SDL*

SDL Machine Translation

Quality Report

Data and results collected and compiled Q4 2019 through Q1 2020

Executive Summary

This report provides a summary of how market-leading MT software solutions perform on two popular automated metrics (BLEU and LEPOR) that attempt to measure relative MT quality.

These metrics provide a rough guide to the automatic translation quality users can expect from these products.

BLEU (bilingual evaluation understudy) attempts to measure how close an automatic translation is to what a human translator would produce.

LEPOR is an alternative automatic test designed to correct for issues commonly found with BLEU, such as language bias.

This report provides a point in time analysis on how these software solutions rank on automated quality scores. This is a single measure of a single dimension of product performance and fit. Buyers and evaluators of MT should consider their overall requirements and business objectives when evaluating any technology, including machine translation.

People want "just one measure" , or a single letter grade to understand system quality, but the actual reality is much more complex than that.

The danger of single measures that summarize complex performance characteristics is that they are almost ALWAYS inadequate.

Automated metrics such as BLEU and LEPOR, both of which were used to capture the results presented in this report, should be complemented with human evaluations of translation quality (segment ratings, edit distance, error classification) along with business outcomes such as productivity gains, cost savings, and time to market. Published MT comparisons using automated scoring typically use news domain and other easily obtained data sets. The testing data used for generating the quality comparisons may not resemble data that an enterprise is using.

In an effort to provide a transparent quality evaluation using the most popular automatic metrics, SDL ran side by side comparisons of SDL Machine Translation against three common MT vendors: Google, Microsoft and DeepL. The translations were obtained using the available APIs and covered twenty-eight (28) language pairs and spanned European, Asian and Cyrillic language sets.

Methodology and Coverage

The SDL MT research team developed an evaluation data set drawn primarily from news domain and public sources typically used for testing. This is similar to data used for evaluations that are available in the public domain.

The data set was run against Google, Microsoft and DeepL. Not all MT providers offer all of the language pairs that were tested. DeepL supports European languages (German, French, Spanish, Portuguese, Dutch, and Italian), a limited set of East European languages (Russian and Polish) and no Asian languages. Google, Microsoft and SDL offer languages beyond those which are shown here.

The table on the right shows which language pairs were used for this test.

English (eng) from/to	**Google**	**DeepL**	**Microsoft**	**SDL**
Arabic (ara)	✓	N/A	✓	✓
Chinese (chi)	✓	N/A	✓	✓
Czech (cze)	✓	N/A	✓	✓
Dutch (dut)	✓	✓	✓	✓
Finnish (fin)	✓	N/A	✓	✓
French (fra)	✓	✓	✓	✓
German (ger)	✓	✓	✓	✓
Italian (ita)	✓	✓	✓	✓
Korean (kor)	✓	N/A	✓	✓
Portuguese (por)	✓	✓	✓	✓
Romanian (rum)	✓	N/A	✓	✓
Russian (rus)	✓	✓	✓	✓
Spanish (spa)	✓	✓	✓	✓
Turkish (tur)	✓	N/A	✓	✓

Translation quality is evaluated by computing LEPOR and BLEU scores between the MT output and the evaluation reference translations.

The test sets consist of data extracted from the following public datasets:

- WMT 2013
- WMT 2015
- WMT 2018
- WMT 2019
- News Commentary 2011
- JHE

* WMT: Workshot on Statistical Machine Translation, **aclweb.org/anthology/venues/wmt/**

* JHE: Junior High Evaluation data for Korean-English, **zenodo.org/record**

SDL showed significant score improvements over previous testing results. These improvements are attributable to:

- Continually updated training data sources
- Continually improved techniques for data cleaning and processing
- Development and application of custom training techniques for improving the handling of specific translation aspects (URLs, numbers, unknown words, etc.)

Results

Results show that **SDL Machine Translation consistently outperformed** Google, DeepL and Microsoft in both LEPOR and BLEU scores.

The biggest differences were observed in the following language pairs (both to and from English): Arabic, Czech, Finnish and Korean. The results are consistent across BLEU and LEPOR. For some language pairs such as Spanish and Dutch, the score differences between the vendors is minor. Other language pairs such as Korean and Arabic show a dramatic difference, with SDL a clear leader.

SDL's innovative work to break the industry **quality standard** for Russian to English shows. While English to Russian scores are fairly consistent across the vendors, SDL scores significantly higher for Russian to English.

Figure 1: LEPOR scores showing translations from English

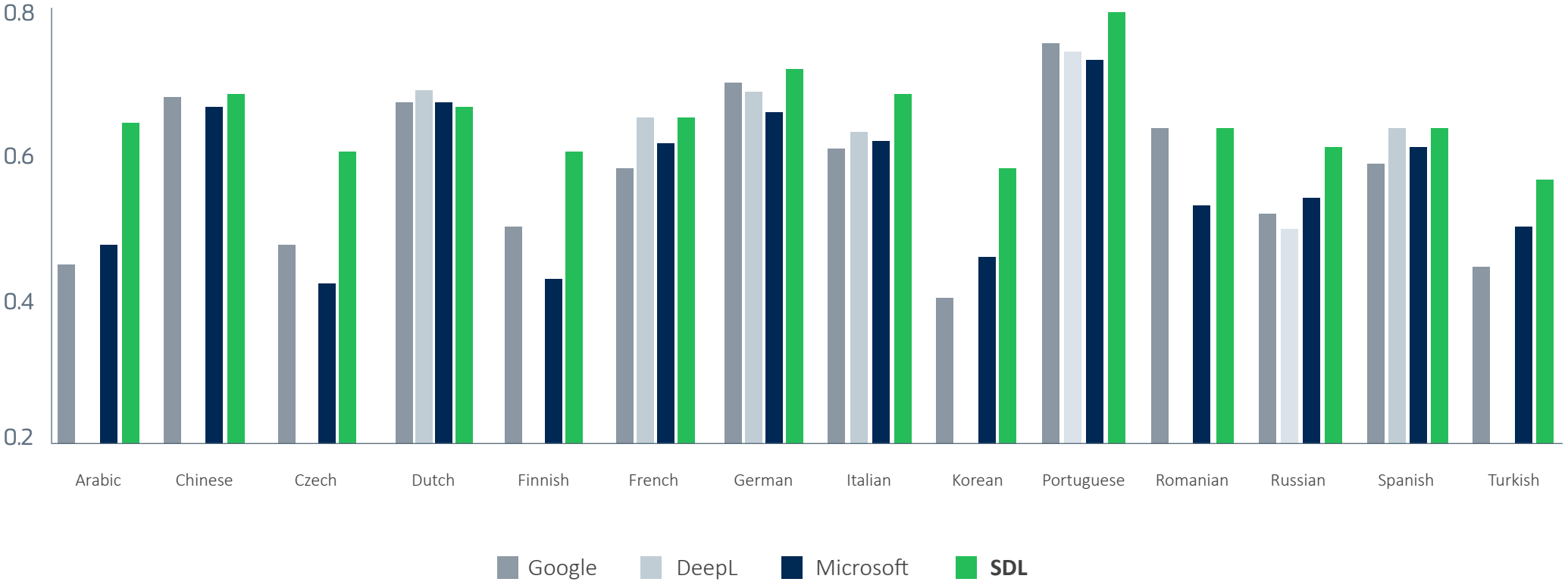


Figure 2: LEPOR scores showing translation to English

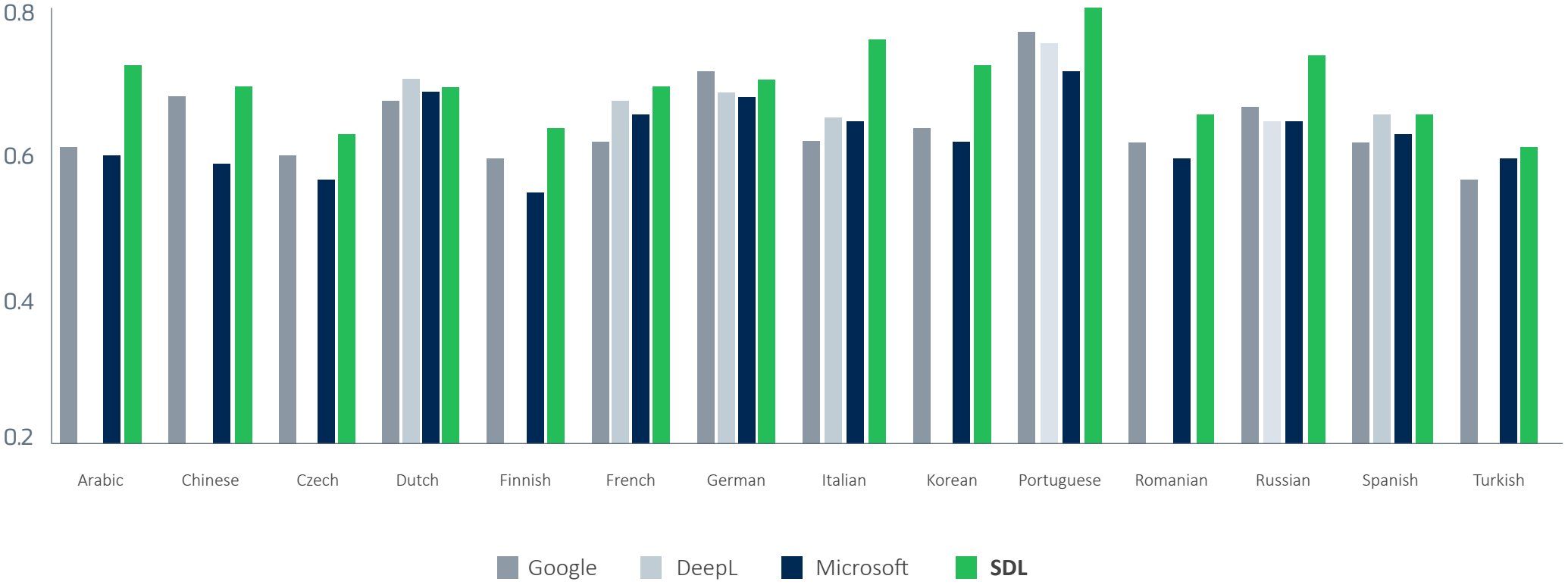


Figure 3: BLEU scores showing translations from English

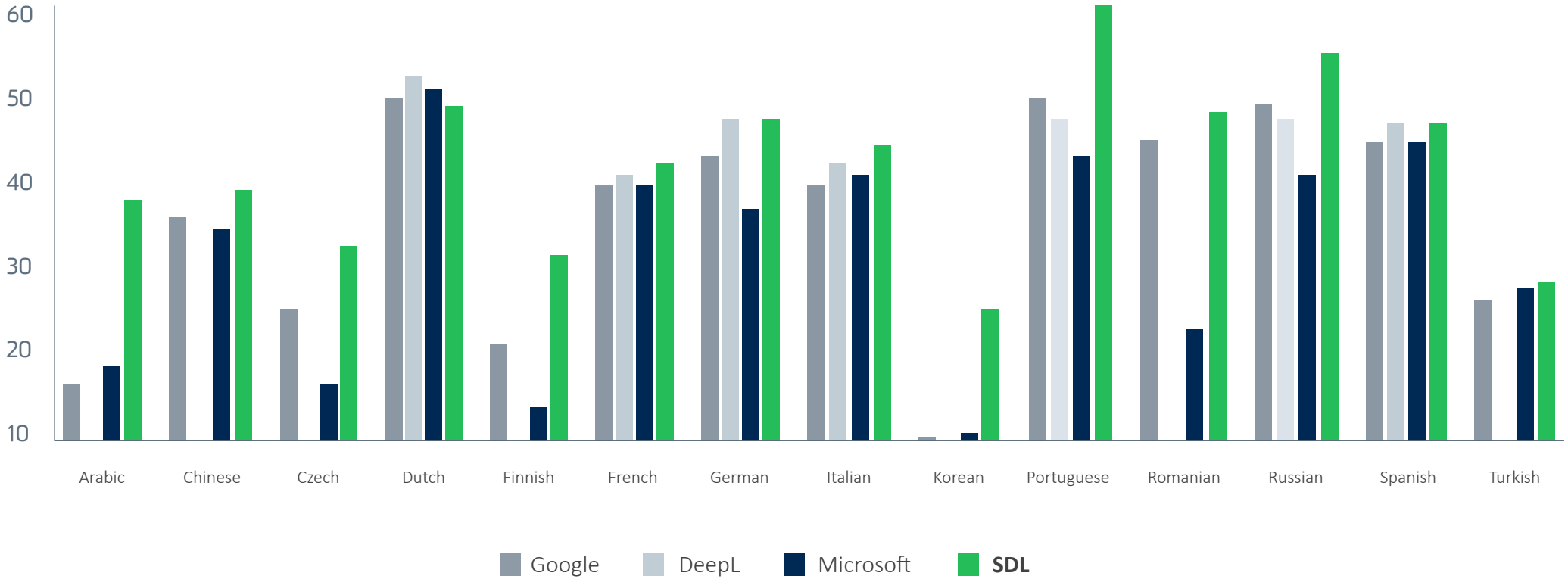
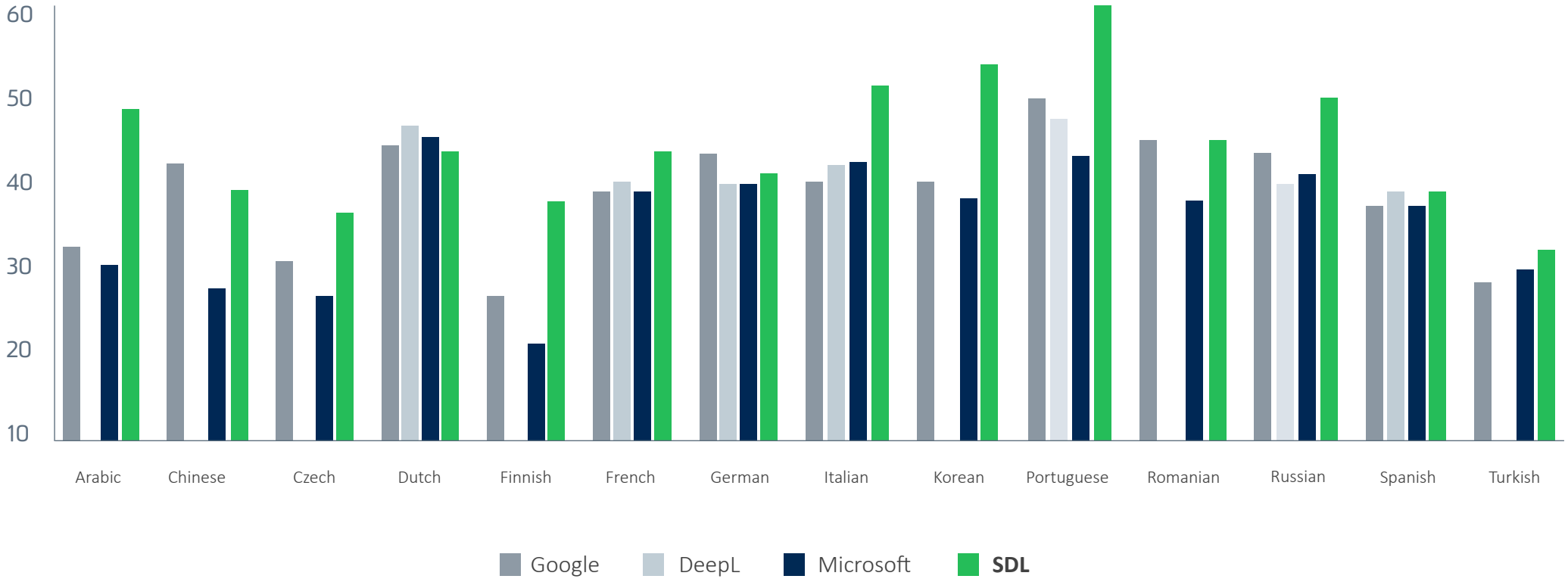


Figure 4: BLEU scores showing translation to English



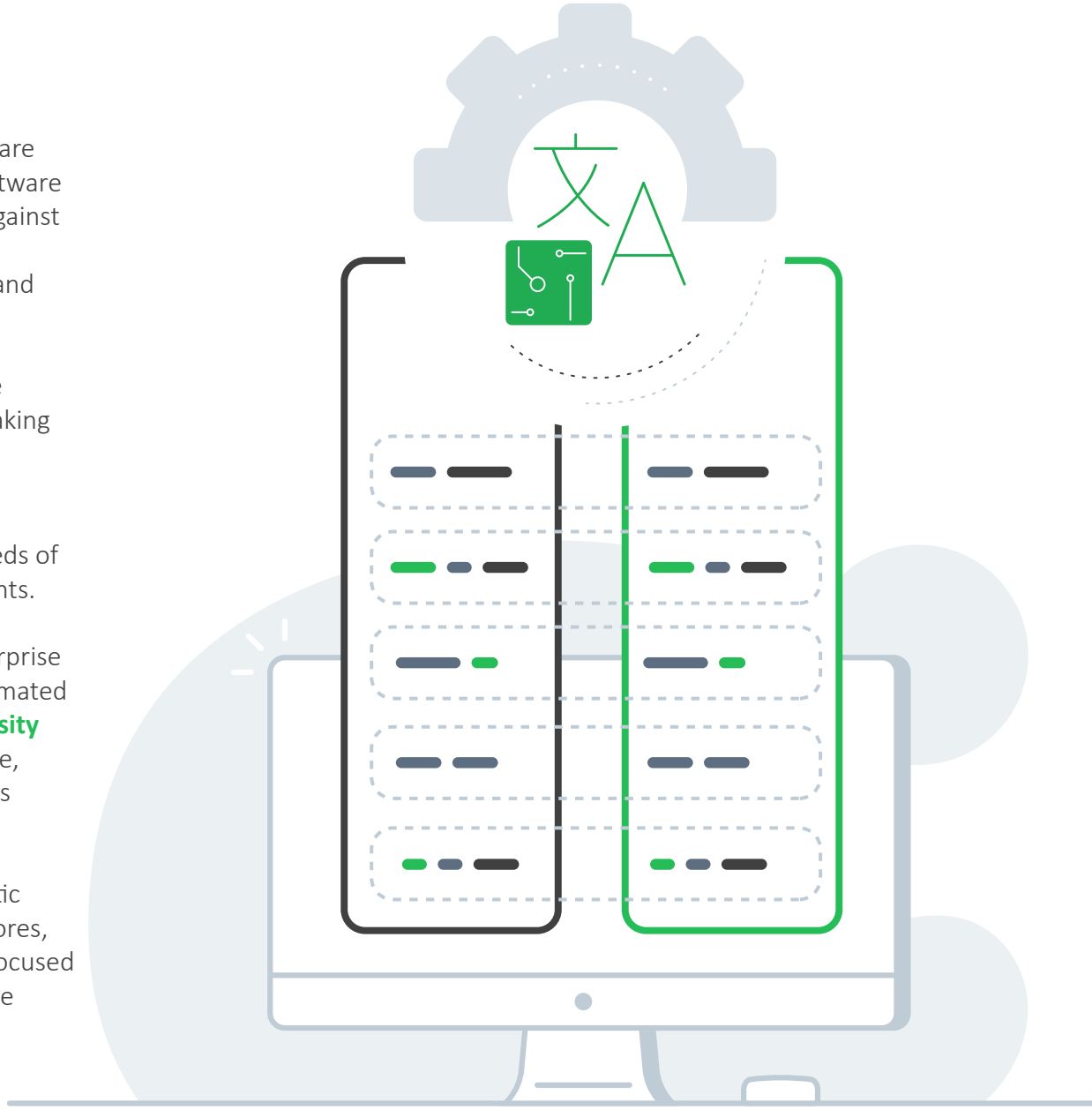
Conclusion and Recommendation

Enterprise buyers continue to seek a way to easily compare MT software performance by using automated quality scores. However, overall software performance characteristics are more complex than a single metric against a single dimension. Experts caution that relying on a single measure may obscure critical details and lead to sub-optimal implementation and business impact.

Unfortunately, the reality of quality assessment is bigger than a single number. Informed buyers should consider additional factors when making their purchase decisions.

In most cases, MT use within the enterprise is going to require some degree of adaptation of the generic MT technology to the unique needs of the enterprise, both in terms of terminology and use case requirements. The capabilities of different MT systems to accommodate and enable these customizations are a critical requirement of MT use in the enterprise business mission context, and this is often more important than automated metric scores. **Research conducted by SDL together with the University of Texas** shows that use case related performance matters much more, and that a few BLEU points do not necessarily result in better business outcomes.

The MT "quality" discussion needs to evolve beyond targeting linguistic perfection, chasing proximity to human translation, BLEU or Lepor scores, and focus more on business outcomes. When enterprise buyers are focused on those factors that will impact the business value delivered, they are better able to measure quality that matters.

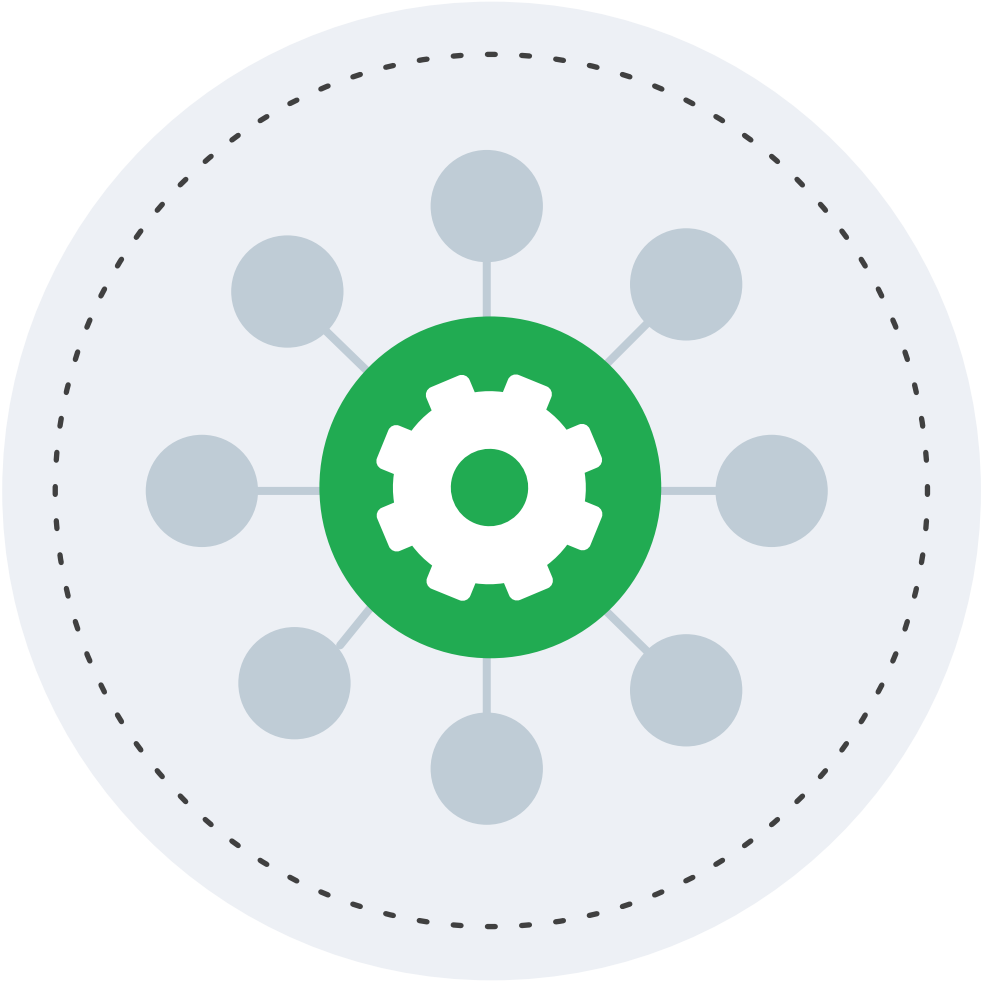


The following factors vary significantly between the many MT systems in the market. These factors are independent of any linguistic considerations yet are critical to the successful application of MT to enterprise business functions:

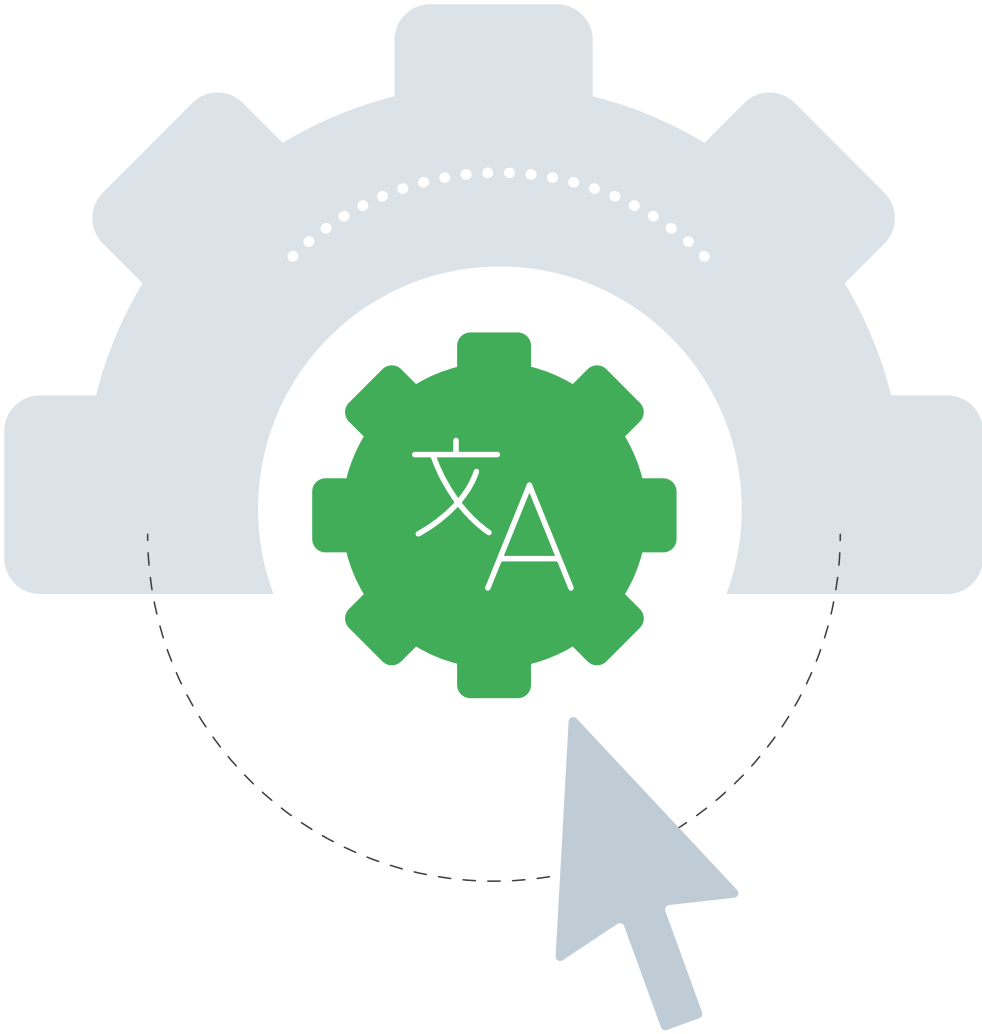
- Data Security and Privacy
- Ability to incorporate and adapt to enterprise terminology
- Integrate directly into enterprise IT infrastructure to maximize efficiency and ease
- Solution deployment flexibility
- Availability of expert MT and linguistic resources

To summarize, it is more about the ability to control key elements of an MT solution and the evaluation process than automated metric quality scores.

The factors that matter can vary greatly and buyers should look for MT vendors that can provide a broad range of capabilities to bear and that have a clear enterprise focus.



To learn more about SDL Machine Translation, please visit **sdl.com/mt**



SDL*

SDL (LSE:SDL) is the global leader in content creation, translation and delivery. For over 25 years we've helped companies communicate with confidence and deliver transformative business results by enabling powerful experiences that engage customers across multiple touchpoints worldwide.

Are you in the know? Find out why the top global companies work with and trust **sdl.com**.
Follow us on **Twitter**, **LinkedIn** and **Facebook**.

SDL_eb_MT_Quality_Report_EN_A4_190220