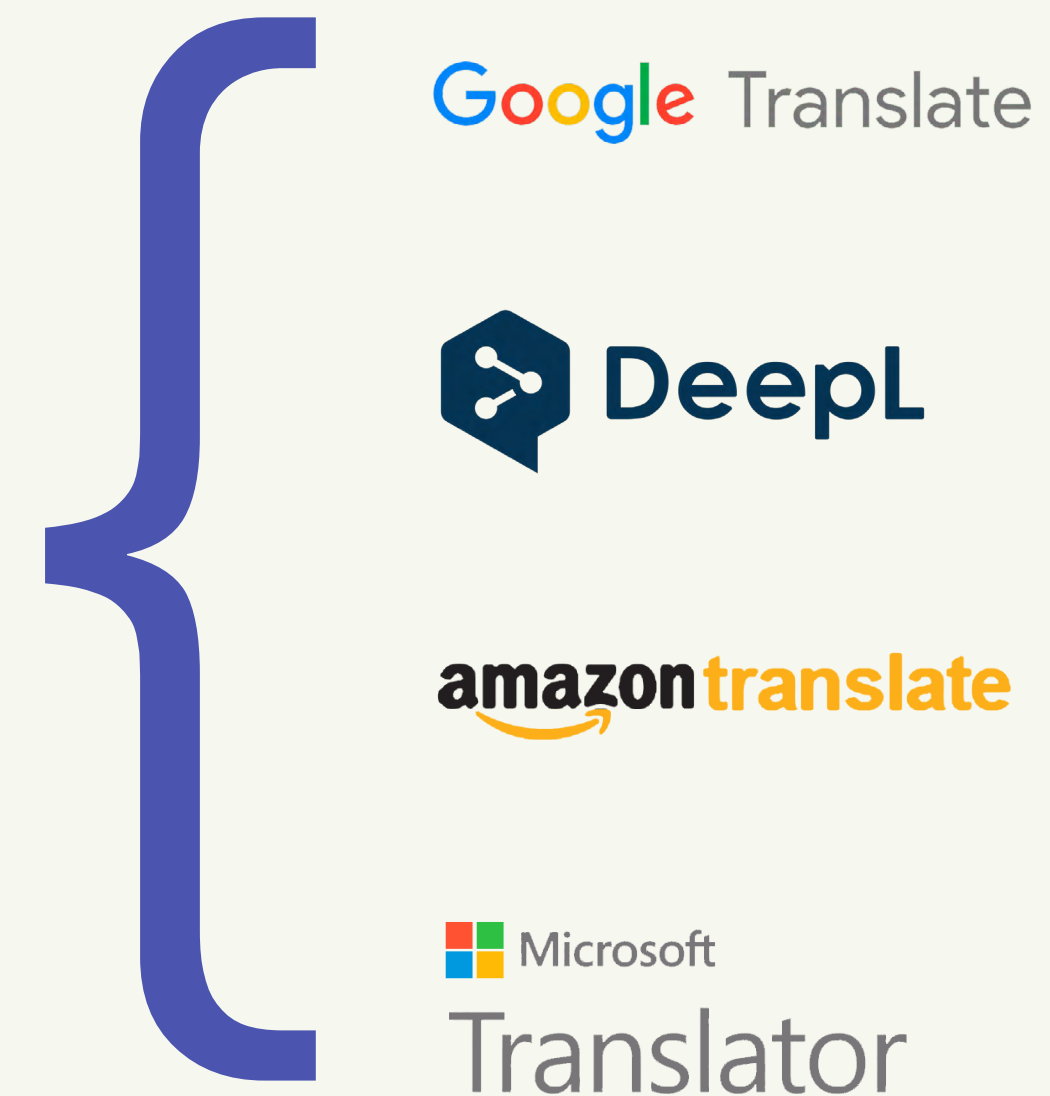


What are the best Machine Translation API's ?

*Evaluating the Quality and Response
Time of Commercial Machine
Translation API's*



*Gabriel
Melo,*

*Luciano
Barbosa,*

*Fillipe de
Menezes,*

*Vanilson
Buregio,*

*Henrique
Cabral.*

THANK YOU!

What is this paper

This paper is the result of a thorough benchmark study that we carried out with the best machine translations APIs on the market: Google, Amazon, DeepL and Microsoft.

This paper was written by Bureau Works engineers.

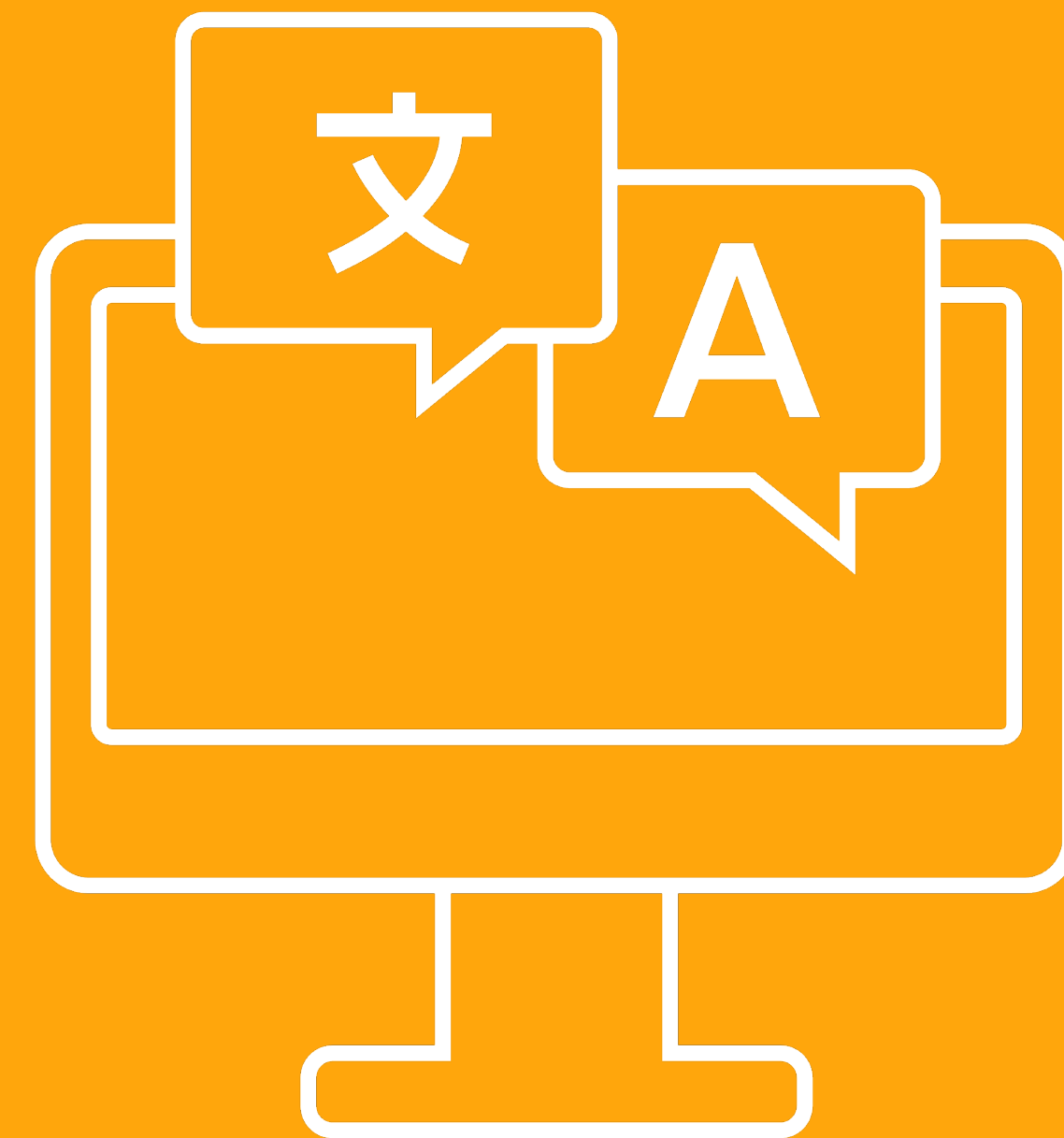
Bureau Works delivers comprehensive in-house translation services on our localization platform that allows for in-depth reporting, evolving translations memory, and automated localization.

Most importantly we combine the business and technical elements of localization under one roof.

Gabriel Melo, Luciano Barbosa, Fillipe de Menezes,
Vanilson Buregio, Henrique Cabral.

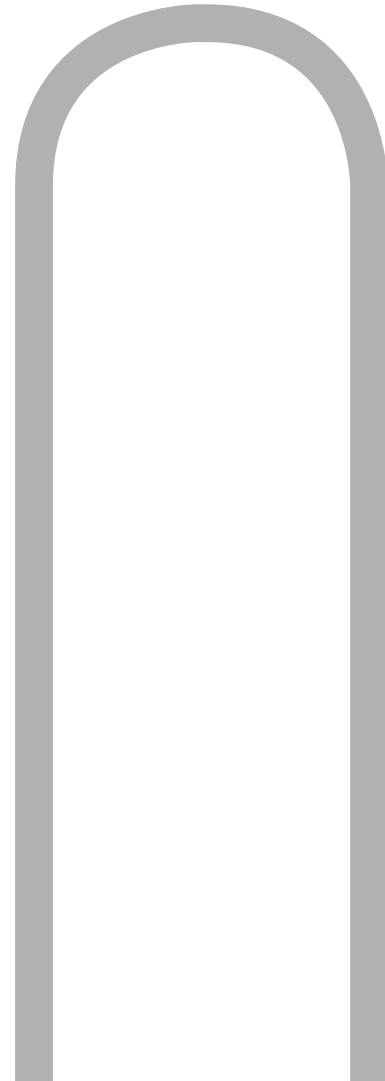
Bureauworks, Universidade Federal de Pernambuco,
Universidade Federal Rural de Pernambuco
3685 Mt Diablo Blvd, Lafayette, CA, United States,
Av. Prof. Moraes Rego, 1235, Recife, PE, Brazil,
Rua Dom Manuel de Medeiros, s/n, Recife, PE, Brazil

{gabriel.melo, filipe, henrique}@bureauworks.com
luciano@cin.ufpe.br, vanilson.buregio@ufrpe.br





This paper is for...



Every company that is planning to implement any kind of translations needs to read this paper because we outline the various advantages and disadvantages of each Machine Translation tool in terms of quality and response time. This in-depth content is geared towards professionals who are actively involved in improving

their translation related products and services, such as:

- Product Managers,
- Project Managers,
- Localizations Managers,
- Engineering Leaders,
- Translators,
- Translation Agencies.



Summary

In this paper we evaluate the quality and translation time of four popular machine translation engines: Amazon Translate, DeepL, Google Translate and Microsoft Translator.

To assess their translation quality, we calculate the **BLEU** score of their translations in comparison to human translations, analyzing different aspects such as target language and the size of the sentence in the source language, which in this study is English. In addition, we measure the response time of those translation APIs, since this is an important feature for applications that require realtime translations, such as traveling apps and translation agencies.

The results show:

- 1** *DeepL and Amazon Translate were the top performers, DeepL achieved the best results for most of European Languages and Amazon Translate for the Asian ones;*
- 2** *in general, the longer the sentence, the better the translation; and*
- 3** *the engines' API provided low translation time, with the exception of DeepL, in which the median time to translate a single sentence was close to 1 second.*

BLEU

(Bilingual Evaluation Understudy) is a measurement of the differences between an automatic translation and one or more human-created reference translations of the same source sentence.

<https://www.bureauworks.com/blog/what-is-bleu-score/>

1. Introduction

Translation services are essential in a wide range of industries and applications. For instance, multinational companies provide content in multiple languages for their customers; and translation apps, such as TripLingo¹ and iTranslate², produce real-time translations to their users.

To attend this demand, in the recent years a growing number of machine translation (MT) engines have become available via public

APIs provided by big techs, e.g., Google, Amazon and Microsoft, and specialized translation companies, such as DeepL³ and Systran⁴.

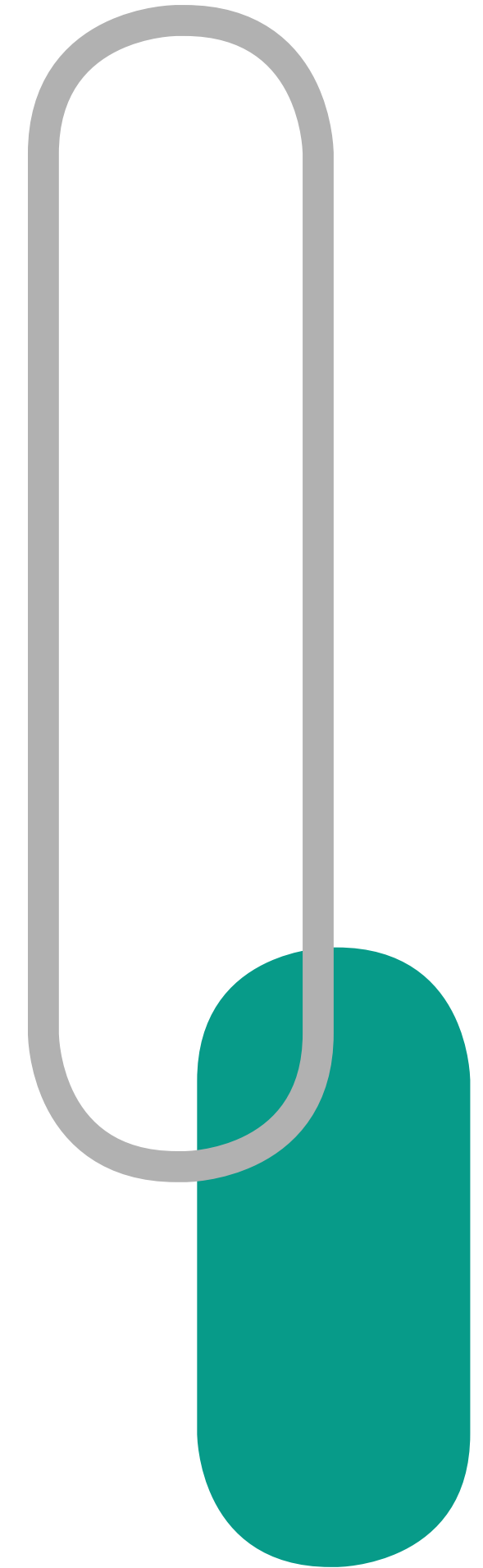
The great challenge for a solution that intends to use those MT engines is to choose which ones are more suitable for its needs. Some aspects to be taken into consideration in this decision are: translation quality, cost and turn around time.

¹ <http://www.triplingo.com/>

² <https://www.itranslate.com/>

³ <https://www.deepl.com/>

⁴ <https://www.systransoft.com/>





Previous approaches have assessed public machine translation APIs with respect to gender bias (Stanovsky et al., 2019), software quality (He et al., 2020; Gupta et al., 2020) and model vulnerability (Wallace et al., 2020). Regarding the translation quality of MT engines in particular,

a recent study⁵ showed there is no single winner for all languages, and commercial engines have a superior performance in comparison to open-source ones.

⁵ <https://try.inten.to/machine-translation-report-2021/>

In the same direction, we assess in this paper the quality of commercial MT engines and, in addition, measure the translation time of their API. More specifically,

we collect more than 200K segments from translation memories in different topics (e.g., health and law) created by professional translators, and use them as ground-truth to evaluate the quality

of the translations of four commercial MT engines (Amazon Translate, DeepL, Google Translate and Microsoft Translator) across seven language pairs having English as the source language.

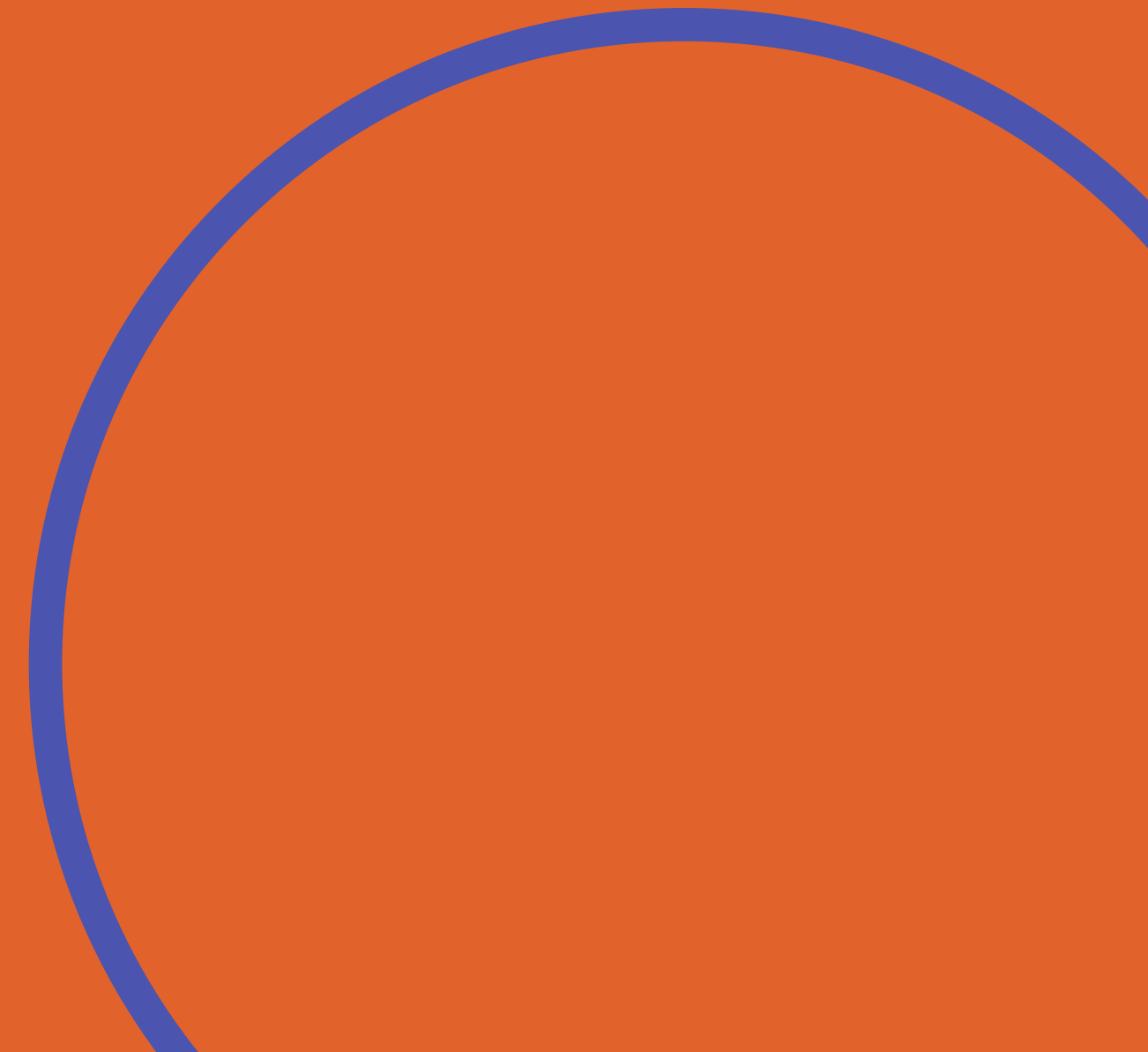
We also analyze the impact of aspects as sentence size and target language in the translation quality. Furthermore, we evaluate the translation time of those engines, since this is a critical factor for real-time applications. For that, we send translation requests of a single sentence (single call) and a batch of sentences (bulk call) to the machine translation APIs during a period of time.

The main results of our analysis are:

- The translation quality of the MT engines are similar across target languages, but DeepL and Amazon produced the best translations: DeepL for European languages and Amazon for Asian languages.
- In general, the longer a sentence, the better the translation quality. And DeepL and Amazon generated the highest quality translations for long sentences;
- The engines' API provided low translation time, which make them suitable for real-time translation applications, with the exception of DeepL, in which the median time to translate a single sentence was close to 1 second.
- The translation time for all engines grows linearly with respect to the number of segments to be translated. But DeepL has a much higher linear coefficient than the other engines in the single call scenario, and Amazon in the bulk scenario.

2. Experimental Setup

In this section, we present the setup we used in our experimental evaluation. More specifically, we describe the ground-truth dataset, the machine translation engines, and the metrics used to evaluate the engines.



2.1 Data

The dataset used in this evaluation, originating from 13 translation memories from different companies generated by professional translators, has English as the source language and seven target languages: German (de), Spanish (sp), French (fr), Italian (it), Japanese (ja), Brazilian Portuguese (pt) and Chinese (zh). Every sentence in English has at least one correspondent pair with one of the mentioned target languages. There is a total of 224,223 segments in English in the dataset and 315,073 pairs.

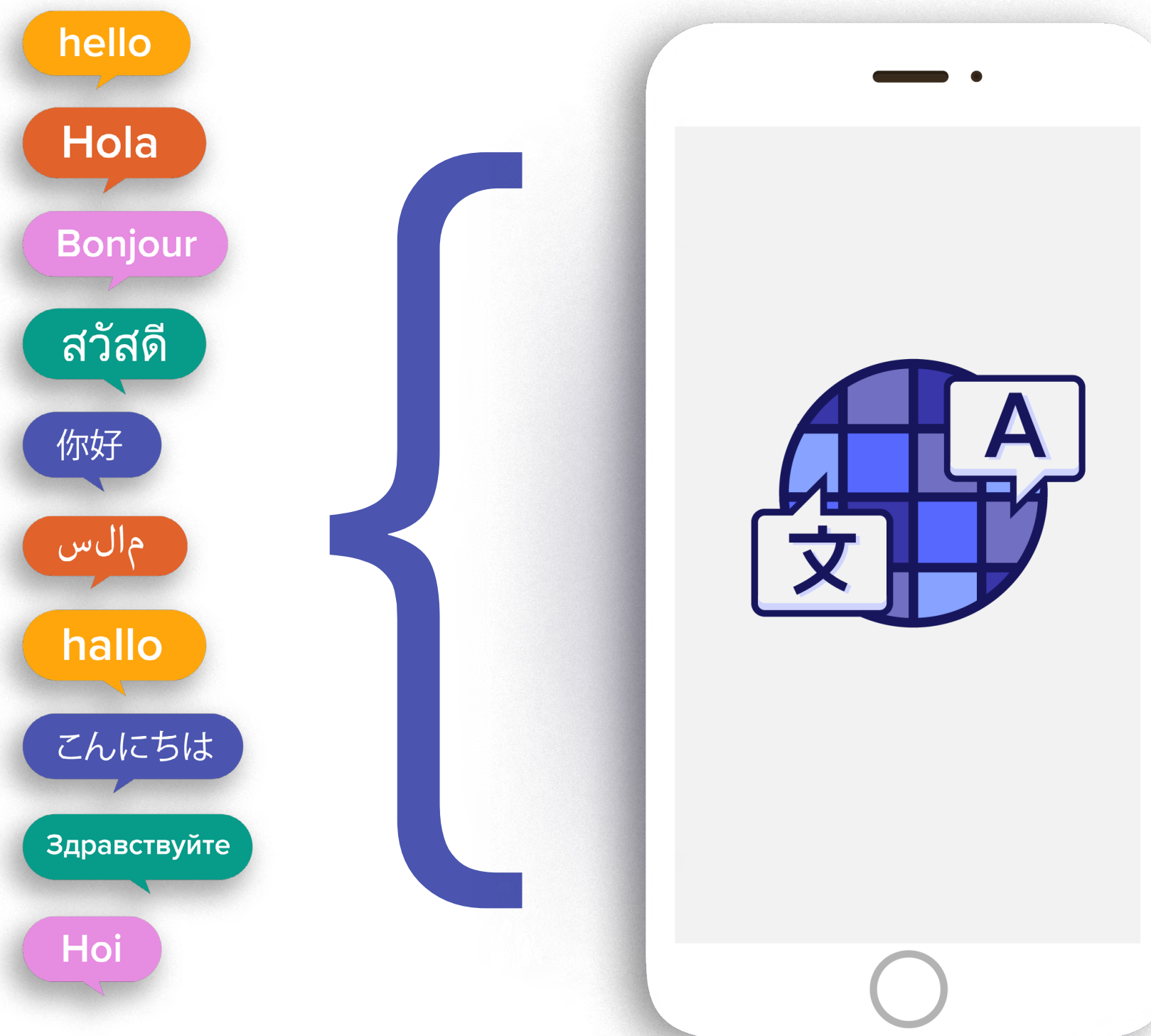


Figure 3 presents the distribution of number of segments for each target language. Brazilian Portuguese has the highest number of segments (near 60k), whereas Japanese and Spanish the lowest one, around 20k segments.

An important feature of this dataset for this evaluation is that it covers a great diversity of topics.

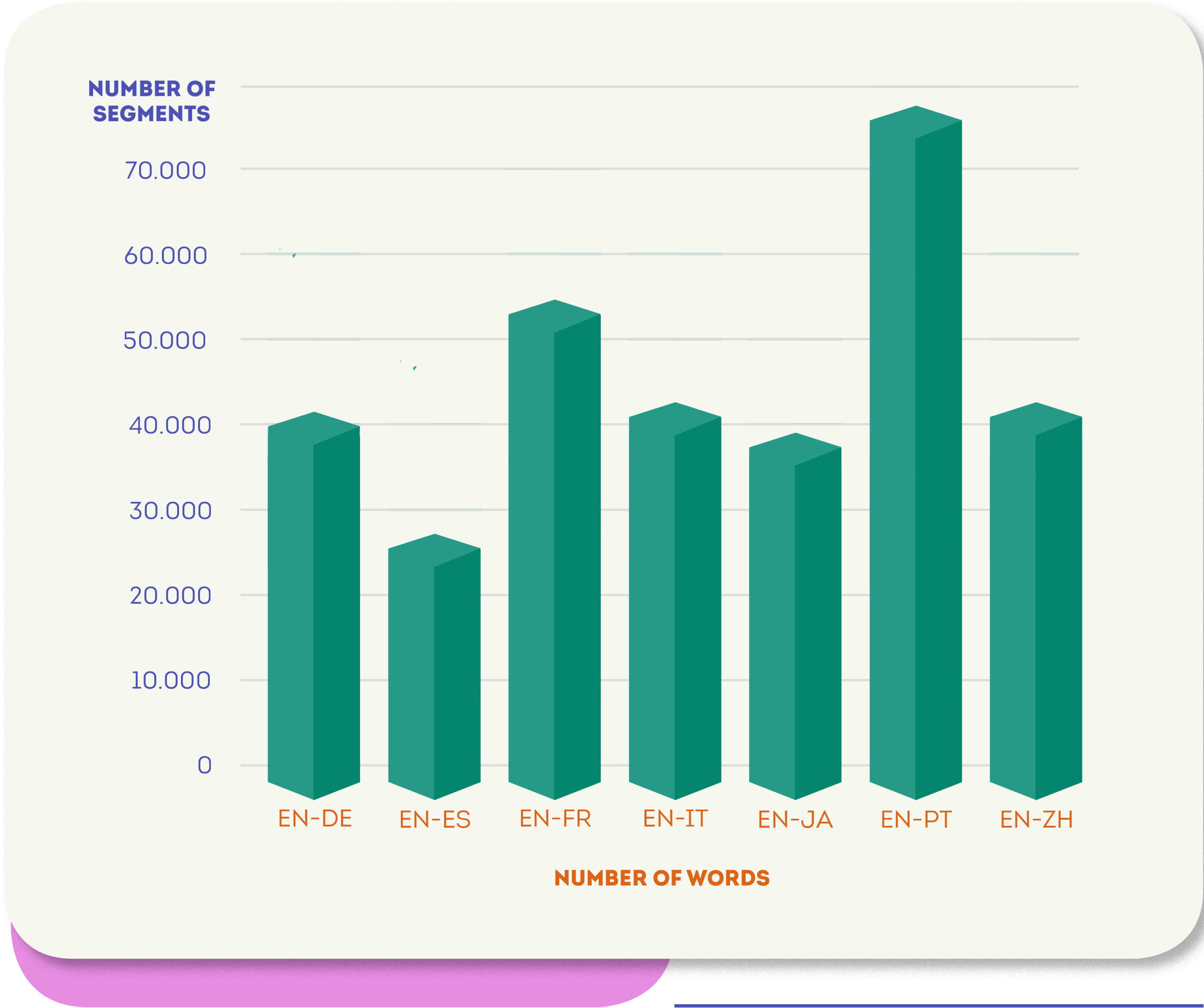


Figure 03
Number of segments for each language pair.

Figure 1 shows a word cloud of the English segments. As one can see, there is content related to health, law, information technology etc.

The dataset is structured with a text segment in the source language, and a reference list with the translations in the target languages. These reference lists have at least one translation associated with the original text, although it could have more than one, as a segment can have more than one possible translation.

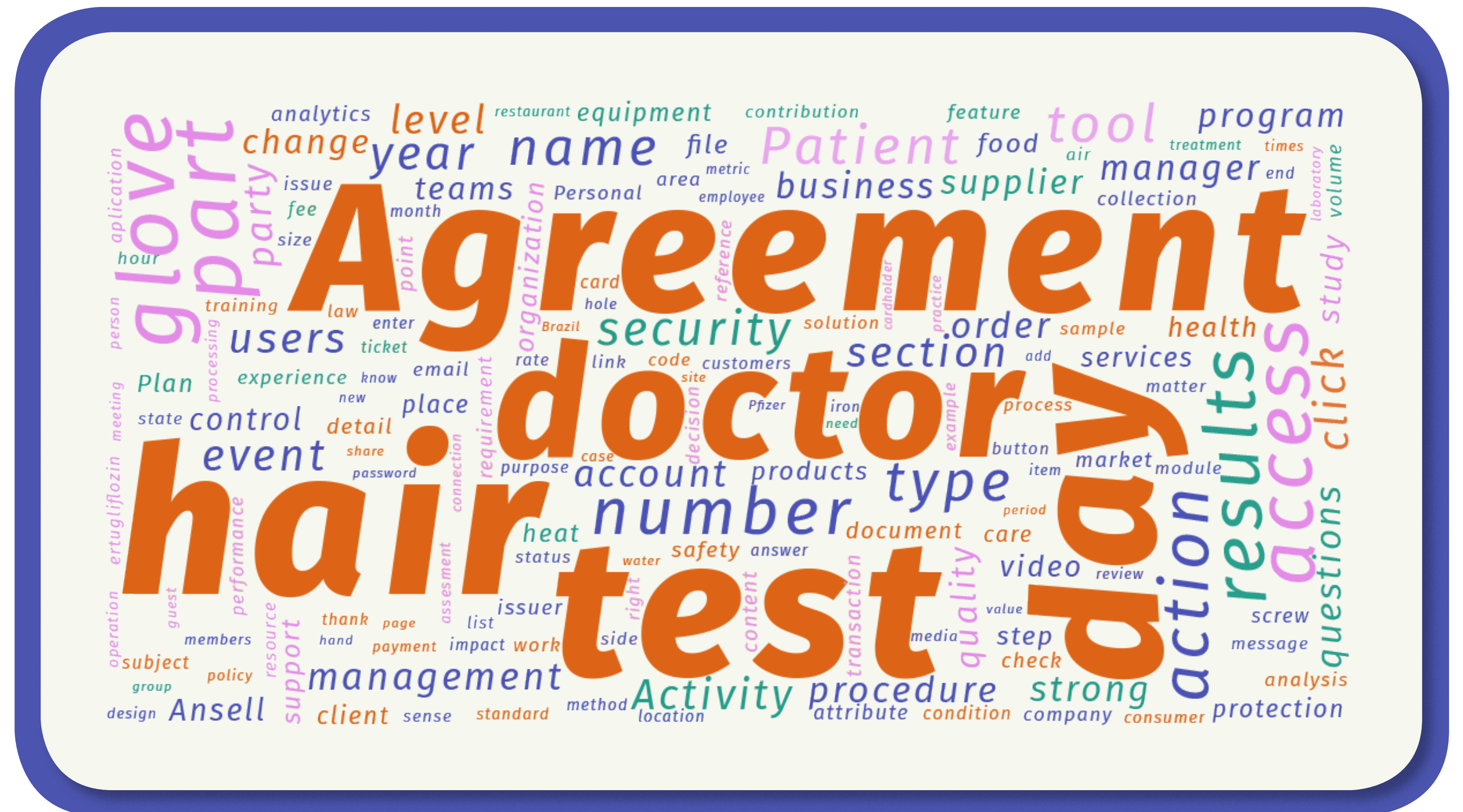


Figure 01

Wordcloud of the segments.

To simplify our analysis, we grouped the segments in ranges of size 10, as shown in Figure 2, in order to evaluate the impact of the segment size in the quality of the engines' translation.

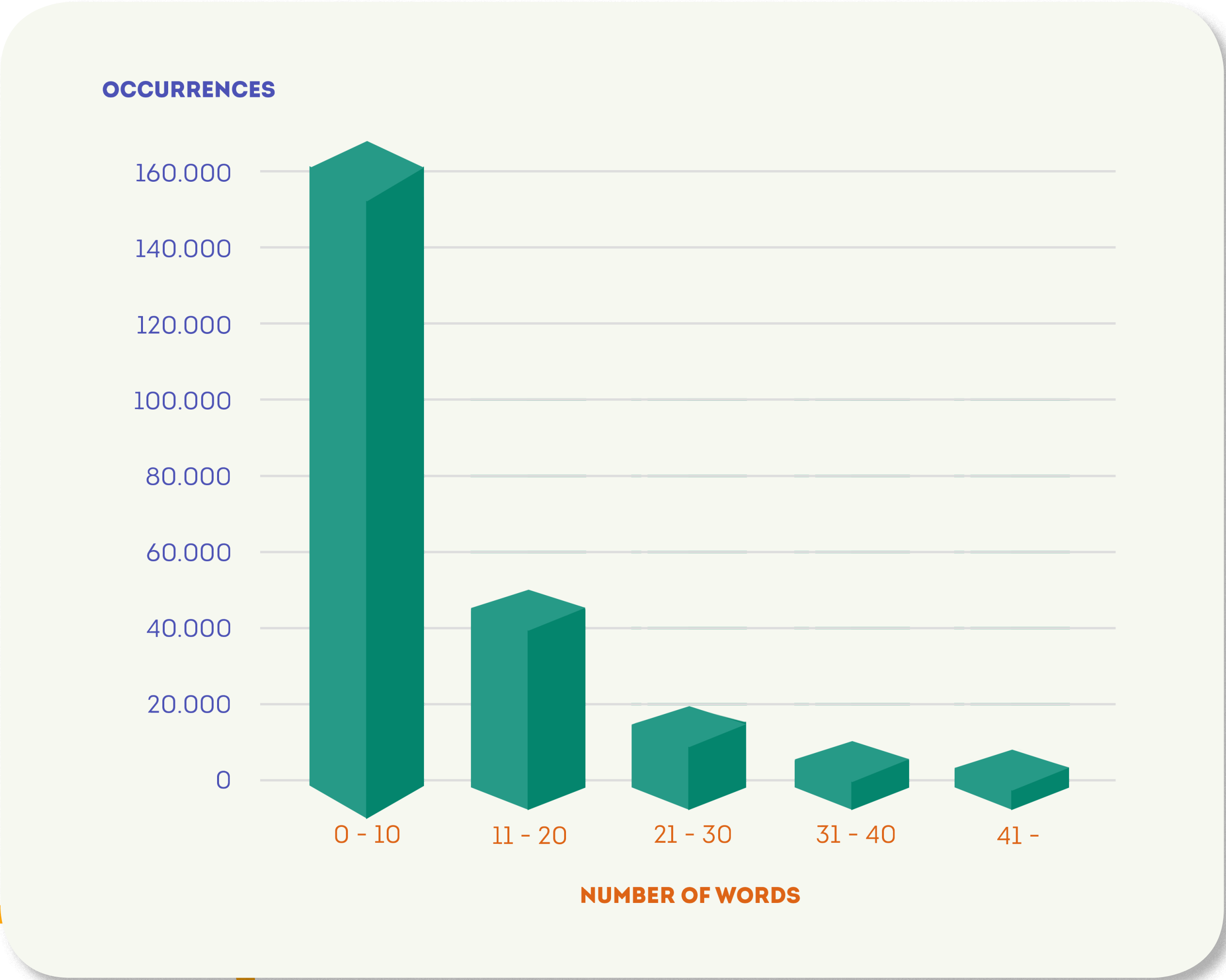


Figure 02
Number of source segments per interval size.

2.2 Machine Translation Engines

For this evaluation, we selected four commercial machine translation engines that support all language pairs in our dataset. We describe them below with their associated cost values as of January 2022.

⁶ <https://aws.amazon.com/translate/>

⁷ <https://www.deepl.com/>

⁸ <https://cloud.google.com/translate/>

⁹ <https://www.microsoft.com/en-us/translator/>



Developed by Amazon, it provides support for machine translation in more than 70 languages. Its Python API is fully integrated with AWS services, at a cost of USD 15 per million characters.



It is company focused on machine translation. Its API supports 26 languages, at a cost of USD 25 per million characters. We used its Python API which enables from and to English translations.



It provides machine translation support for over 100 languages, being the engine with the wider reach in regard to supported languages. It also provides a Python API integrated with all Google Cloud services. The translation pricing is USD 20 per million characters.



It is the machine translation service provided by Microsoft at a cost of USD 10 per million characters, being the lowest pricing among all evaluated MT engines. This engine supports near 90 languages.

The selected MT engines are all able to translate a single segment through their respective API, and except for Amazon Translate, they can also respond to a bulk call, when a list of segments are submitted and returned at once.

To deal with the bulk limitation of Amazon Translate, we made a minor coding optimization in the single call in order to eliminate the need to establish a connection to the API at every translation, which is not near a bulk translation but helped to reduce the gap between this and the other engines with bulk translation support.

Although all mentioned MT engines were suitable for tuning their models with parallel data or a glossary for specific terms, we decided to put these options aside for this evaluation.

We also try to evaluate other MT engines (e.g., Baidu Translate¹⁰, Tencent¹¹, Systram PNMT¹², Apertium¹³, Alibaba¹⁴), but we could not use them for one of the following reasons: API unavailability, lack of documentation, or no support for all target languages.

¹⁰ <https://fanyi.baidu.com/>

¹¹ <https://ai.tencent.com/ailab/en/index/>

¹² <https://www.systran.net/en/translate/>

¹³ <https://www.apertium.org/>

¹⁴ <https://www.alibabacloud.com/product/machine-translation>



2.3 Metrics

We evaluate the translation quality of the engines using BLEU score (Papineni et al., 2002). We used Friedman's test (Friedman, 1940) to compare the scores of different engines, and the post hoc Nemenyi test (Nemenyi, 1963) to verify statistical significant differences between individual MT engines.



To calculate the APIs' response time, we selected a sample of 100 segments of our dataset, respecting the distribution of intervals of segment sizes (Figure 2), and translated them in each engine from English to Portuguese.

We hit the engines with the selected sentences once a day for one week to assess the APIs' methods: single and bulk. We did not use the whole dataset and only translated in one target language to evaluate the response time, because it would be financially costly to hit the engines for one week with 200k segments in seven languages.

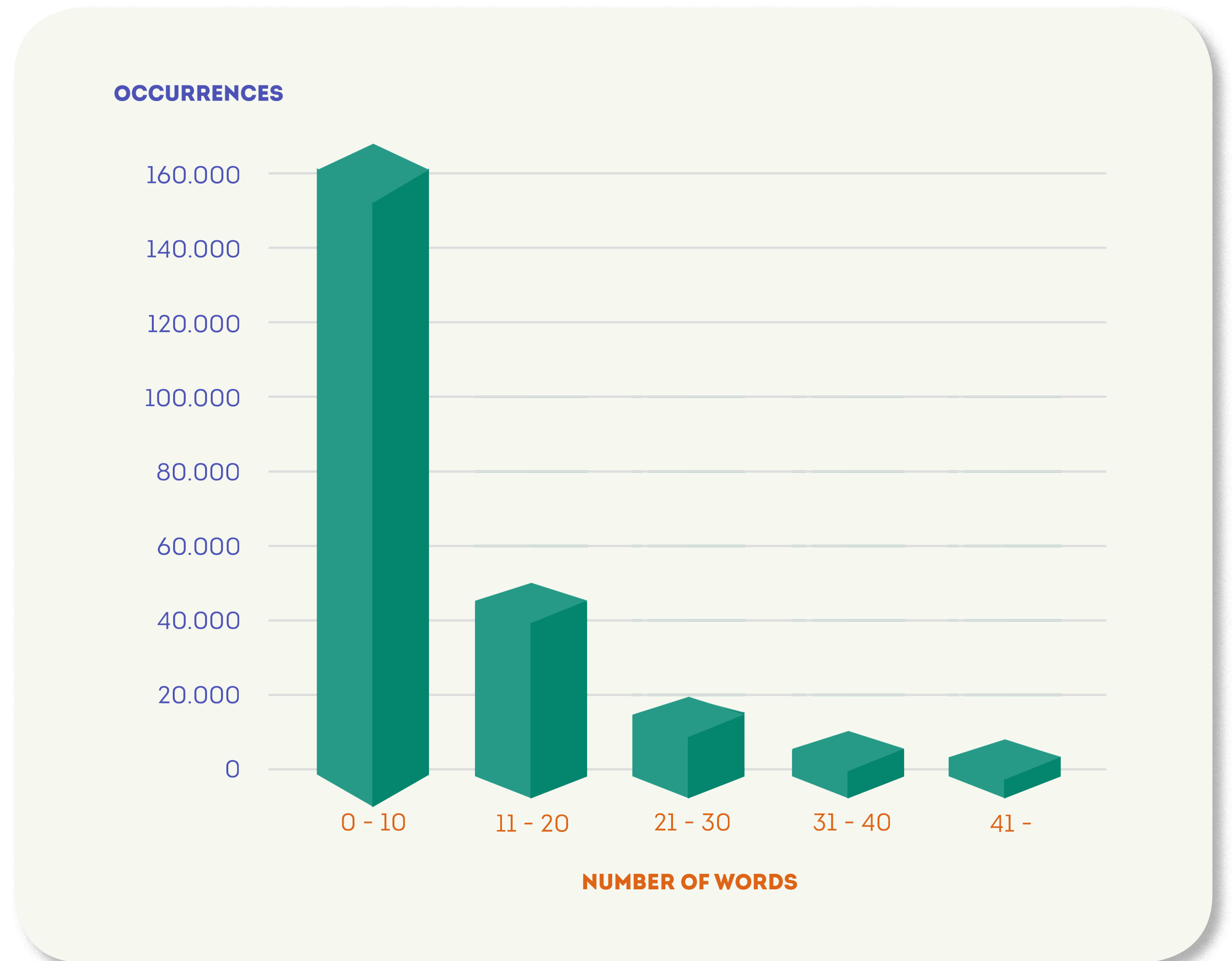


Figure 02
Number of source segments per interval size.

3. *Experimental Results*

In this section, we present the results of our investigation about the performance of the machine translation engines described in Section 2.



3.1 Quality Evaluation

Table 1 presents the mean BLEU score of the four engines on each target language. For all languages, the p-values of Friedman’s test were smaller than the significance level (0.05), meaning that there are statistically significant differences in the scores of the engines. In addition, the engines with best scores for each language had performance statistically different of the other ones, according





				
German (de-de)	0,63	0,65	0,64	0,62
Spanish (es-es)	0,73	0,74	0,74	0,73
Italian (it-it)	0,67	0,69	0,67	0,67
French (fr-fr)	0,72	0,72	0,70	0,71
Portuguese (pt-br)	0,74	0,71	0,72	0,71
Japanese (ja-jp)	0,61	0,59	0,60	0,59
Chinese (zh-cn)	0,54	0,51	0,54	0,53

Table 1
Average BLEU score of the machine translation
Wengines in the target languages.

to the post hoc Nemenyi test with p-values lower than the significance level of 0.05. Amazon and DeepL achieved the best overall results with the highest scores in 4 target languages. Google tied with DeepL in Spanish and with Amazon in Chinese, whereas the Microsoft translation engine did not outperform any MT engine in any language.

In Figure 4, we present the BLEU score distribution for different segment sizes on each target language. A common trend in these plots is that the longer a sentence, the better the BLEU score.



Figure 04
BLEU score distribution per segment size.



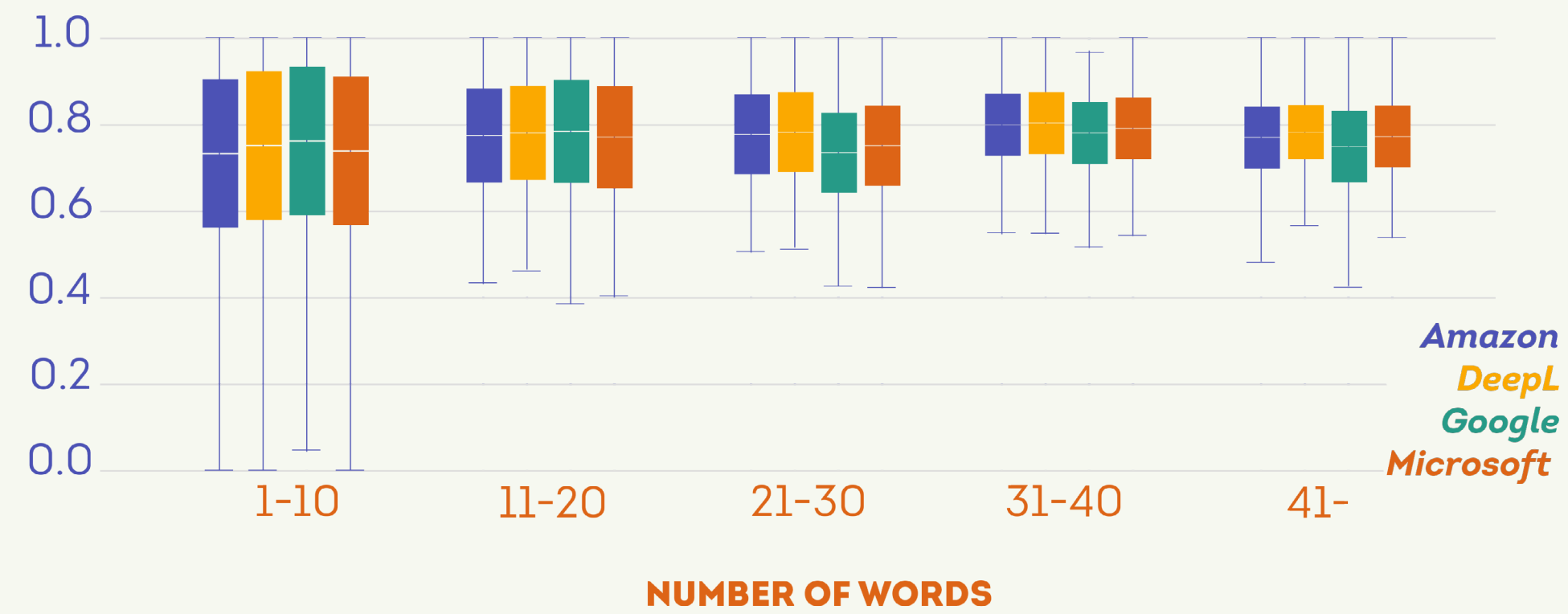
For instance, the median scores of all MT engines for German as the target language (Figure 4a) were around 0.6 for segments with size between 1 and 10 and close to 0.7 for the segments greater than 40 words.



Japanese is the only exception (Figure 4f): the segment size did not affect the translation quality of Amazon and DeepL, but affected the quality of Microsoft (median BLUE score of 0.61 for the 1-10 interval and 0.58 for the 40- interval) and Google (median BLUE score of 0.62 for the 1-10 interval and 0.6 for the 40- interval).

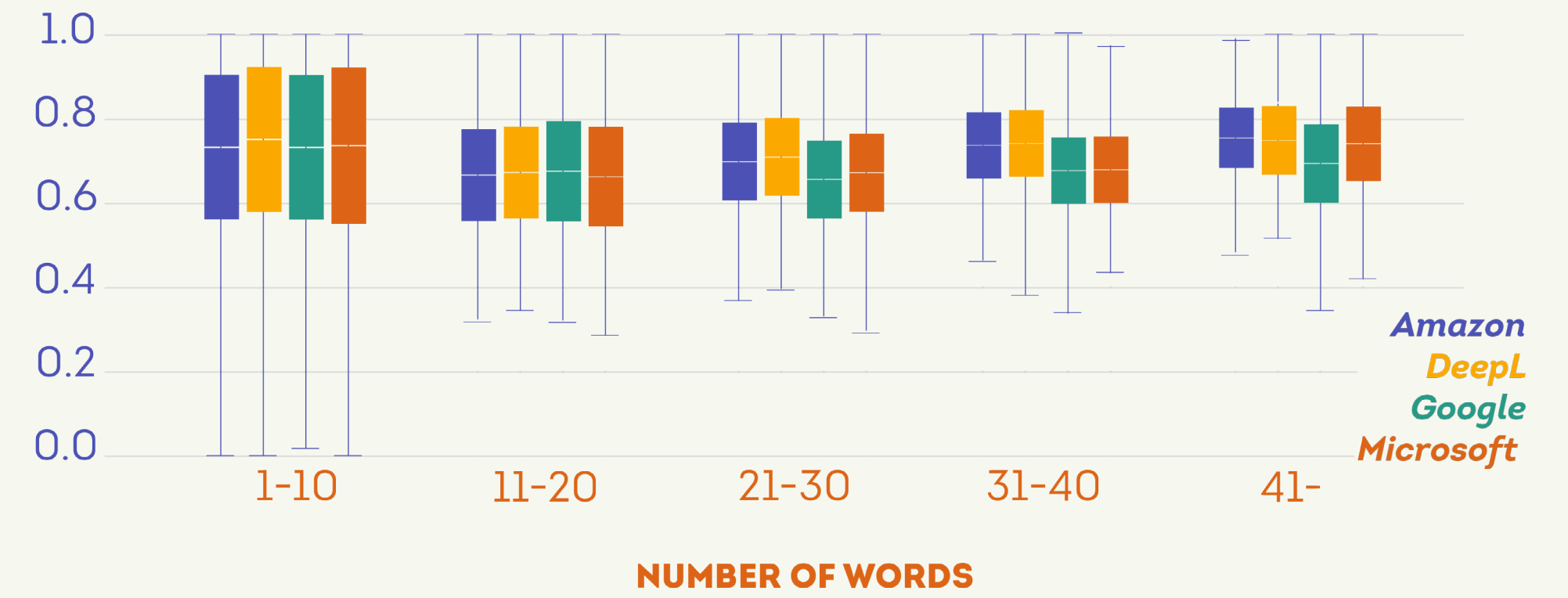
BLUE SCORE

(b) Spanish



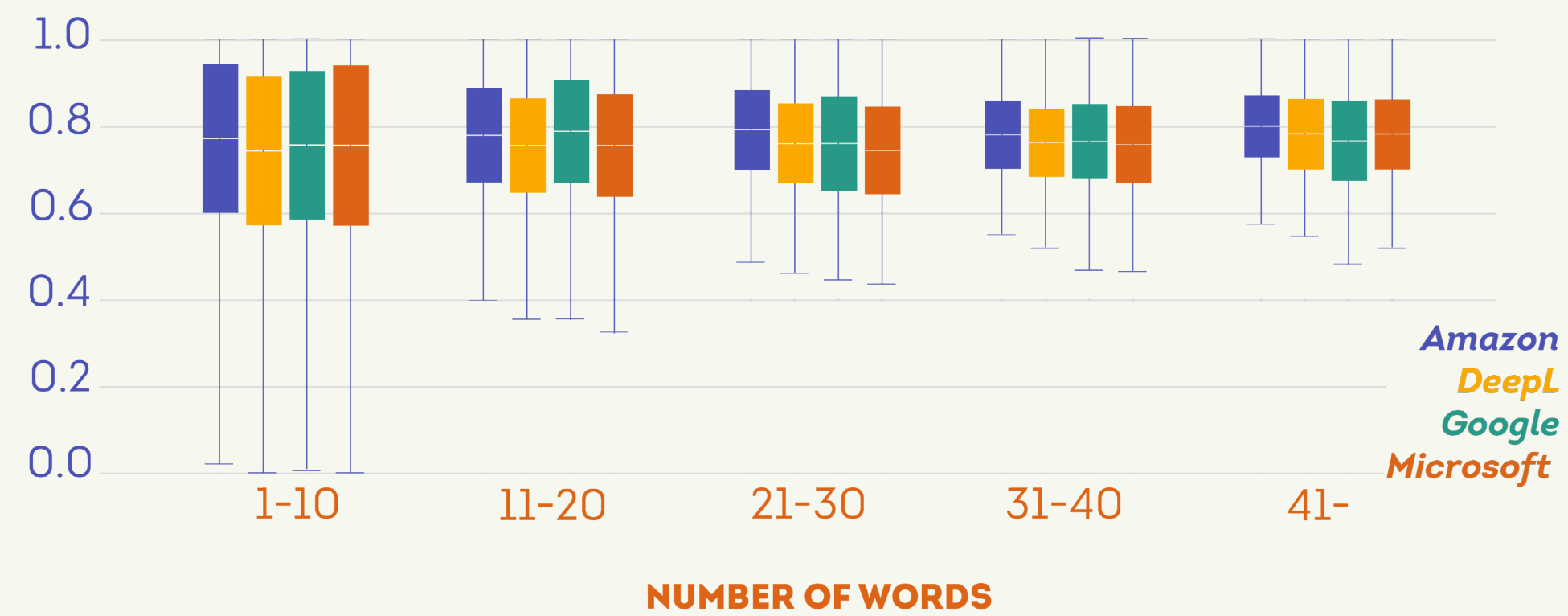
BLUE SCORE

(d) Italian



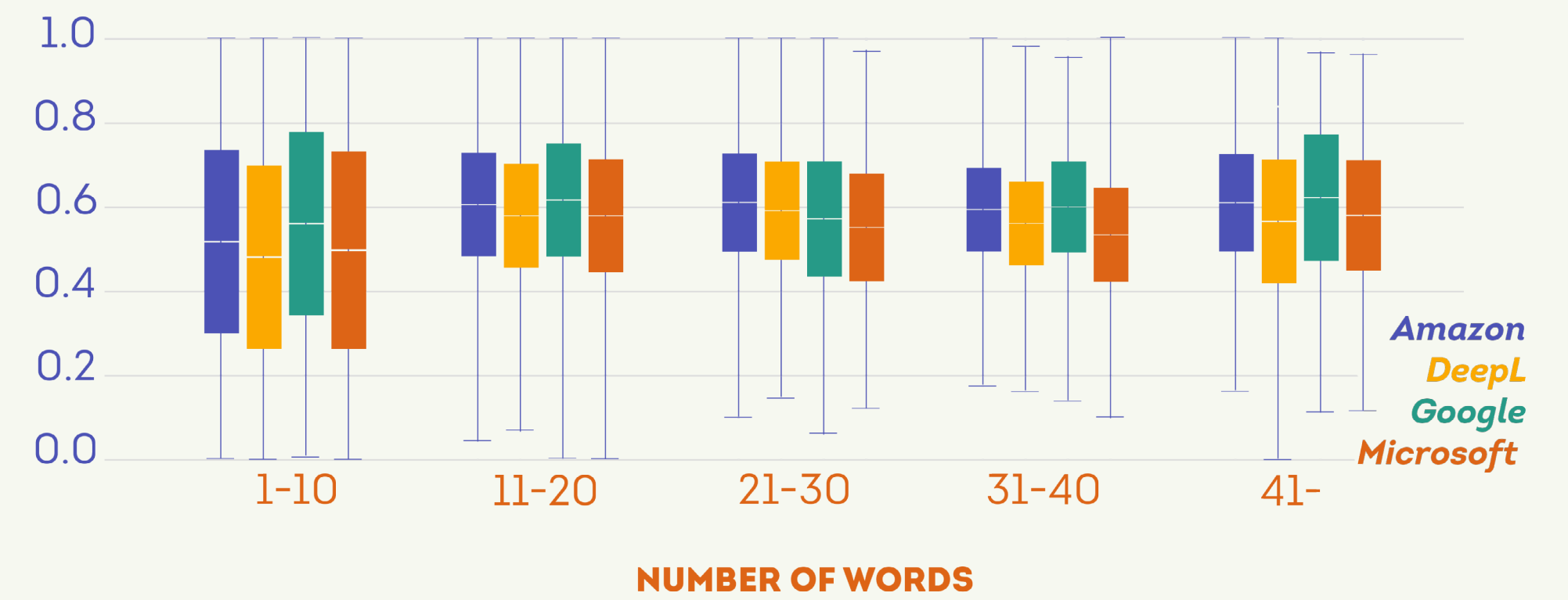
BLUE SCORE

(e) Portuguese



BLUE SCORE

(g) Chinese



3.2 Translation Time Evaluation

Figure 5a presents the distribution of the translation time per segment for each MT engine sending one segment at the time (single), and Figure 5b sending 100 segments at once (bulk).

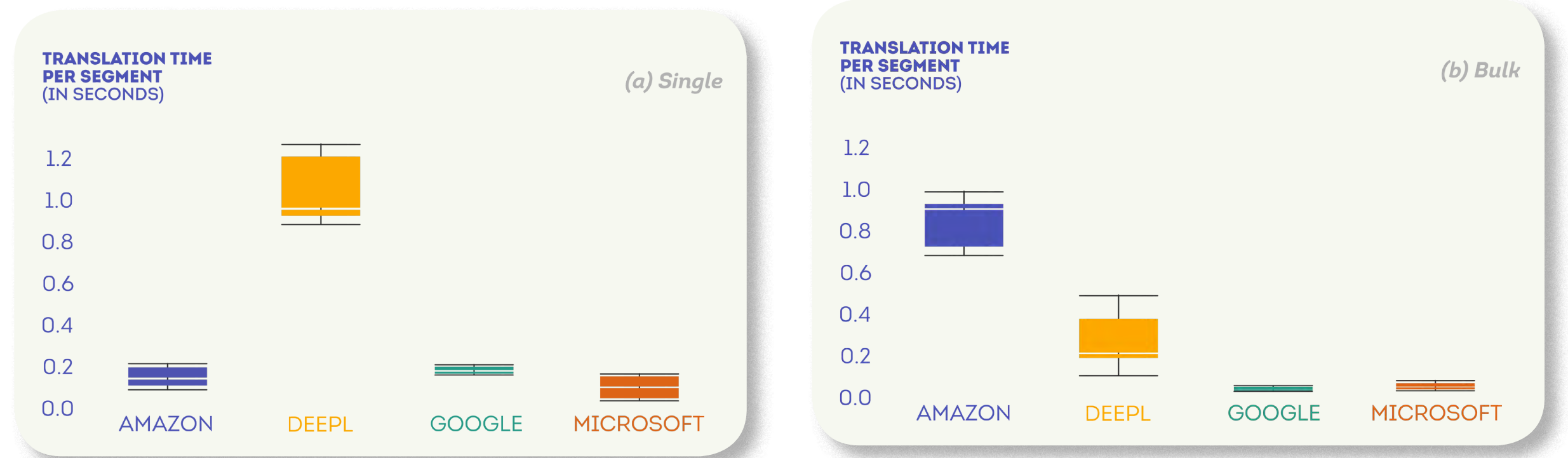


Figure 05
Translation time of the engines.

In the single scenario, Microsoft provided the fastest translation (median of 0.09 second per segment). Amazon and Google were around two times slower (medians close to 0.2 second), and DeepL was the slowest one (median of 0.96 second per segment), almost ten times higher than Microsoft.

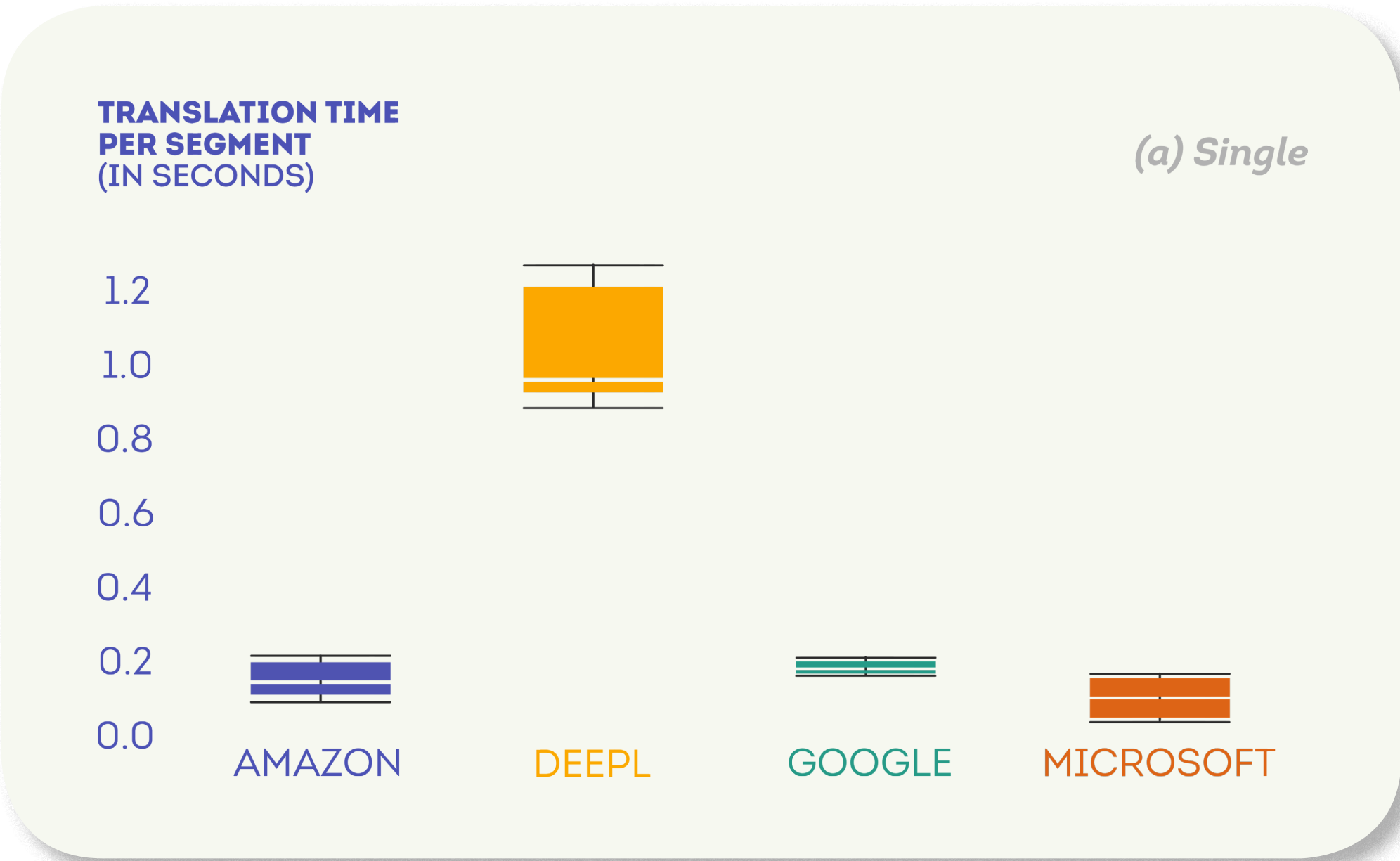


Figure 05
Translation time of the engines.

The first thing to notice when using the bulk call of the APIs (Figure 5b) in comparison to the single one (Figure 5a) is that there was a great reduction in the translation time per segment. For DeepL, for instance, the median time of translation per segment decreased from 0.95 second, in the single execution, to 0.02 second in the bulk one. These results clearly show that the bulk operation is much more efficient than sending segments individually for translation. Regarding the individual performances of the engines, Microsoft and Google obtained the lowest translation

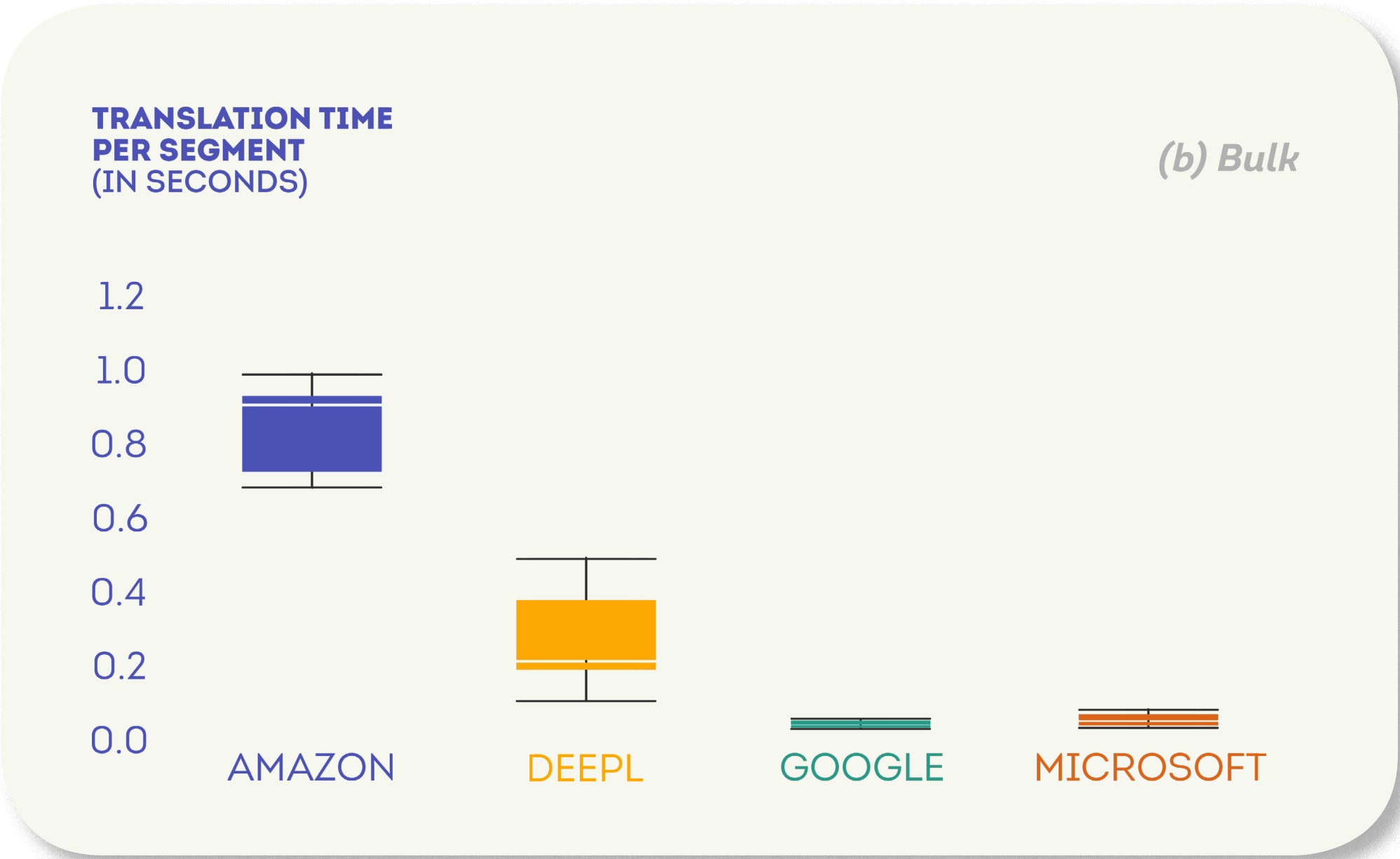


Figure 05
Translation time of the engines.

times (median of 0.003 and 0.002 second per segment, respectively), whereas the highest translation time was from Amazon (median of 0.09 second). We believe the reason for this poor performance of Amazon is that it does not provide a real bulk call, which we had to approximate in our experiments as aforementioned.

The evaluated MT engines, therefore, presented low translation time per segment which make them suitable for real-time translation applications. The only exception was DeepL in the single scenario in which the median translation time of a single sentence was close to 1 second.

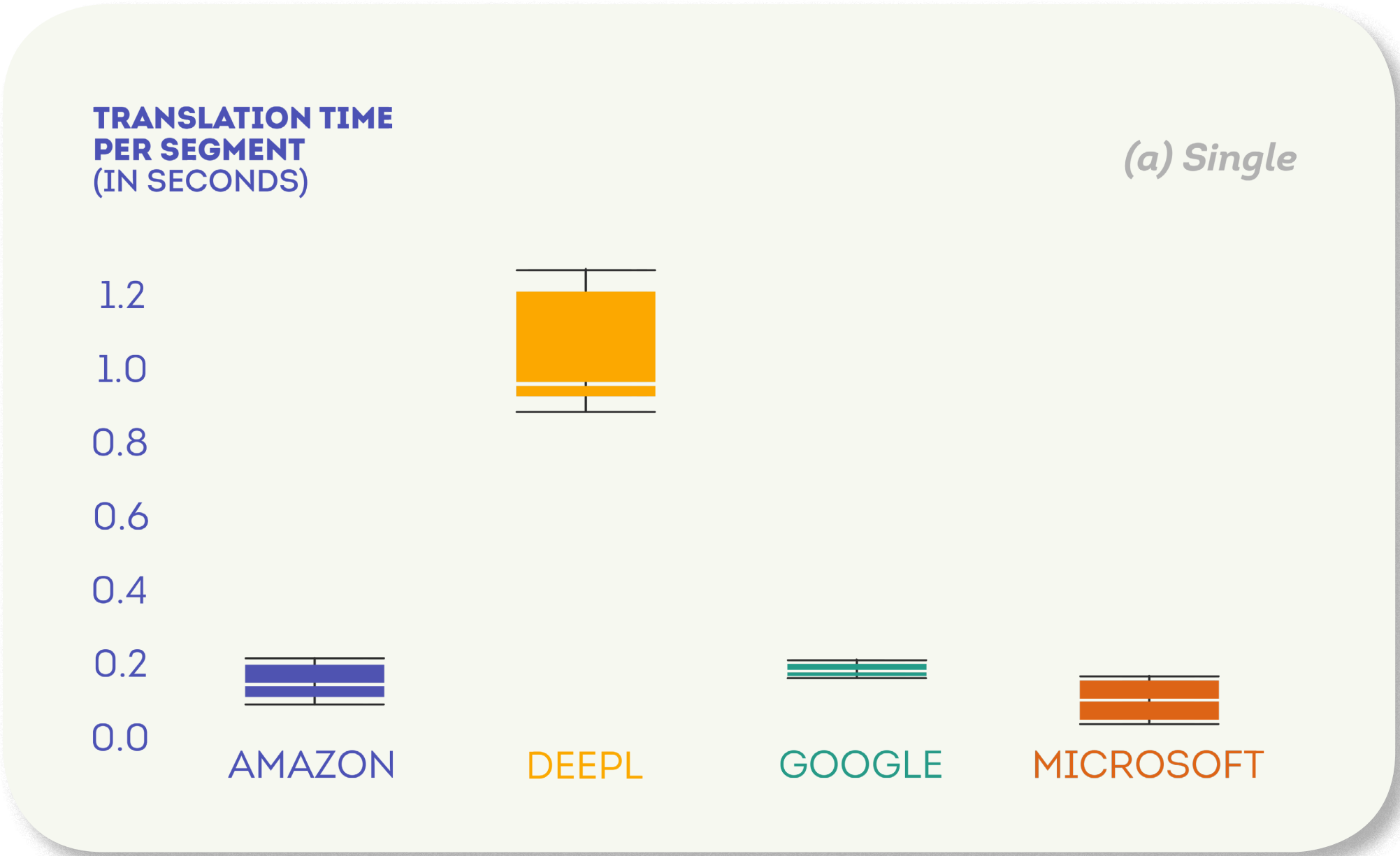


Figure 05
Translation time of the engines.

To analyze the scalability of the engines, we present in Figure 6a and 6b the response time of the MT engines when we vary the number of segments. In all curves, the time grows linearly with the number of segments.

However, the linear coefficient of some of the engines is much smaller than the others. For instance, DeepL has the highest coefficient in the single scenario and Amazon the highest in the bulk one meaning that they do not scale as well as their competitors in each respective scenario.

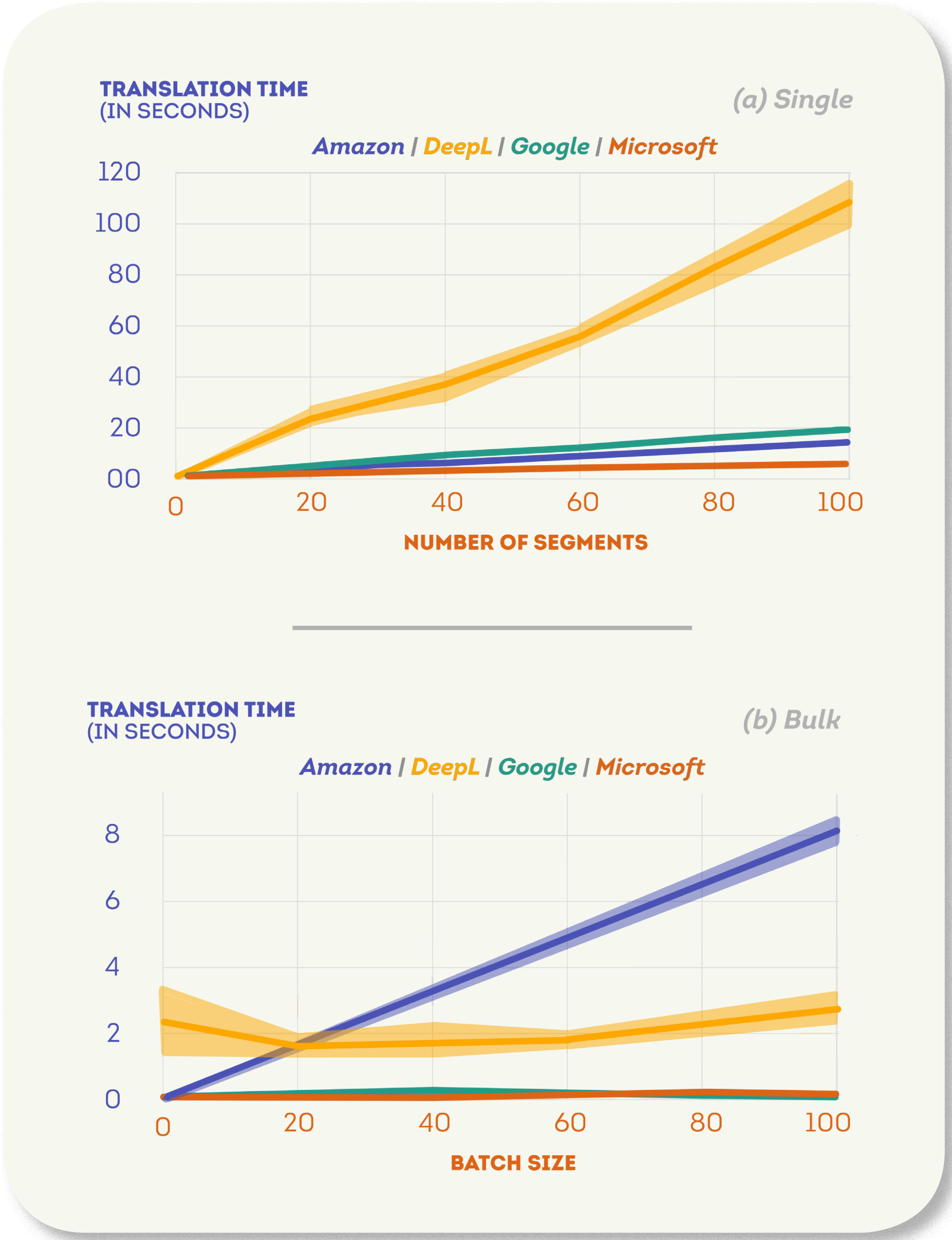


Figure 06
Translation time of the engines
varying the number of segments.

4. Conclusion

In this paper, we presented an evaluation of four machine translation engines with respect to their quality and response time. Our evaluation showed the quality of the engines are similar, but having Amazon and Deepl as top performers. Regarding response time, overall the engines presented good performance, with exception of DeepL, when sending one segment at the time, and Amazon in the batch call.



5. *Bibliographical References*

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92.

Gupta, S., He, P., Meister, C., and Su, Z. (2020). Machine translation testing via pathological invariance. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 863–875.

He, P., Meister, C., and Su, Z. (2020). Structure invariant testing for machine translation. In *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*, pages 961–973. IEEE.

Nemenyi, P. B. (1963). *Distribution-free multiple comparisons*. Princeton University.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Wallace, E., Stern, M., and Song, D. (2020). Imitation attacks and defenses for black-box machine translation systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5531–5546, Online, November. Association for Computational Linguistics.

What are the best Machine Translation APIs ?

*Evaluating Quality and Response
Time of Commercial Machine
Translation APIs*



*Gabriel
Melo,*

*Luciano
Barbosa,*

*Fillipe de
Menezes,*

*Vanilson
Buregio,*

*Henrique
Cabral.*



This paper was developed by Bureau Works engineers and despite all the efforts our team has put into the project; it's just a sample of the wealth of potential we can deliver to our clients. No matter the job, your success matters to us, and we excel at what we do.

Schedule a conversation and find out how our translation services and platform will engage your audience globally.



Talk to us

**We look forward to providing you
with an amazing experience!**

