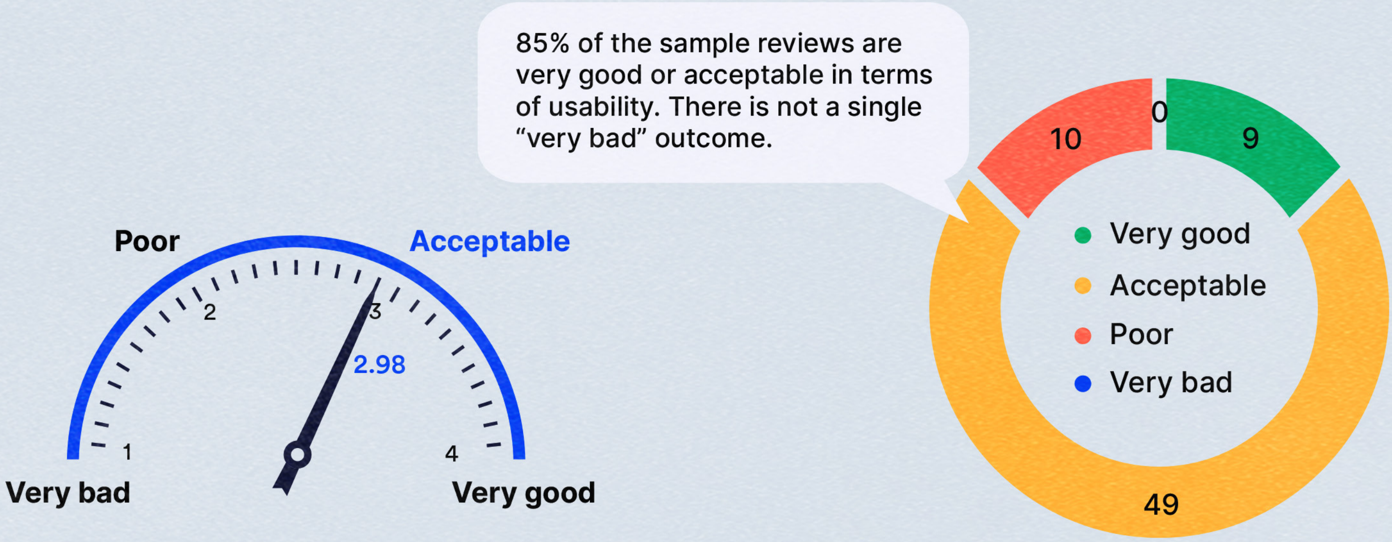
Principles and constraints in this report

- Stock integrations of the 5 NMT technologies were used through a translation management system. No dictionaries or additional AI, which some of the engines utilize, were used.
- We've adopted a user-centered approach. While evaluating accuracy and reliability, we did not rely to any great extent on quality parameters that are typically provided in an LQA report, but more on the content experience and usability in the given context. The website was provided to the editors for full reference.
- The perspective of solely measuring MT against human translation is no longer relevant.
- Edit distance is only a complementary measure and not the main focus.
- Since each editor was asked to review 5 MT translations of the same source, we included waiting time in the review process to reduce potential bias.

Insights and findings

Overall usability from different perspectives

We have examined a number of MT providers and mapped out the space to see how stock NMT technologies perform in the scope described above. The following charts illustrate that the engines tested show high overall usability for the company’s website content.



85% of the sample reviews are very good or acceptable in terms of usability. There is not a single “very bad” outcome.

Figure 5: Average overall rating of samples and sum of reviews by overall usability (68 human reviews across all languages and engines)

Average overall usability rating per language and engine

	ar-EG	de-DE	es-ES	fr-FR	it-IT	pt-PT	zh-CN
Amazon Translate	2.5	3.5	3.5	3	2.5	3	3
DeepL	-	3.5	4	3	2.5	3.5	2.5
Google Cloud Translation	3.5	4	2.5	2.5	2.5	2.5	3
Microsoft Translator	3	3	3	3	3	3	3
ModernMT	3	3	3	3	2.5	3	2.5

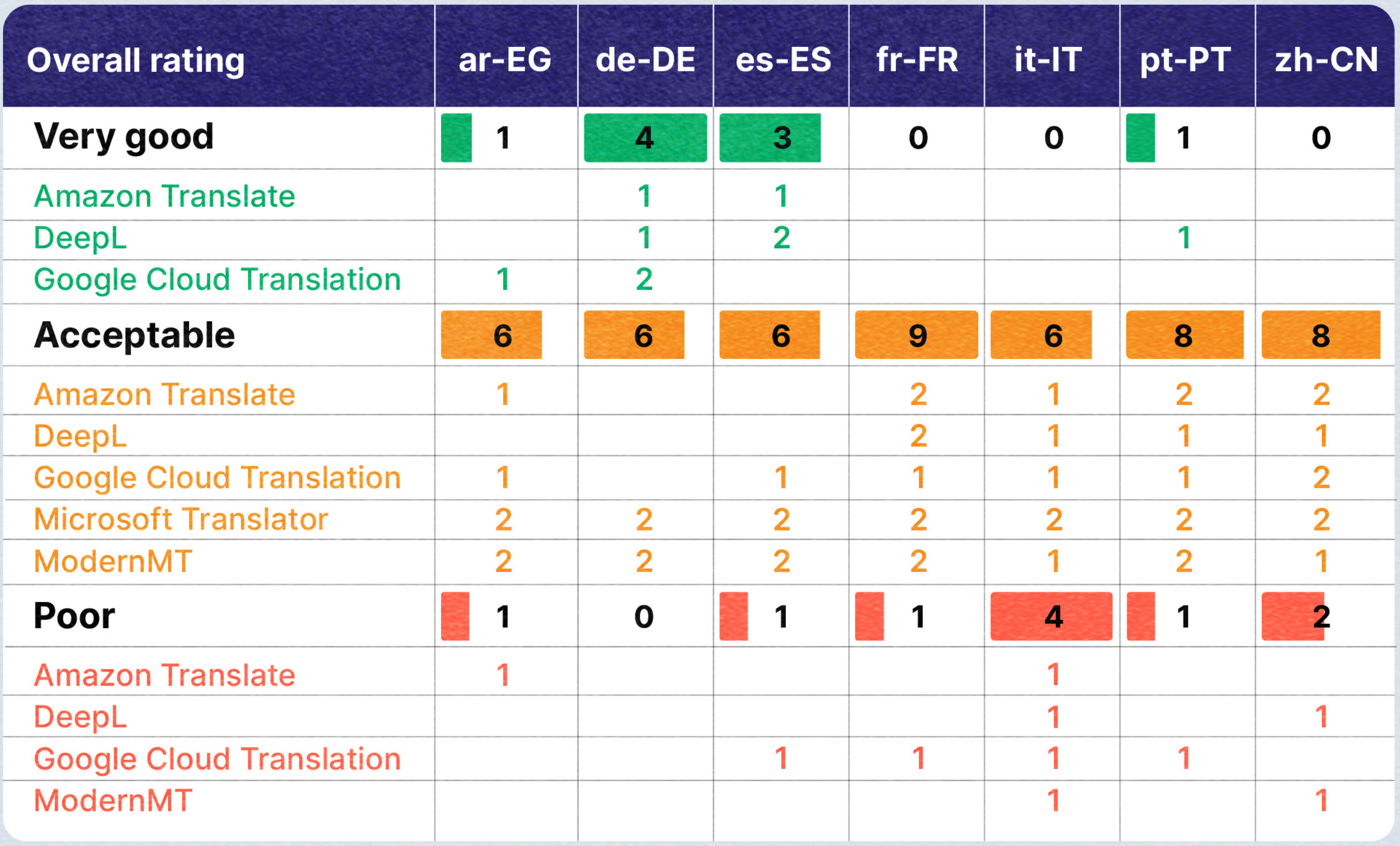
Rating scale

Very good	4	Acceptable	3	Poor	2	Very bad	1

Figure 6: Each cell of the table shows the average score of 2 sample reviews performed by 2 editors.

Ratings distribution

Here is an additional perspective on the ratings. French, Portuguese, Chinese, and Arabic appear to be balanced across the engines (having the fewest outliers). German and Spanish score high in very good ratings (highly usable), while Italian, across all engines, has a relatively high number of scores that are poor, but come from one single editor.

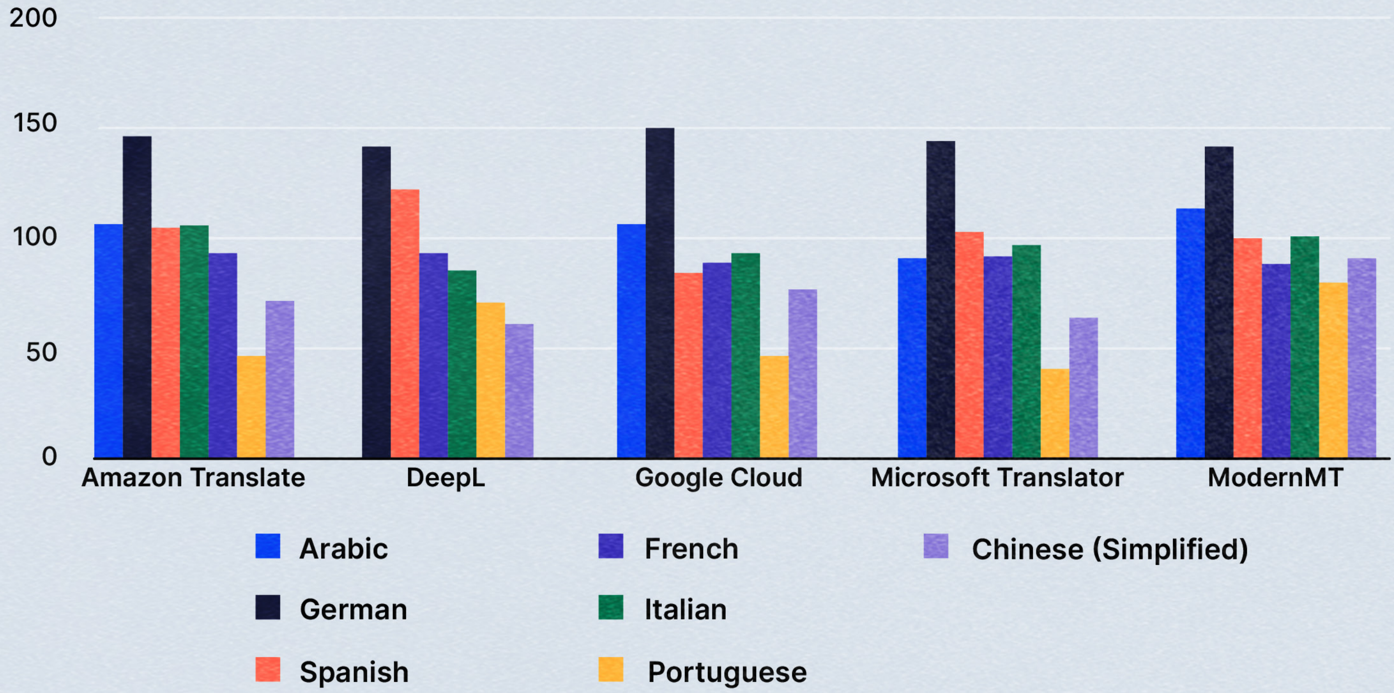


Overall rating	ar-EG	de-DE	es-ES	fr-FR	it-IT	pt-PT	zh-CN
Very good	1	4	3	0	0	1	0
Amazon Translate		1	1				
DeepL		1	2			1	
Google Cloud Translation	1	2					
Acceptable	6	6	6	9	6	8	8
Amazon Translate	1			2	1	2	2
DeepL				2	1	1	1
Google Cloud Translation	1		1	1	1	1	2
Microsoft Translator	2	2	2	2	2	2	2
ModernMT	2	2	2	2	1	2	1
Poor	1	0	1	1	4	1	2
Amazon Translate	1				1		
DeepL					1		1
Google Cloud Translation			1	1	1	1	
ModernMT					1		1

Figure 7: Ratings matrix

No-touch translations

Each sample, after editing, produced a number of segments that required no editing. The following chart illustrates for each language and engine how many of the 168 segments contained in the sample were not touched (i.e. were used as they were). For German, it is about 145 segments, while for Portuguese, the median value is 58. Also, it is notable that ModernMT has the highest relative count of no touch segments plus the lowest variation when compared with the other engines.



de-DE no touch segments count relatively the highest, while pt-PT are relatively the lowest.

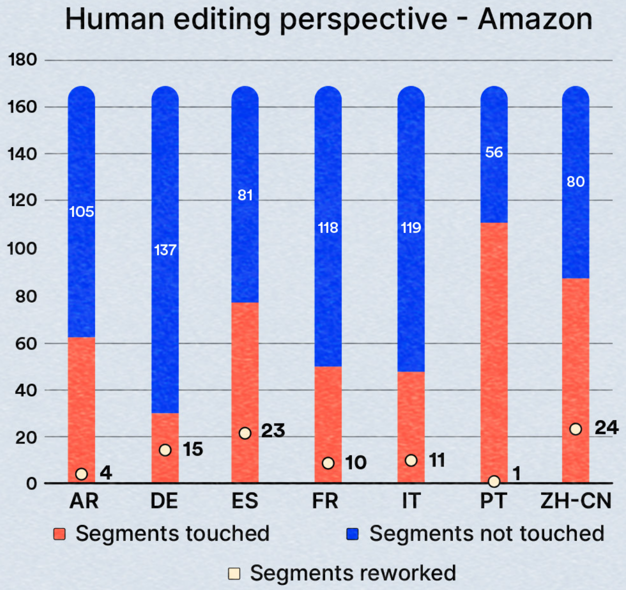
Figure 8: Count of machine-translated segments in each sample that required no editing

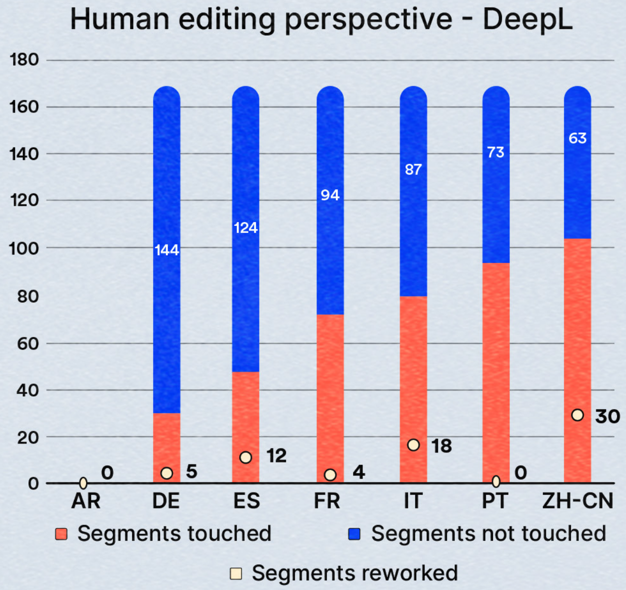
Accuracy and reliability

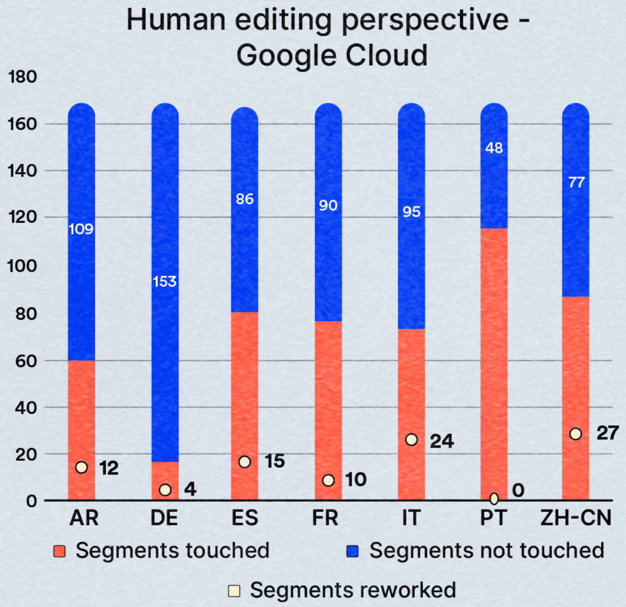
We have focused on evaluating and comparing the overall usability rating. Although accuracy and reliability criteria were also assessed, the data did not show any significant differences or outliers there.

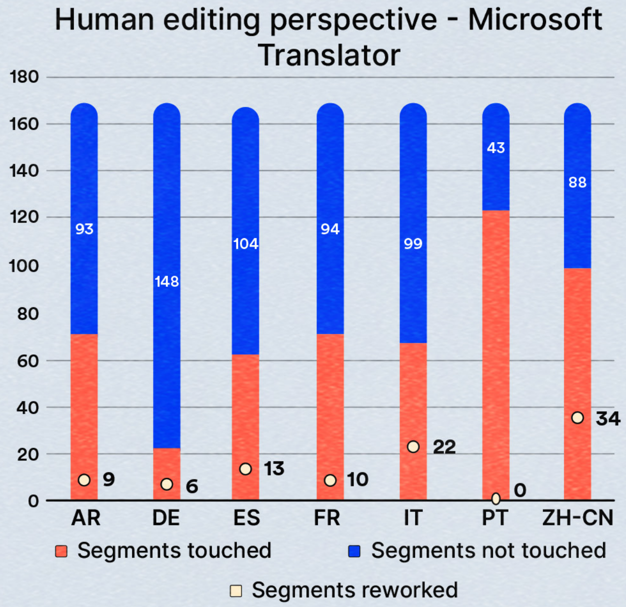
Human editing chart

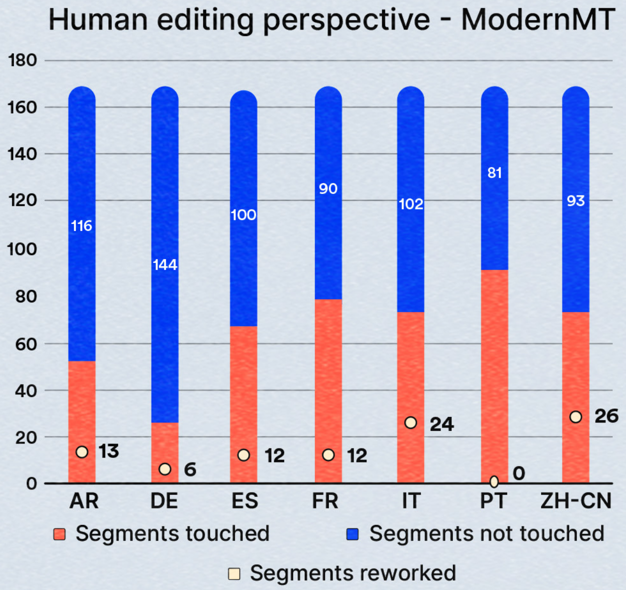
In each sample, there were segments that were great, segments that had to be light-edited, and segments that had to be reworked. Below are the charts for each MT engine.











Across the engines, Portuguese has the lowest number of segments that required reworking, while Chinese has the highest number. Similarly, German has the highest number of great translations, whereas Portuguese counts the lowest in this category.

Figure 9: Number of segments not touched, touched, and reworked (a subset of the touched segments)

Predisposition

People's preconceptions and trust of algorithms may influence their evaluations of an algorithm's effectiveness. To learn about any prior editor expectations regarding the machine-translated content, we also asked each editor (only once during the whole project) to rate the extent to which they were positively surprised by the quality of the MT.
As the table below shows, 10 out of 14 editors were on the side of being positively surprised by the quality of the translation presented to them. This illustrates that for most of them, the MT output was of better quality than they expected. As other studies have demonstrated, lower evaluations of an MT product lead to a stronger desire to intervene in the translation by introducing changes to the original message.

	Negatively surprised					Positively surprised
Rating Scale	1	2	3	4	5	6
Editor 1 ar-EG				1		
Editor 2 ar-EG			1			
Editor 1 de-DE					1	
Editor 2 de-DE						1
Editor 1 es-ES			1			
Editor 2 es-ES				1		
Editor 1 fr-FR		1				
Editor 2 fr-FR				1		
Editor 1 it-IT					1	
Editor 2 it-IT				1		
Editor 1 pt-PT					1	
Editor 2 pt-PT					1	
Editor 1 zh-CN				1		
Editor 2 zh-CN			1			

Figure 10: Disconfirmation report

Debunking MT myths

Companies who want to implement an MT strategy are often reluctant due to the prejudices that have arisen over the years in the localization industry. Google Translate has never had the best reputation, people think that machine translation doesn't work for marketing content, and "exotic" languages like Arabic and Chinese are often avoided because of the allegedly lower translation quality.

The data presented in this study debunks some of the persisting myths and confirms that machine translation can be deployed as an accelerator in the translation process, even for marketing content. Furthermore, the question that kept the industry up at night for quite some time – "Will machine translation ever replace human translators?" – seems to have disappeared. Machine translation has reached a certain level of maturity and people realize that it has made it possible to translate high volumes of content that would otherwise not be translated because of budget and time constraints, thus complementing (rather than replacing) the human linguist. Today, it is not uncommon for a single company to translate millions of words of web content using stock NMT.

Editing aspects

In this section, we'll use the data that was measured in the MT assessment and present how much effort the editing represented in an objective manner. We'll work towards two conclusions:

1. MT as a technology is maturing and the results are good.
2. Nonetheless, MT as a service still has a lot of maturing to do. We can't indicate whether the editing effort was high or low because our industry doesn't have parameters for it. How can we support the statement that MT works as an accelerator? We need more standardization in terms of quality, as the traditional metrics (edit distance, BLEU score) are not always enough.

Recurring mistakes

Here is an overview of the most frequently occurring mistakes that were reported by the editors in their evaluation forms.

Context

It was reported repeatedly by the editors that certain text fragments were translated out of context. This actually speaks in favor of machine translation because context is just as important to machines as it is to human linguists. Human translators are usually provided with all kinds of reference material allowing them to understand the context of a given source text fragment. If you ask a translator to translate a short sentence without any context, they might make mistakes too. This is what some of the editors said about context:

If the MT does not know a term in its context, it will provide a general translation, which does not fit the context. - **Editor 2, FR, ModernMT**

Overall, the information of the source text is conveyed in the target texts. There were few cases in which a mistranslation has distorted the meaning of the sentence as it referred to another context. - **Editor 1, IT, DeepL**

Overall, the translation was accurate when there was enough context. When the context was missing, the translation was incorrect. - **Editor 2, FR, Google Cloud**

In stock machine translation, context is gained by the amount of content that is sent to an engine for translation. The more input, the better the output. This implies a disruption for translation memories and segmentation rules in TMS. Sentence-based segmentation takes away vital information, which might result in low-quality translation output.

The “&” character (ampersand)

In general, the editors agreed that the MT engines manage to transfer information accurately from the source to the target. Nevertheless, we identified a recurring semantic problem caused by one specific character: the ampersand (&). It was reported more than 10 times by the editors that the MT engines have a hard time processing the character appropriately.

Overall, the MT delivers accurate translations but in some instances, the target sentence does not make sense. For example, if a source sentence contains ‘&’, the MT doesn’t succeed in conveying the same meaning. - **Editor 2, FR, ModernMT**

Example:

Source:

[Brand]'s state-of-the-art automated warehouses in France, Spain, and the UK enable ultra-fast order preparation & shipping to delight your customers and allow you to conquer Europe.

MT:
Les entrepôts automatisés à la pointe de la technologie de [Brand] en France, en Espagne et au Royaume-Uni permettent une & expédition ultra-rapide de la préparation des commandes pour ravir vos clients et vous permettre de conquérir l'Europe.

It is clear that the MT engines don't interpret the ampersand character as a semantic equivalent of the word 'and', but it is difficult to determine the exact reason. Maybe the ampersand is not common enough in the target languages, or the MT engines don't manage to process the underlying XML entity (&amp;) correctly. Nonetheless, since the pattern can be found across all language combinations and MT technologies, the effort to fix it shouldn't be too big.

Flavors of Portuguese

The Portuguese editors reported several times that some of the MT engines inject Brazilian terms, despite the fact that the target language was consistently set to European Portuguese. Here are some pieces of feedback provided by the Portuguese editors:

> *Overall quality was good, but as happens with many MT engines, the Portuguese variant used was Brazilian Portuguese instead of European. This means that several corrections regarding terminology, spelling and sentence construction had to be made.* **- Editor 1, Amazon Translate**

> *I have only made some minor changes to improve readability, style, and terminology since some terms were in Brazilian, but overall, the MT translation was rather accurate.* **- Editor 2, ModernMT**

> *There were some Brazilian terms mixed with the Portuguese: "logotipo", "rastreamento", "aplicativo". There was some old Portuguese spelling too: "factura", "activada".* **- Editor 1, DeepL**

We can ask ourselves the question whether MT providers are responsive enough to locale-specific solutions. The need to differentiate between Brazilian and European Portuguese in machine translation is high, as they should be treated as two separate languages in localization. This is likely to apply to other languages such as Spanish.

Other errors

Additional errors that occurred in MT content include missing punctuation and spaces.

Edit distance

For each machine-translated segment, we measured how much the raw machine translation differs (in characters) from its reviewed counterpart. In other words, we computed how much the editors needed to modify the samples to get to the final versions.

We're presenting the results per MT provider per target language as "normalized edit distance statistics," a percentage indicating the distance between the two versions of a sample. 0% means that there are no changes, while 100% means that everything has changed. Note that we assessed the two samples per target language as an entity and calculated average edit distance statistics for them since averages give a better idea of the general performance of the engines.

The results are very consistent across the engines and the target languages. The edit distance is the highest for zh-CN and the lowest for de-DE, which corresponds to the results presented in the Human editing perspective chart.

Since edit distance doesn't track the time spent by the editors, we're only providing it as a complementary metric.

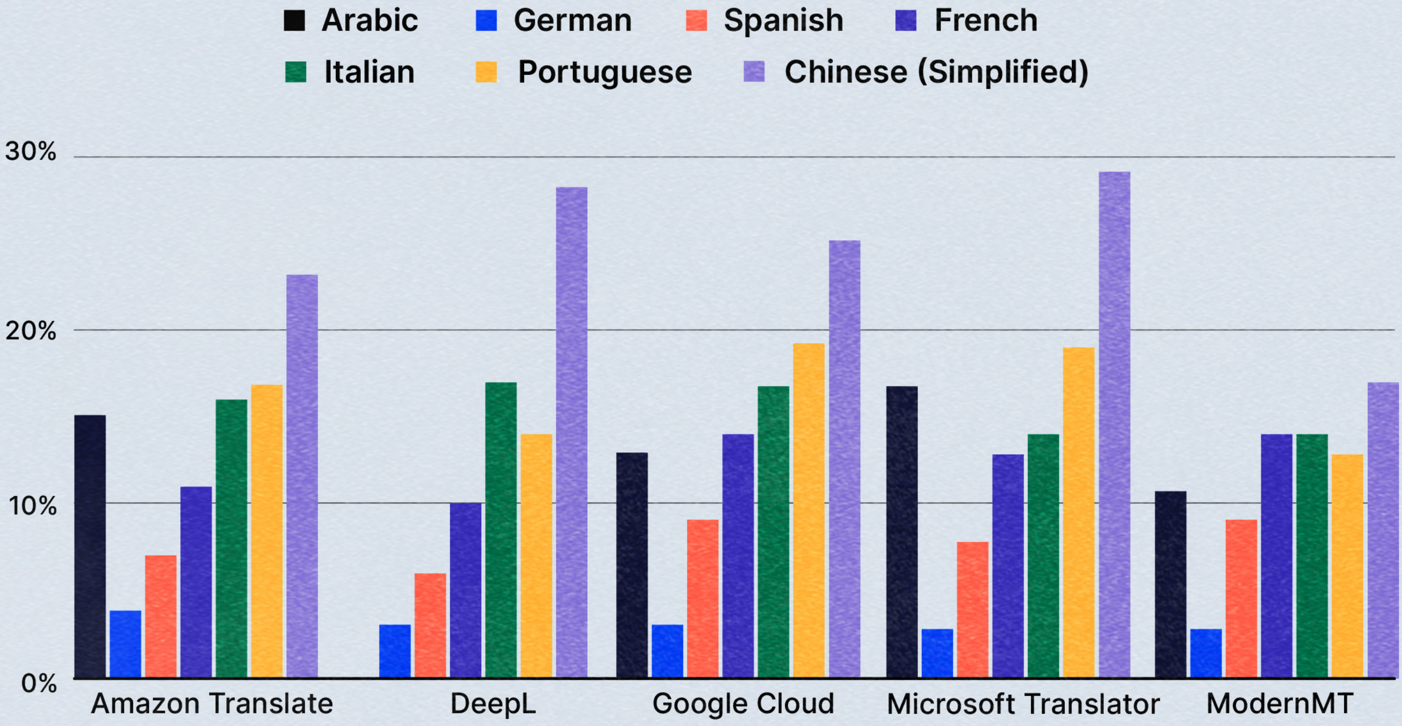


Figure 11: Normalized edit distance statistics

Engine performance highlights

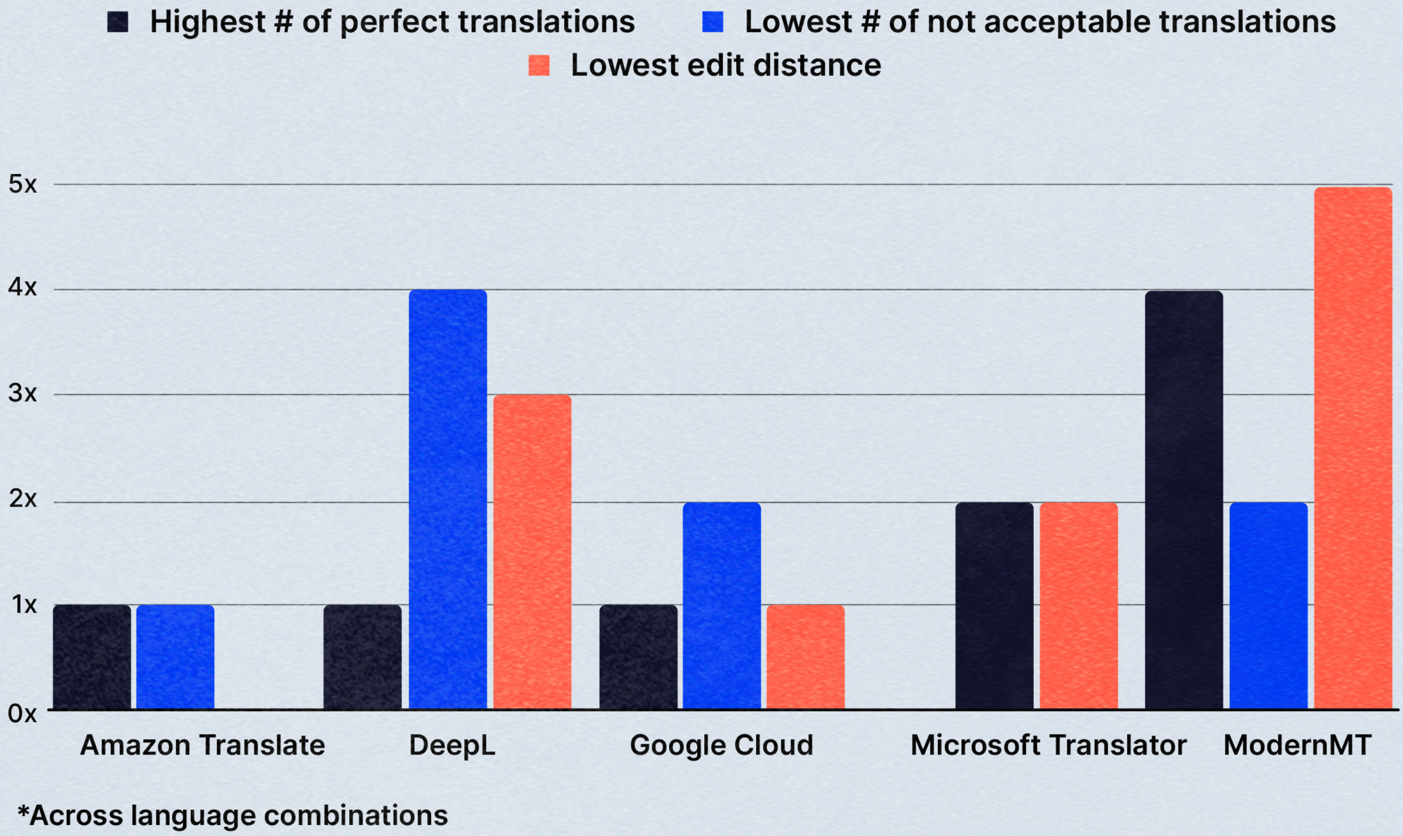


Figure 12: Technology performance comparison in 3 categories

The last thing we analyzed is the performance of the different MT technologies relative to each other. In this part of the study, we chose to focus on objectively measured data, i.e. data one can't argue about. We evaluated three categories: the number of no-touch translations (translations that didn't require post-editing), the number of unacceptable translations (translations that required post-editing to be usable), and edit distance. Just like in Edit distance, we assessed the two samples per language pair as an entity and calculated average statistics from them since averages enable us to take potential conflicting opinions between two reviewers about a sample into account.

The chart indicates that ModernMT produced the highest number of no-touch translations in four samples and the lowest edit distance in five samples, thus outperforming the other MT technologies. DeepL on the other hand had the lowest number of unacceptable translations in four samples. Amazon Translate is lagging behind the other providers, as it generated the highest number of no-touch translations and the lowest number of unacceptable translations in one sample only.

We could say that DeepL and, even more so, ModernMT are the best performing engines, but this statement must nevertheless be further nuanced. While DeepL has the lowest number of unacceptable translations for it-IT, it also has the lowest number of no-touch translations for the same language. Furthermore, Amazon Translate, the least performing engine according to the chart, still outperforms the other providers in two categories: the highest number of no-touch translations in fr-FR, and the lowest number of unacceptable translations in zh-CN. Furthermore, the data in this analysis shows that there's a best performing engine across the evaluation categories for three samples only: Google Cloud for de-DE, DeepL for es-ES, and ModernMT for pt-PT.

Based on this analysis, we can conclude that comparing the baseline quality of different MT engines might become an obsolete practice. Neural machine translation as a technology has reached a certain level of maturity and the market-leading MT providers produce decent stock quality. At the same time, more and more MT programs are based on multi-provider models that combine the best of different engines into one framework. Such frameworks are supported by advanced algorithms that select the best performing engine for a certain piece of content in a given language pair.

Conclusion

State of the MT technologies today

MT has existed since the 1950s, but it was the paradigmatic shift from rules-based MT to statistical MT in the last 20 years, and particularly the recent advent of deep learning, neural machine translation (NMT), that raised its quality level immensely. NMT is a proverbial blackbox that lacks algorithmic transparency. It draws on very large parallel corpora of existing human translations paired with their source text segments (usually at the sentence level), using a deep learning approach to determine probabilities of translation outputs by means of complex recursive neural networks (Asscher, Glikson, 2021).

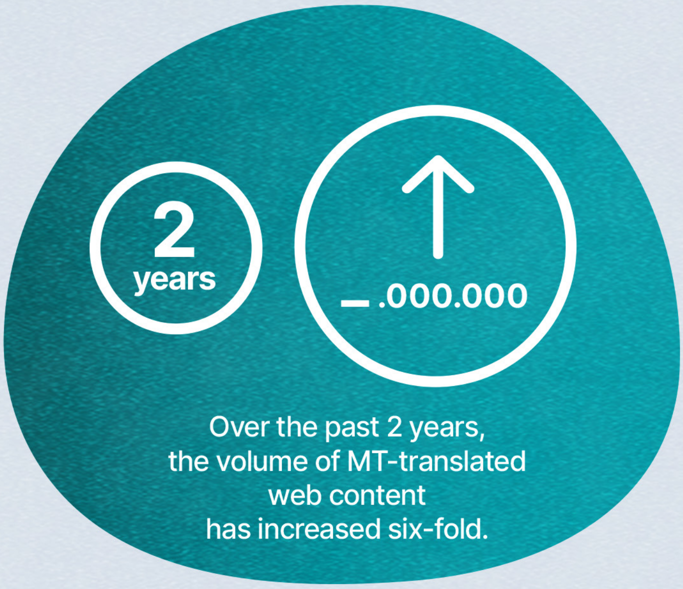
If used with additional AI and, for example, dictionaries to eliminate DNT (do not translate) items and brand/product names, most of the errors encountered would be eliminated. The design and flavor errors are definitely something to address and improve on.

Business impact

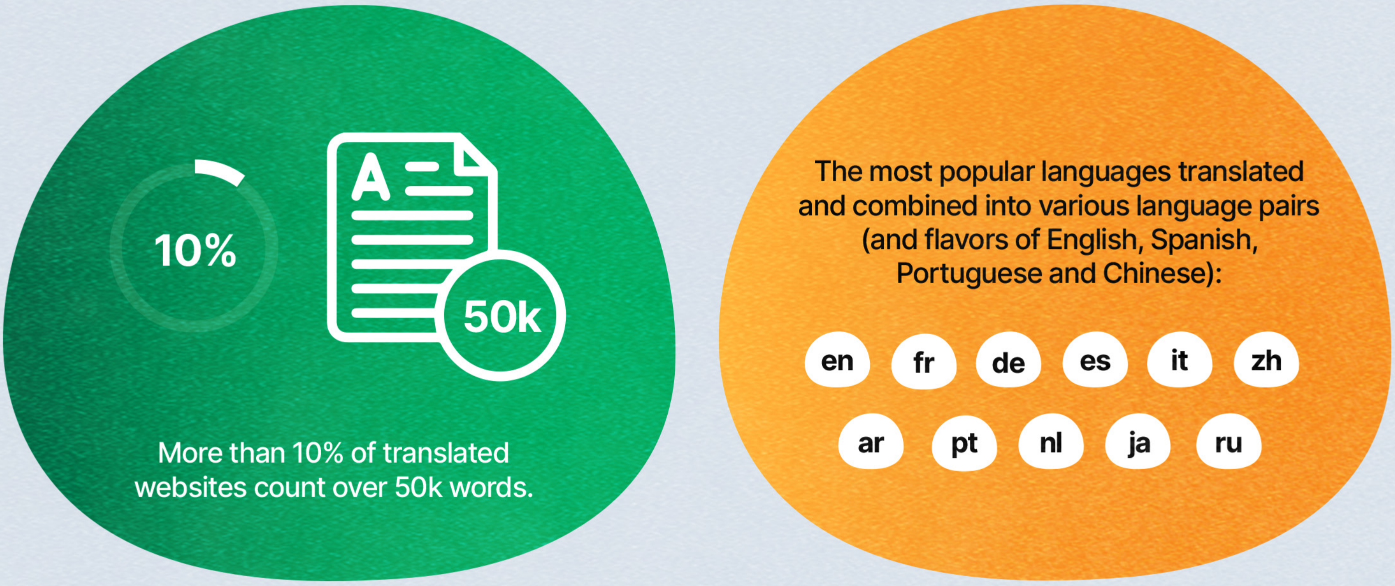
It has been recently estimated that 99% of the translations produced globally are not done by professional human translators. The value of using MT effectively lies in having state-of-the-art engines handy. As this study illustrates, website translations by contemporary NMT are highly usable and require mostly minor editing.

Website translation solution, Weglot, is used by more than 60,000 global brands to translate their websites using a mix of machine translation and post-editing. It chooses the most suitable MT engine for a given language pair based on the most accurate outcome. The data provided by Weglot gives substance to the results provided in this study.

On average,
30% of machine-translated
content is edited.



Over the past 2 years,
the volume of MT-translated
web content
has increased six-fold.



10%

50k

More than 10% of translated websites count over 50k words.

The most popular languages translated and combined into various language pairs (and flavors of English, Spanish, Portuguese and Chinese):

en fr de es it zh

ar pt nl ja ru

Figure 13: Market data provided by Weglot

Inspiration for your business

If you intend to grow a successful, highly forward-looking organization and use machine-translated content effectively, start your next strategic initiative using these revealing insights to challenge your assumptions.

Expand
MT is fit for translating marketing content like company websites

Optimize
There is no "winner MT", there are advanced choices for best business impact.

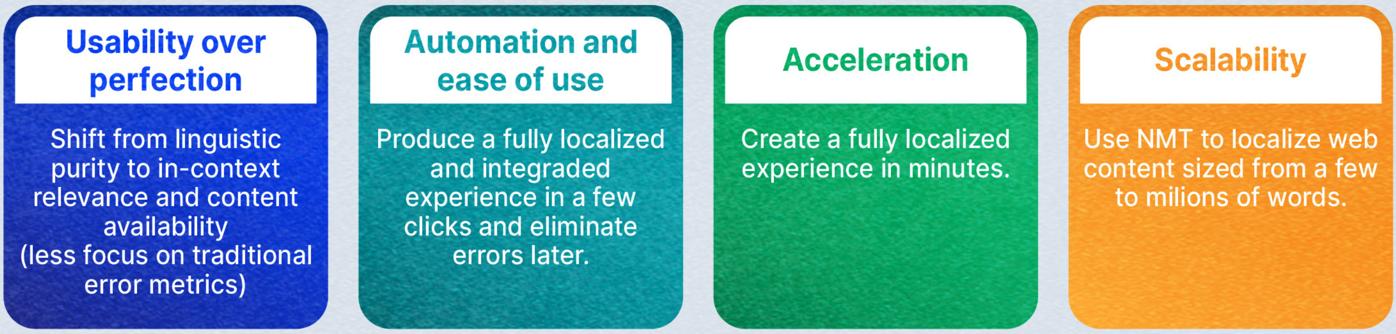
Challenge
To overcome perceptions, managers need to understand true MT capabilities.

Abandon
A localization model that relies on post-editing as a process step is old news.

Figure 14: MT-related insights to inspire your growth mindset

Perspectives that matter

Many of the findings of this study reflect the following localization business trends:



Usability over perfection

Shift from linguistic purity to in-context relevance and content availability (less focus on traditional error metrics)

Automation and ease of use

Produce a fully localized and integraded experience in a few clicks and eliminate errors later.

Acceleration

Create a fully localized experience in minutes.

Scalability

Use NMT to localize web content sized from a few to milions of words.

Figure 15: Trends

Effectiveness

Solutions like these hit a sweet spot of content usability, instant deployment, and ease of use. This creates a great opportunity for localization managers to expand and optimize without taking risks to choose the best possible technology. The findings presented in this report open a perspective of machine translation as a highly effective way to translate a website. If you have somewhat reserved feelings towards MT being used to translate website content, hopefully our findings dissolve some of the reservations.

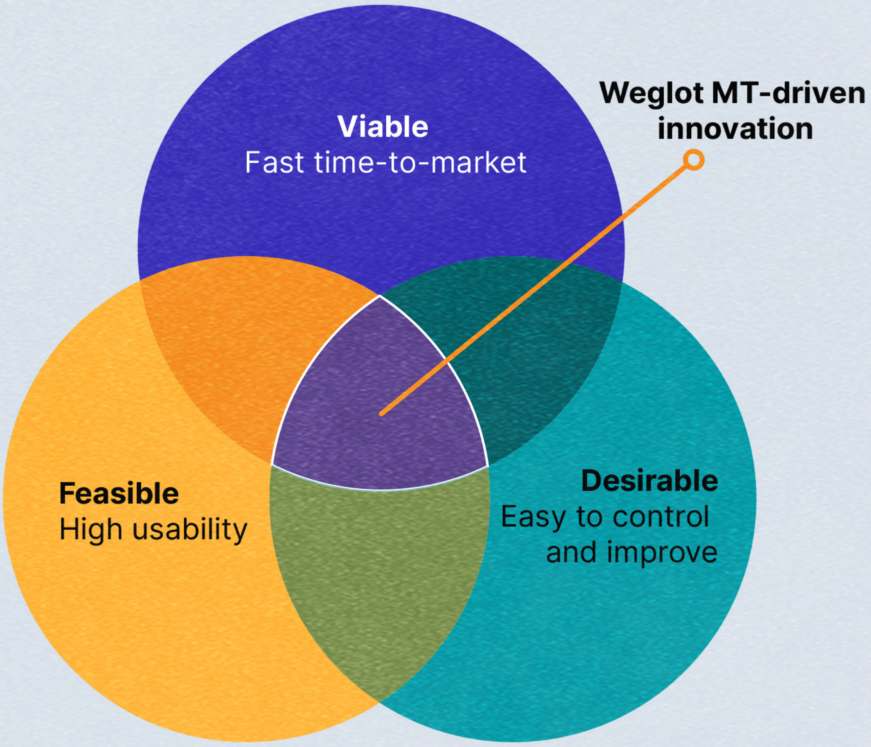


Figure 16: Design perspectives

Study brought to you by Nimdzi Insights and Weglot, 2022

Weglot is a no-code website localization solution that allows you to launch a multilingual website instantly. It both translates and displays the content of your website removing the pain of having to manage multiple websites for multiple markets. Manage the translation of your website translation project in days not months with a first layer of machine translation for speed and automation, then use Weglot's post-editing features to control the quality of your translations. Easily collaborate with teammates, order professional translators from the Weglot Dashboard or add your own translator. Complete the website localization process by translating more than just the words on your website, including images, metadata and content coming from outside of your website (e.g. a review app). Weglot is powering 60,000+ multilingual websites around the world. Some of our customers include Microsoft, Spotify, Steve Madden, Murad, Crabtree & Evelyn and Volcom. Learn more about Weglot's features and capabilities from our website or contact us.

Nimdzi creates the knowledge to empower your success. For your company to dominate the competition and be ahead of the game, you require insights – and that is exactly where we come in. We come from diverse backgrounds in the language industry. We are a market research and international consulting company made up of analysts, consultants, experts, and researchers. But we are all connected with one united goal – helping our clients succeed. And yes. We know the industry. We build actionable insights and reports that cater specifically to your products and services. Working with us means building relationships with influential players – an essential part of penetrating your market.