Market Assessment of AI Translated Captions

Independent Technology Assessment

slator

March 5, 2026

Table of Contents

Key Takeaways

1. DeepL Voice leads in both translation quality and caption stability

DeepL Voice achieved the highest scores in both human linguistic evaluation and automated caption stability measurement. DeepL Voice for Zoom achieved a quality score of **96.4/100**, while DeepL Voice for Teams scored **96.3/100**, compared with **87-89** across other platforms. DeepL Voice products also produced the most stable captions, achieving stability scores of **88.6** and **85.8** respectively.

2. DeepL significantly reduces critical translation errors

Across all language pairs, DeepL Voice reduced the rate of critical or major translation errors by **76% on average** compared with other evaluated platforms.

Segments that fully passed linguistic evaluation occurred **79% of the time using DeepL Voice**, compared with **42% across competing tools**.

Across all language pairs, DeepL Voice **reduced** average translation errors per segment by **66% vs. Microsoft Teams and 64% vs. Zoom**.

3. Caption stability varies significantly across platforms

Caption churn — where translations flicker or rewrite repeatedly on screen — was observed across all platforms. However, DeepL Voice products demonstrated the lowest levels of caption churn.

Across all language pairs, **DeepL Voice reduced caption churn by 37.6%** on average compared to **Microsoft Teams**, and **54.7%** on average compared to **Zoom**.

4. Linguists overwhelmingly preferred DeepL Voice

After completing blind evaluations, **96% of linguists** ranked a **DeepL Voice** product as their **preferred platform** for translated captions.

5. DeepL Voice products are Leaders in AI Translated Caption Capabilities

To synthesize results across both linguistic accuracy and caption stability, Slator mapped platforms against two evaluation dimensions:

1. Translation Quality, measured through human linguistic evaluation.
2. Caption Stability, measured through automated frame-level analysis.

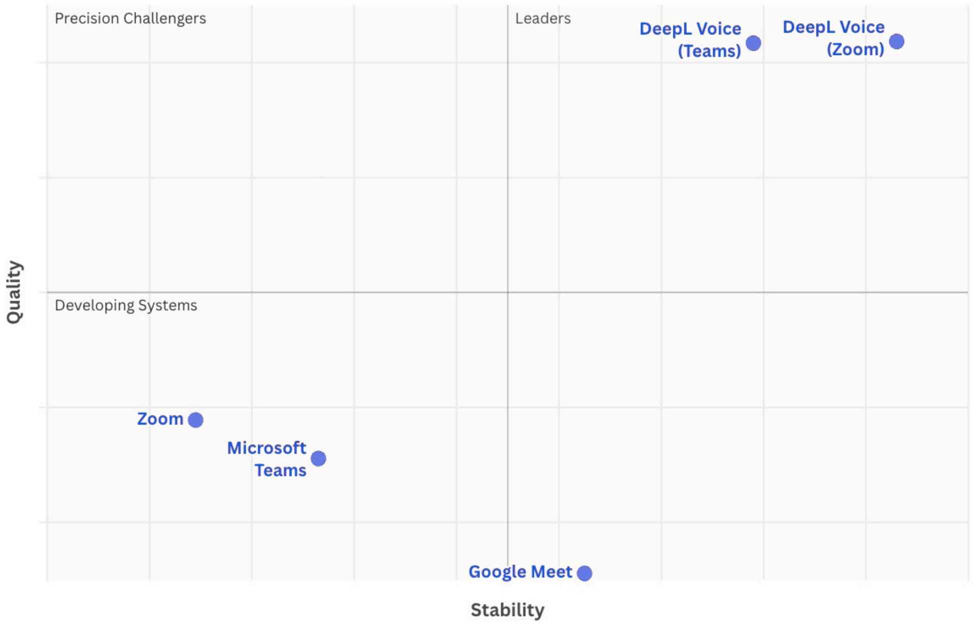
This creates four possible system profiles:

Quadrant	Description
Leaders	High translation quality and high caption stability
Precision Systems	High quality, but lower stability
Efficiency Performers	Lower quality, but relatively stable captions
Developing Systems	Lower performance in both quality and stability

The results position DeepL Voice products in the **Leader** quadrant of this evaluation framework:

Quadrant for AI Translated Captions Platforms

Comparative positioning of real-time caption translation systems based on linguistic quality and caption rendering stability.



Introduction

Translation is increasingly being added as a feature of enterprise tools, and AI generated translated captions have been incorporated into software for global meetings. Real-time caption translation is now available in platforms such as Google Meet, Microsoft Teams, and Zoom to enable communication across languages.

However, despite the growing prevalence of these systems, there is limited independent benchmarking of how well they perform in real-world meeting scenarios. In practice, the user experience of live translated captions depends on two critical dimensions:

- Translation quality: whether captions accurately convey the meaning of the spoken content; and

- Caption stability: whether captions appear consistently on screen without frequent rewriting or flickering as speech is processed.

Frequent caption updates, partial rewrites, or oscillating translations can negatively affect comprehension, even when the final translation is accurate. Measuring both linguistic accuracy and visual stability therefore provides a more complete picture of how real-time captioning systems perform for end users.

To address this gap, Slator conducted an independent evaluation of AI-generated translated captions across five platforms:

1. Google Meet
2. Microsoft Teams
3. Zoom
4. DeepL Voice for Microsoft Teams
5. DeepL Voice for Zoom

The study assessed both translation quality and caption stability across 14 language combinations, covering seven languages translated into English and seven languages translated out of English: Spanish, French, German, Italian, Portuguese, Korean, and Japanese.

This study evaluates the actual captions visible to users on screen, rather than underlying speech recognition transcripts or backend translation outputs. Slator captured and analyzed screen-recorded meetings and extracted captions directly from rendered video frames, allowing the analysis to measure the real user-visible caption experience.

Twenty-eight professional linguists conducted a blind evaluation of translated captions to compare platform performance. Linguists were not informed that customized DeepL Voice systems were included in the evaluation, ensuring that assessments were based solely on caption quality and usability.

The platforms were tested using standard, out-of-the-box caption translation settings in Google Meet, Microsoft Teams, and Zoom. DeepL Voice for Teams and DeepL Voice for Zoom were evaluated using native product capabilities available to end users, including translation glossaries and — in the case of DeepL Voice for Teams — spoken-term recognition features that reinforce transcription of proper nouns and technical terminology.

Audio samples were sourced from podcast recordings featuring two speakers discussing business-related topics in conversational settings. Each recording was edited in length to produce approximately 12 minutes of continuous speech per language.

This approach ensured that the evaluation captured natural dialog patterns, domain-specific terminology, and real-world speech characteristics typical of professional meetings. The evaluation methodology was designed to measure the actual user-visible caption experience across platforms under comparable conditions.

Because linguistic structure varies significantly across languages, some translation directions may naturally produce more interim caption revisions than others. For example, languages with different word-order patterns or sentence-final verbs, such as Japanese or Korean, may require translation systems to adjust captions as additional context becomes available. For this reason, results are analyzed both in aggregate and on a language-by-language basis.

In addition to human linguistic evaluation, Slator developed an automated measurement pipeline to quantify caption stability by analyzing frame-by-frame changes in rendered captions.

The pipeline was as follows: Video recording > Frame extraction (10 fps) > Caption region crop > OCR (Tesseract) > Text normalization > Frame comparison > Change event detection.

Together, these human and automated analyses provide a comprehensive evaluation of how AI-translated captions perform in real-time multilingual meetings.

A full description of the methodology and process can be found in the Appendix.

Translation Quality

Slator commissioned 28 native linguists to evaluate AI-translated captions across 14 language combinations (7 into English, 7 from English) in a blinded human assessment. More details on the methodology can be found in the Appendix.

Across all tested languages and platforms, **DeepL Voice consistently outperformed native captioning tools** in both overall quality and reduction of critical translation errors.

The results are as follows:

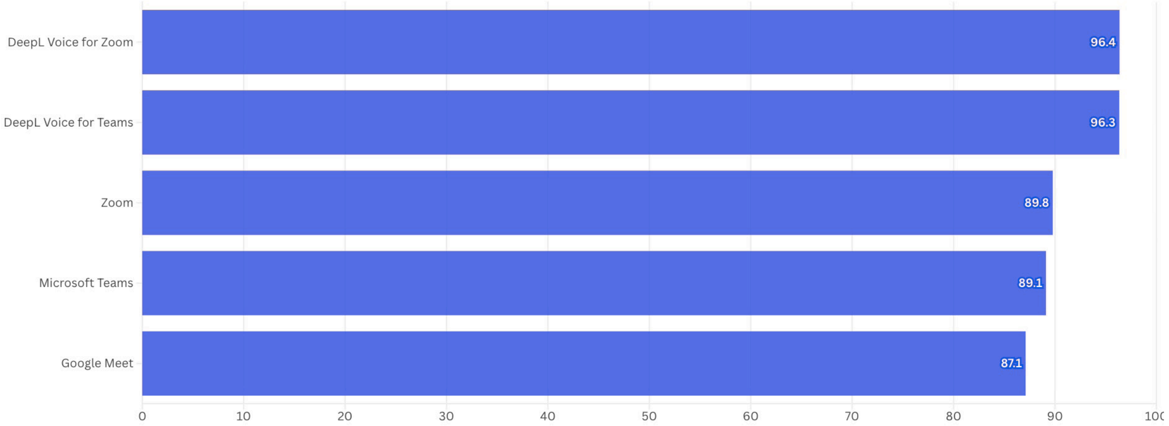
Overall Performance

Slator consolidated results into a single 0–100 Quality Score reflecting the overall severity of translation errors across all evaluated segments. This allowed Slator to answer the question: “How good is the quality of AI translated captions overall across all five platforms?”

The results of the assessment are as follows:

DeepL Voice for Zoom Scored Highest in Human-Led Quality Assessment of AI Translated Captions

Overall quality score (/100) for AI translated captions across all tested language combinations and platforms.



Source: Slator · • The overall quality score was calculated from blind quality assessments of AI translated captions of accuracy and fluency from two native linguists in each language combination. The Quality Score reflects the average severity of errors across all evaluated segments, normalized to a 0–100 scale. Tested language combinations were English into Spanish, French, German, Italian, Portuguese, Korean and Japanese, and vice-versa.

- **DeepL Voice for Zoom** achieved the **highest quality score** in human-led assessments of AI translated captions, with **96.4/100**.
- This was closely followed by **DeepL Voice for Teams** with a score of **96.3/100**.
- Google Meet scored the lowest with 87/100.

Severity Performance

Slator drilled down further into the above-mentioned quality scores to answer the question: “How often does AI translation quality materially affect the end-user understanding of on-screen captions?” This enables us to understand how critical translation errors are within each platform.

Slator tracked the percentage of segments containing critical or major accuracy errors. Linguists were asked to rate the criticality of translation errors according to accuracy (mistranslations, omissions, and additions) and fluency (style, grammar, spelling). This enabled Slator to track the overall **fail rate** of AI translated captions, i.e., the % of segments that had critical or major accuracy errors, resulting in the complete loss or mis-translation of the original meaning.

This provides additional nuance to the quality score provided above, as the fail rate shows how often captions materially distort or lose the meaning of the source audio. Equally the pass rate shows how often each platform accurately conveys the meaning of the source audio, and fluency in the target language.

Below are examples of each category (translations from Google Meet):

Example 1 (Fail):

- Source audio (English): “Let me start by saying that I think when you sit in this position as CEO of a company like Merck, there are many stakeholders who have interests.”
- Target caption (Spanish): “Entonces, permítanme comenzar diciendo que creo que cuando uno se sienta en esta posición, como director ejecutivo de una empresa como Killing, hay muchas partes interesadas que tienen intereses.”
- Error categorization: Mistranslation (Major), Style (Major)
- Linguist comment: A literal translation results in semantic errors, specifically “when you sit in this position,” and awkward language such as “stakeholders who have interests.” “Merck” was interpreted as “Murder” in Spanish.

Example 2 (Pass with Issues):

- Source audio (English): “If you have to pay that much money, all of the things that I just said are not going to make you pleased with the system.”
- Target caption (Spanish): “Y si tienes que pagar tanto dinero, todas las cosas que acabo de decir no te harán sentir satisfecho con el sistema.”
- Error categorization: Style (Major)
- Linguist comment: The AI translated caption switched to the informal you (“tú”). This is an odd translation as natural Spanish uses a neutral third person (“Y si hay que pagar tanto dinero…”).

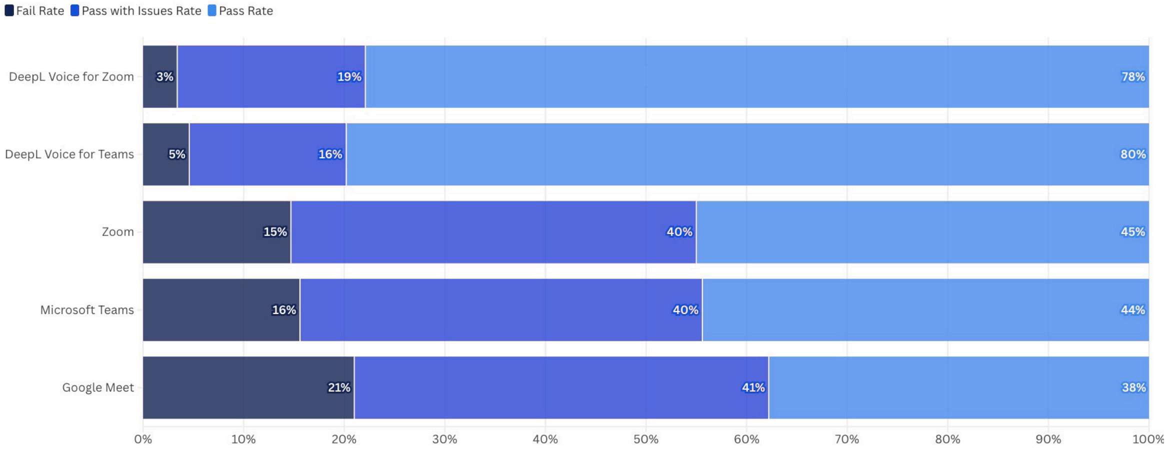
Example 3 (Pass):

- Source audio (English): “I think it’s my job to ensure that the company functions in a sustainable way to create long term value for all of its stakeholders, including its shareholders.”
- Target caption (Spanish): “Creo que mi trabajo es garantizar que la empresa funcione de manera sostenible para crear valor a largo plazo para todas sus partes interesadas, incluidos los accionistas.”
- Error categorization: N/A
- Linguist comment: “The Spanish in this segment reflects the English as spoken.”

The results of this analysis are as follows:

DeepL Voice Reduces Critical Errors by 13% on Average, has Average Pass Rate of 79%

Overall fail / pass with issues / fail rate for AI translated captions across all tested language combinations and platforms.



Source: Slator • Rates were calculated from blind quality assessments of AI translated captions of accuracy and fluency from two native linguists in each language combination. Rates represent the severity of translation errors across all evaluated segments. Tested language combinations were English into Spanish, French, German, Italian, Portuguese, Korean and Japanese, and vice-versa.

- AI translated captions on **DeepL Voice products** produced, on average, a **4% fail rate**, compared to **17%** on average across all other tools. This represents a **76% reduction in critical or major accuracy errors using DeepL Voice products.**
- DeepL Voice products produced, on average, a **79% pass rate** on all AI translated segments, compared to 42% on average across all other tools. This represents an **88% relative increase** in fully passing segments compared to other market tools.
- Approximately 60% of segments on Google Meet, and half of segments on Zoom and Microsoft Teams contained either major errors or material translation issues.

Language-by-Language Performance

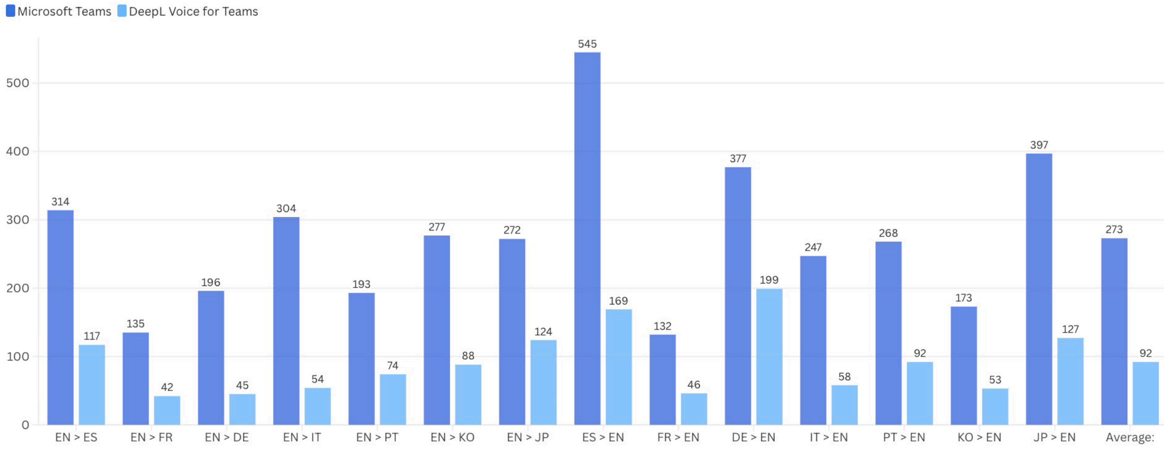
Slator drilled down further to understand the language differences between out-of-the-box platforms (Microsoft Teams, Zoom) and custom platforms (DeepL Voice for Teams, and DeepL Voice for Zoom), to answer the question: “How much better or worse is DeepL Voice compared to out-of-the-box platforms?”

Across all language pairs, **DeepL Voice reduced** average translation errors per segment by **66%** on average compared to **Microsoft Teams**, and **64%** compared to **Zoom**.

When comparing Microsoft Teams and DeepL Voice for Microsoft Teams:

Using DeepL Voice for Teams Reduces the Average Caption Error Rate by 66%

A comparison of the error rate in AI translated captions across language combinations between Microsoft Teams, and DeepL Voice for Teams.



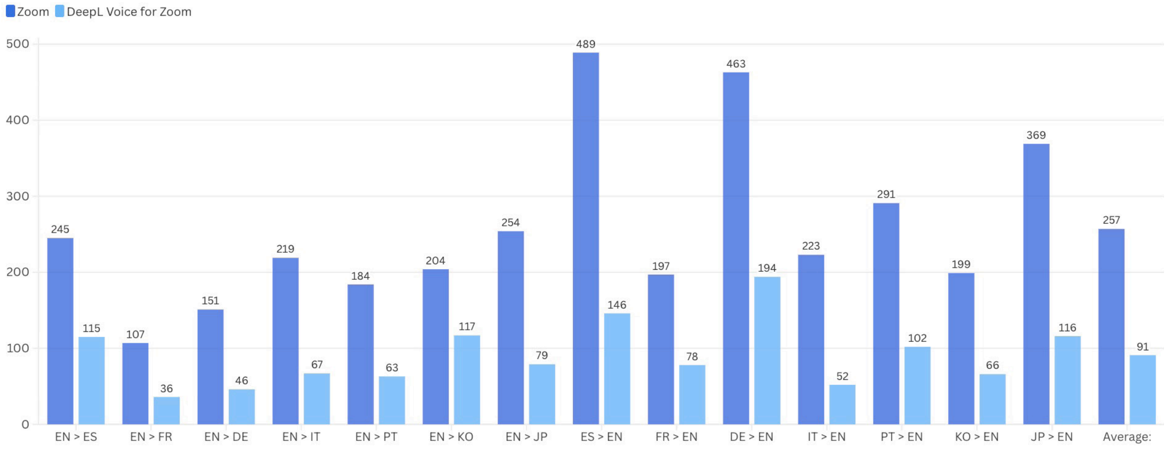
Source: Slator • Lower numbers represent higher quality. Translated captions were reviewed by two linguists per language combination, and scored according to accuracy and fluency. As part of the scoring fluency errors were weighted 0.5x.

- The **greatest reductions** in average translation errors per segment with **DeepL Voice for Teams** were in:
  - English to Italian (82%)
  - English to German (77%)
  - Italian to English (76%)
- Only one language pair dropped below a 50% improvement when using DeepL Voice for Teams (German to English, 47%).

When comparing Zoom and DeepL Voice for Zoom:

Using DeepL Voice for Zoom Reduces the Average Caption Error Rate by 65%

A comparison of the error rate in AI translated captions across language combinations between Zoom, and DeepL Voice for Zoom.



Source: Slator • Lower numbers represent higher quality. Translated captions were reviewed by two linguists per language combination, and scored according to accuracy and fluency. As part of the scoring, fluency errors were weighted 0.5x.

- The **greatest reductions** in average translation errors per segment with **DeepL Voice for Zoom** were in:
  - Italian to English (77%)
  - Spanish to English (70%)
  - English to German (70%)
- Only one language pair dropped below a 50% improvement when using DeepL Voice for Zoom (English to Korean, 43%).

Linguist Preferences

After completing quality assessments, Slator asked all 28 participating linguists to rank their preferred platform for AI translated captions based on user experience. These linguists carried out their analysis blindly, and provided the name of their preferred platform based on a code name for each platform.

The results are as follows:

- **52%** of linguists (15) ranked **DeepL Voice for Teams** as their preferred platform.
- **44%** of linguists (12) ranked **DeepL Voice for Zoom** as their preferred platform.
- All in all, **96%** (27) of linguists ranked either one of DeepL Voice for Teams or DeepL Voice for Zoom as their **top-two preference.**
- Zoom typically ranked third, but received one vote as the most preferred tool. Microsoft Teams and Google Meet received no votes in the top two positions, and consistently appeared in 4th or 5th place.

Caption Stability

Slator automatically measured the number of change events across all languages and platforms. Below are three examples of these changes taken from the sampled audio files, to demonstrate how captions rendered before fully stabilizing:

Example 1 (Incremental Sentence Growth, French to English):

- Frame 1: “Hello Nathalie the topic of the”
- Frame 2: “Hello Nathalie the topic of the day is the”
- Frame 3: “Hello Nathalie the topic of the day is the subject of pharmacy”
- Frame 4 (stable): “Hello Nathalie the topic of the day is the subject of pharmacy groups.”

In this example, the live captions display partial translations before the speaker has completed a sentence. As additional words are processed, captions are extended until the full sentence stabilizes.

Example 2 (Caption Re-Writing, Spanish to English):

- Frame 1: “When you make a demand what is the minimum economic amount”
- Frame 2: “When you make a claim what is the minimum economic amount”
- Frame 3: “When you make a claim what is the minimum economic amount so that”
- Frame 4 (stable): “When you make a claim what is the minimum economic amount so that you can file a case?”

In this example, the live captions revise previously displayed words as additional linguistic context becomes available.

Example 3 (Caption Flicker, Korean to English):

- Frame 1: “the fourth quarter ended about 10 days”
- Frame 2: “the fourth quarter is now about 10 days”
- Frame 3 (stable): “the fourth quarter ended about 10 days ago”

In this example, the live captions completely revise previously displayed words before changing again to another final rendition (A>B>A, A>B>C, or similar variations in behavior).

Example 4 (High Stability, Korean to English):

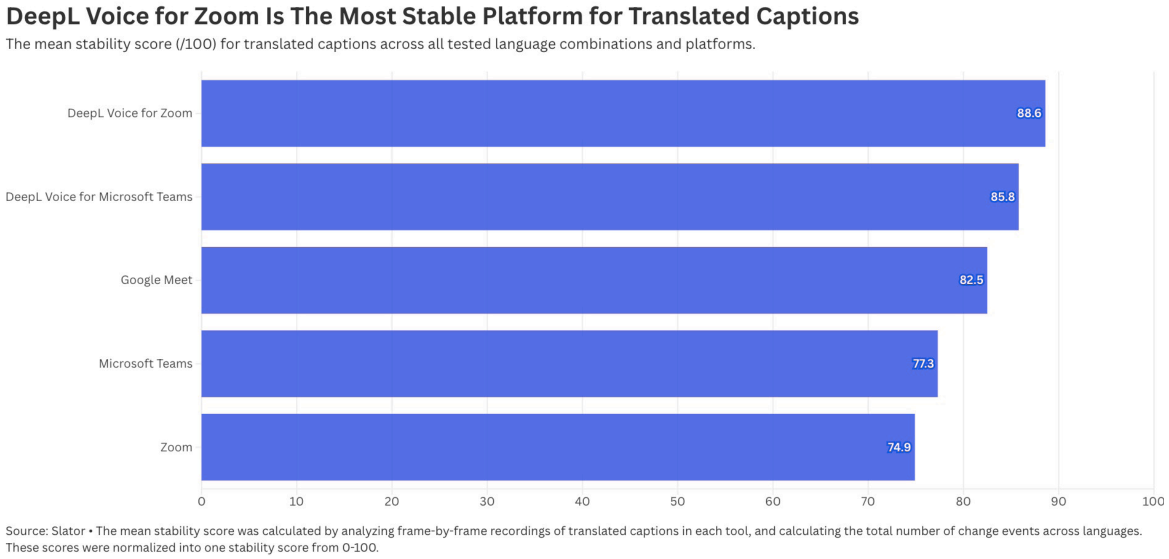
- Frame 1: “Let’s talk about Samsung Electronics’ performance.”
- Frame 2: “Let’s talk about Samsung Electronics’ performance.”
- Frame 3: “Let’s talk about Samsung Electronics’ performance.”

In this example, the final translated caption was displayed and unchanged over time.

Overall Performance

Slator consolidated results into a single 0–100 Quality Score reflecting the overall stability of translated captions across all platforms. This allowed Slator to answer the question: “To what extent does each platform flicker or update the translations until the captions fully stabilize?”

The results of the assessment are as follows:



- **DeepL Voice for Zoom** achieved the **highest score** in an automated assessment of AI translated caption stability, with **88.6/100**.
- This was closely followed by **DeepL Voice for Teams** with a score of **85.8/100**.
- Zoom scored the lowest with 74.9/100.

Language-by-Language Performance

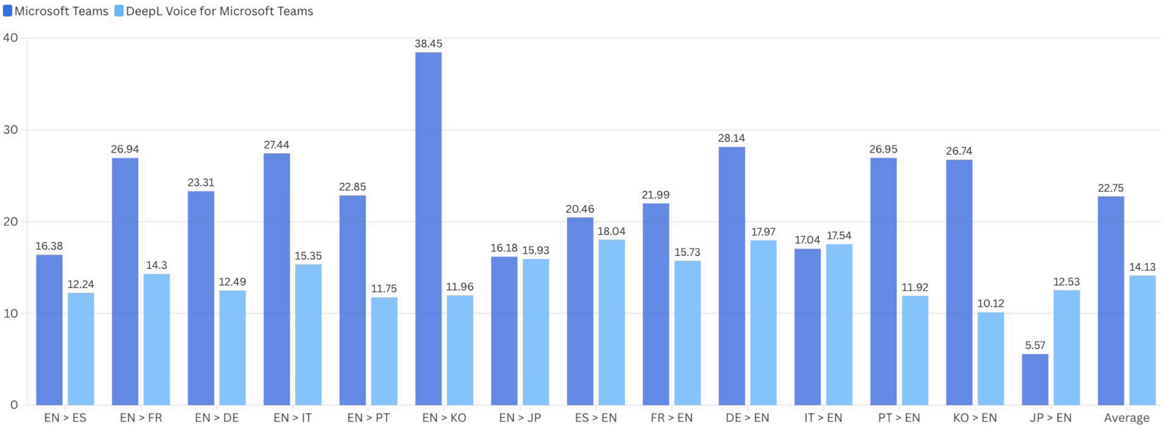
Slator drilled down further to understand the language differences between out-of-the-box platforms (Microsoft Teams, Zoom) and custom platforms (DeepL Voice for Teams, and DeepL Voice for Zoom), to answer the question: “Are DeepL Voice products more or less stable on a language-by-language basis compared to out-of-the-box platforms?”

Across all language pairs, **DeepL Voice reduced caption churn by 37.6%** on average compared to **Microsoft Teams**, and **54.7%** compared to **Zoom**.

When comparing Microsoft Teams and DeepL Voice for Microsoft Teams:

Using DeepL Voice for Teams Improves Stability by 38% on Average

A comparison of the caption churn rate in AI translated captions across language combinations between Microsoft Teams, and DeepL Voice for Teams.



Source: Slator • The caption churn rate is defined as the percentage of frames in which a displayed caption changed. This directly quantifies how often users experience caption updates or disruptions. Figures were calculated automatically using Python.

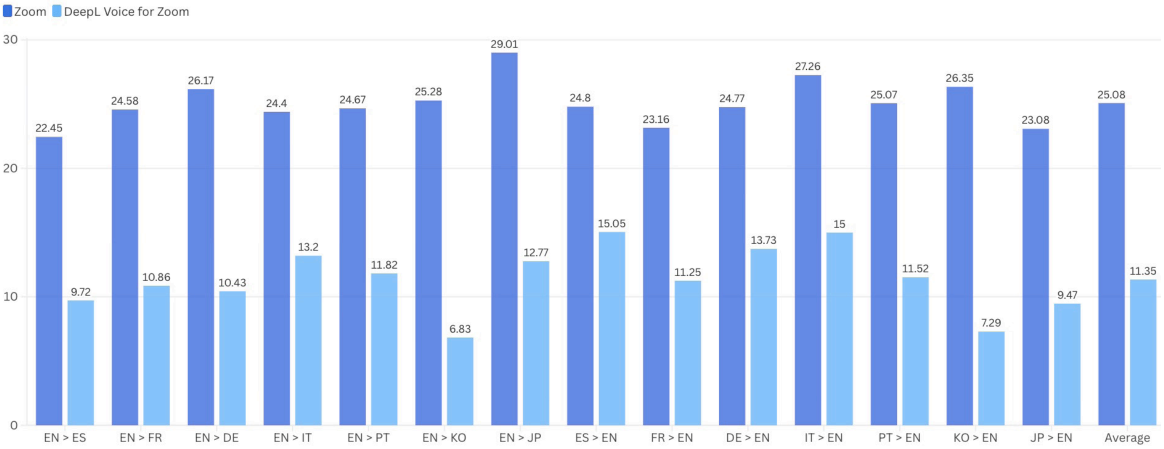
- The **greatest improvements** in translation caption stability with **DeepL Voice for Teams** were in:
  - English to Korean (+69%)
  - Korean to English (+62%)
  - Portuguese to English (+56%)

- Only two language pairs were more stable in Microsoft Teams’ AI translated captions — Italian to English (3% more stable in Microsoft Teams), and Japanese to English (125% more stable in Microsoft Teams).

When comparing Zoom and DeepL Voice for Zoom:

Using DeepL Voice for Zoom Improves Stability by 55% on Average

A comparison of the caption churn rate in AI translated captions across language combinations between Zoom, and DeepL Voice for Zoom.



Source: Slator • The caption churn rate is defined as the percentage of frames in which a displayed caption changed. This directly quantifies how often users experience caption updates or disruptions. Figures were calculated automatically using Python.

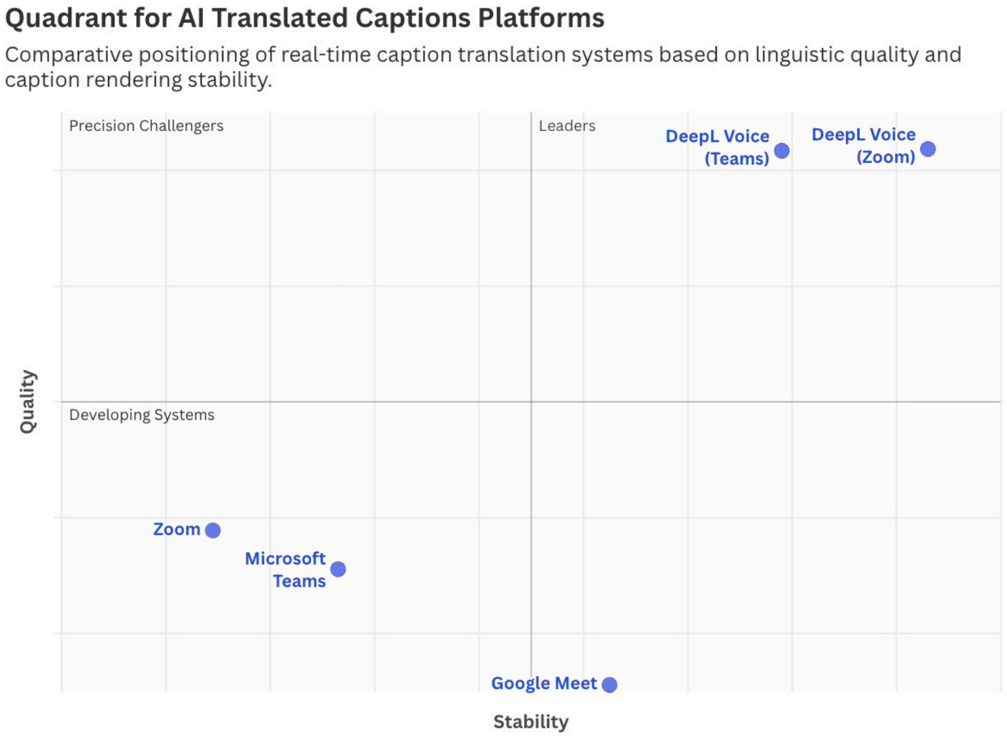
- The **greatest improvements** in translation caption stability with **DeepL Voice for Zoom** were in:
    - English to Korean (73%)
    - Korean to English (72%)
    - English to German (60%)

- No language pairs were less stable in DeepL Voice for Zoom when compared to Zoom.

Conclusion

Overall Results

Below is the overall ranking of all five AI translation caption platforms:



DeepL Voice for Teams and **DeepL Voice for Zoom** achieved high scores in both quality and stability, positioning these platforms as **Leaders** for **AI translated caption solutions**. Real-time captioning systems may face a trade-off between translation accuracy and caption stability, by providing value in high accuracy translations while limiting the latency in rendering these translations. However, the results of this study suggest that DeepL Voice improves both simultaneously.

Both **Microsoft Teams** and **Zoom** are positioned as **Developing Systems**, having achieved below-average performance in both caption rendering and translation quality. Microsoft Teams performed marginally better in terms of stability, and Zoom performed marginally better in terms of quality, however both tools would require improvements in each category to improve its existing offering.

Google Meet, while relatively stable compared to other market tools, achieved the lowest quality score as identified by Slator's pool of expert linguists. This positions Google Meet as an **Efficiency Performer** in the overall ranking, and could become a Leader if the platform improves quality across language combinations while maintaining or improving its caption rendering.

Implications for Enterprises

Slator concludes the following implications for enterprises upon completing this assessment:

1. **AI caption quality is improving but remains uneven across platforms**
The results show significant variation in translation quality across commonly used meeting platforms. Organizations relying on built-in caption translation may encounter meaningful differences in translation accuracy depending on the tool used.

2. **Caption stability directly affects usability**
Even when translations are accurate, frequent caption rewrites can disrupt comprehension. Measuring caption stability provides a useful proxy for real-world user experience in multilingual meetings.

3. **Specialized platforms with customization can improve translation outcomes**
The results suggest that using specialized AI translation tools, and applying domain-specific glossaries and transcription reinforcement can substantially improve translation performance in real-time AI captions across languages.

Appendix

Methodology

Sourcing Audio Samples

Slator sourced podcast recordings in each language in scope. Podcast recordings were selected because they provide consistent audio quality across languages, enabling a controlled comparison of translation performance rather than differences caused by background noise or microphone quality.

Each podcast featured two speakers engaging in a back-and-forth conversation on business-related, regulated topics (finance, life sciences, legal), with minimal to zero background noise. This approach set up the experiment to measure the translation quality of specialized terms, and to assess how the platforms handle natural conversations between native speakers of the source language.

Each podcast was edited to create an isolated snippet of source audio lasting approximately 12 minutes per language. This source audio file was processed through an automated transcription tool to generate a source audio transcript.

Technical Setup

Slator set up a virtual cable to route the source audio to each platform (Google Meet, Microsoft Teams, Zoom). This ensured that no additional or unintentional background noise was added during the recording process. Slator used a screen capture tool to record an .mp4 video file of the translated captions in each platform, showing live captions exactly as a user would see them.

Standard, out-of-the-box translated captions settings were used for Google Meet, Microsoft Teams, and Zoom. Specific to Google Meet, English into Italian was identified as Beta.

When testing DeepL for Voice in Microsoft Teams and Zoom, Slator applied a glossary of terms, and configured an informal voice from DeepL's formality settings given the familiar nature of the podcast conversations.

In addition, Slator switched on DeepL Voice's Spoken Terms feature — available in Microsoft Teams only. This feature is a monolingual list of spoken terms (i.e., acronyms or proper nouns) that reinforces the tool's ability to recognize monolingual terms for the speech-to-text transcription process.

Slator processed all recordings through an automated processing engine (as described in Automated Quality Assessments), and shared these recordings with linguists.

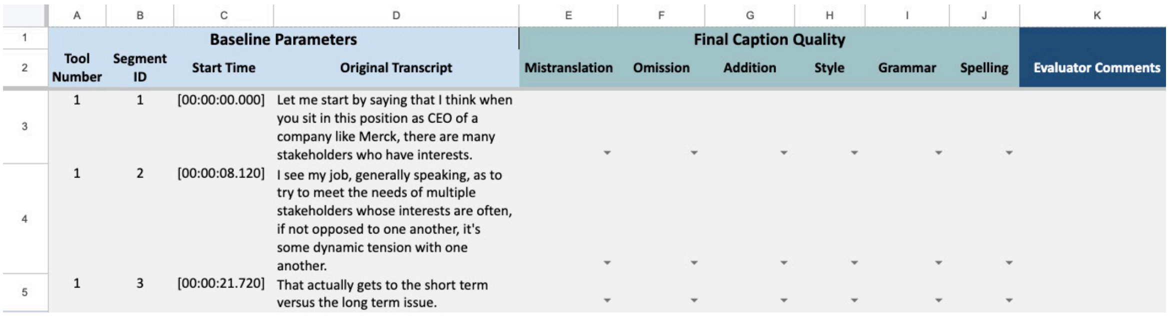
Sourcing Linguists

Slator sourced two linguists for each language combination, ensuring that each linguist was a native speaker or near-native of the target language. Sourced linguists were hand-picked according to subject matter expertise related to the audio files, and are qualified linguists.

Selected linguists primarily had experience in media localization, translation, interpreting, and quality assessments (including AI training / AI assessments). This mix of expertise reinforced the need to listen to the audio file, while simultaneously assessing the quality of on-screen captions.

Human-Led Quality Assessments

Slator created a Quality Assessment Scorecard for each source audio file. Below is a sample:



Baseline Parameters				Final Caption Quality						
Tool Number	Segment ID	Start Time	Original Transcript	Mistranslation	Omission	Addition	Style	Grammar	Spelling	Evaluator Comments
1	1	[00:00:00.000]	Let me start by saying that I think when you sit in this position as CEO of a company like Merck, there are many stakeholders who have interests.							
1	2	[00:00:08.120]	I see my job, generally speaking, as to try to meet the needs of multiple stakeholders whose interests are often, if not opposed to one another, it's some dynamic tension with one another.							
1	3	[00:00:21.720]	That actually gets to the short term versus the long term issue.							

The scorecard contains:

- Baseline Parameters:
    - Tool number. This is an anonymized reference to each tool / platform, to prevent linguists from knowing the name of the platform that they were evaluating, and to prevent linguists from knowing which platform may or may not be customized. Linguists were not informed that this assessment included customized models from DeepL.
    - Segment ID. This is the segment ID of the source audio transcript. This enabled Slator to ground the assessment into segmented chunks of text, for clearer analysis across multiple linguists within the same language combination, and across languages and platforms.
    - Start time. This is grounded in the source transcript, and enabled linguists to easily navigate to specific segments as needed in the source audio file.
    - Original Transcript. This is the original source transcript. Linguists were informed that this may contain errors. Slator was intentional in not cleaning errors in the

transcript, to prevent linguists from assessing the translated captions against the written transcript and instead force linguists to listen to the source audio. Likewise, Slator did not provide a written transcript of translated captions for the same reasons.

- Final Caption Quality:
  - Accuracy Categories:
    - Mistranslation. Slator defined this category as “The translated caption meaning is incorrect or misleading.” Possible categorizations were defined as follows:
      - None – No noticeable issue.
      - Minor – Meaning mostly correct, nuance loss only
      - Major – Meaning partially distorted
      - Critical – Meaning is materially wrong

    - Omission. Slator defined this category as “Important information is missing from the translated caption.” Possible categorizations were defined as follows:
      - None – No noticeable issue.
      - Minor – Meaning mostly correct, nuance loss only
      - Major – Meaning partially distorted
      - Critical – Meaning is materially wrong. Linguists were instructed to categorize segments as Critical if the platform did not translate the segment or displayed the source language in the translated caption.

    - Addition. Slator defined this category as “Information has been added to the translated caption that is not in the audio.” Possible categorizations were defined as follows:
      - None – No noticeable issue.
      - Minor – Meaning mostly correct, nuance loss only
      - Major – Meaning partially distorted
      - Critical – Meaning is materially wrong

  - Fluency Categories:
    - Style. Slator defined this category as “Caption style or register feels inappropriate for live captions. Examples: the captions are overly verbose, too formal / too casual, reads like written prose, etc.” Possible categorizations were defined as follows:
      - None – No noticeable issue.
      - Minor – Noticeable but easy to read
      - Major – Distracting or reduces readability

    - Grammar. Slator defined this category as “Sentence violates grammatical rules of the target language. Examples: agreement errors, wrong tense or conjugation." Possible categorizations were defined as follows:

- None – No noticeable issue.
- Minor – Noticeable but easy to read
- Major – Distracting or reduces readability

- Spelling. Slator defined this category as "Typos or orthographic errors that affect reading. Examples: misspellings, incorrect diacritics, wrong capitalization." Possible categorizations were defined as follows:
    - None – No noticeable issue.
    - Minor – Noticeable but easy to read
    - Major – Distracting or reduces readability

- Evaluator Comments. Slator asked linguists to provide justification for their ratings in English. This enabled Slator to carry out cross-checks across languages and platforms, to check that instructions were correctly followed and ratings are consistent. This also encouraged linguists to be intentional about their ratings and to reinforce consistency during evaluations. This also provided qualitative feedback on the translated captions across languages and platforms.

Linguists were instructed to complete the Quality Assessment Scorecard based on the finalized / stabilized translated captions only.

In addition to the above, Slator also asked linguists to rate the platforms in order of user experience / preference and quality. Slator used these ratings to draw conclusions on which platform is preferred by linguists, and which platform has the worst perceived user experience.

Scoring Human-Led Assessments

Data Cleaning & Weighting

Because each language pair was evaluated by two independent linguists, Slator reviewed the assessments for consistency and outliers before aggregating results. Scores were averaged across evaluators to produce mean segment scores per platform and language pair. While minor variation in severity ratings is expected in human linguistic evaluation, cross-review confirmed a high degree of consistency in error classification across evaluators.

Slator converted the quality assessments into numerical values (None [0 points], Minor [1 point], Major [3 points], Critical [5 points]).

Categories relating to Accuracy (i.e., Mistranslation errors, Omission errors, and Addition errors) were given a Total Accuracy Score.

Categories associated with Fluency (i.e., Style errors, Spelling errors, and Grammar errors) were also given a Total Fluency Score.

These two Accuracy and Fluency scores were combined to produce a Total Segment Score. Scores across different linguists in the same language combination were combined and averaged to produce mean scores across Accuracy, Fluency, and the Total Segment.

As part of the Total Segment Score, the Fluency score carried a 0.5x weighting. This prevented a segment from producing a "Fail" result based on one major spelling, grammar, or style error, and weighted the overall segment score in favor of accuracy issues, such as mistranslations or omissions, which carry a lower perceived quality among linguists.

Data Categorization

The Total Segment Score also indicated whether each segment had "Pass", "Pass with Issues", or "Fail" rates. This enabled Slator to calculate overall Fail rates across platforms and languages.

Given the weighting towards accuracy, as described above, a segment is automatically classified as a "Fail" if any of the following conditions are met:

- An Accuracy-based segment (Mistranslations, Omission, or Addition) is marked as Critical;
- An Accuracy-based segment had two or more "Major" errors (e.g., a major omission, and a major addition);
- A Mistranslation is marked as Major (or Critical, as per point 1)

A segment is classed as "Pass with Issues" if the following conditions are met:

- A Fluency-based segment (Spelling, Grammar, Style) is marked as Major, or;
- An Accuracy-based segment is marked as Major (having not triggered a Fail classification), or;
- There are more than three minor errors across all categories

If none of these conditions are met, the segment is a Pass. If there are only one or two minor errors, the segment meets the conditions to pass.

Data Scoring

Total Segment Scores were calculated for each platform and language combination, enabling a direct comparison between these.

The weighted, Total Segment Scores are as follows (the lower the number, the better the score):

Language Combination	Google Meet	Microsoft Teams	Zoom	DeepL Voice for Teams	DeepL Voice for Zoom
EN > ES	288	314	245	117	115
EN > FR	164	135	107	42	36
EN > DE	179	196	151	45	46
EN > IT	252	304	219	54	67
EN > PT	210	193	184	74	63
EN > KO	237	277	204	88	117
EN > JP	291	272	254	124	79
ES > EN	531	545	489	169	146
FR > EN	274	132	197	46	78
DE > EN	408	377	463	199	194
IT > EN	311	247	223	58	52
PT > EN	390	268	291	92	102
KO > EN	329	173	199	53	66
JP > EN	671	397	369	127	116

Fail / Pass rates were calculated by adding the total number of segments in each category (Fail / Pass with Issues / Pass), as follows:

Tool	Fail Rate	Pass with Issues Rate	Pass Rate
Google Meet	307	603	555
Microsoft Teams	229	586	650
Zoom	216	590	659
DeepL Voice for Teams	67	228	1170
DeepL Voice for Zoom	50	274	1141

These figures were divided into the total number of segments across all languages (1465) to generate the rates as overall percentages:

Tool	Fail Rate	Pass with Issues Rate	Pass Rate
Google Meet	21.0%	41.2%	37.9%
Microsoft Teams	15.6%	40.0%	44.4%
Zoom	14.7%	40.3%	45.0%
DeepL Voice for Teams	4.6%	15.6%	79.9%
DeepL Voice for Zoom	3.4%	18.7%	77.9%

Data Processing

To enable clear comparison across tools and language directions, we converted Total Segment Scores into a normalized 0–100 Quality Score. For each tool, we first calculated the mean penalty per segment, defined as the Total Segment Scores across all evaluated segments (across all language pairs and both directions) divided by the total number of segments assessed.

Error points were assigned according to the predefined scoring framework (Accuracy and Fluency categories with severity-based weights. This means the theoretical maximum possible penalty per segment is 24 points — 15 from Accuracy categories and 9 from Fluency categories). The normalized score was then calculated as: 100 x (1-Mean penalty per segment/24).

Under this formulation, a score of 100 represents a system with no observed errors. A score of zero would represent the theoretical extreme in which every segment incurred the maximum possible penalty.

The data is as follows:

Tool	Total Weighted Score	Total Segments	Mean Penalty	Normalized 0-100 Score
Google Meet	4534	1465	3.09	87.10537543
Microsoft Teams	3828	1465	2.61	89.11333902
Zoom	3592	1465	2.45	89.78313424
DeepL Voice for Teams	1286	1465	0.88	96.34172355
DeepL Voice for Zoom	1277	1465	0.87	96.36732082

Observed tool scores (87–96) reflect performance relative to this maximum theoretical error ceiling, ensuring comparability across tools, languages, and future evaluations.

Automated Quality Assessments

We built an automated measurement pipeline that quantifies how stable or unstable live captions appear on screen over time across Google Meet, Microsoft Teams, Zoom, DeepL Voice for Microsoft Teams, and DeepL Voice for Zoom, and across multiple languages using language-specific OCR.

In short: the system takes all screen-recorded meeting videos (one per language and platform), extracts the caption region from each frame, runs OCR on that region to recover the text that a viewer would have seen at that moment, then computes caption stability metrics and writes a summary report.

The purpose is to measure the actual user-visible caption experience, not the underlying ASR transcript. If captions flicker, rewrite themselves, or oscillate, Slator captures that directly from the rendered video frames.

Specifically, Slator extracted approximately 10 frames per second for each recorded file, saving these as image files and enabling a frame-by-frame analysis of what was shown on screen across languages and platforms, accounting for rapid re-writes that occur at fractions of a second.

For each extracted frame, Slator's system cropped a rectangle that only contains the caption overlay area, to reduce unrelated screen content.

Slator leveraged Tesseract — an OCR engine that recognizes each caption language — that runs on every cropped caption image, extracting the visible caption text, saving one record per frame.

Raw OCR output can vary due to spacing, Unicode variants, and invisible characters. Before comparing frames, the system normalizes text in a controlled way so that comparisons represent real visible changes rather than OCR noise.

Normalization includes:

- Unicode normalization
- Removing timestamps
- Removing speaker names
- Removing UI text (e.g., "Translated captions are on")
- Collapsing repeated whitespace
- Trimming leading/trailing spaces
- Removing zero-width characters
- Language-group-specific handling (Latin vs CJK)

This helps ensure "changes" represent actual user-visible rewrites as much as possible, while still preserving meaningful differences.

The system reads the frame-by-frame OCR results and compares text over time to detect when captions update, how often they rewrite, and whether they flicker / oscillate.

The pipeline writes a summary CSV with key metrics per video, i.e., numbers of active frames and numbers of change events (see Scoring Automated Assessments below). These results are designed to allow direct comparison across platforms, across languages, and across test conditions (DeepL vs. not DeepL).

Scoring Automated Assessments

To measure caption stability objectively, Slator measured how often captions changed from one frame to the next. Any visible modification, including word additions, re-segmentation, or brief oscillations / flickers, was counted as a change event. Specifically, change events captured:

- Character additions or deletions
- Word completions
- Minor substitutions
- Punctuation adjustments
- Formatting changes
- Back-and-forth oscillation
- Text instability, and
- Perceptually disruptive flickers

The proportion of frames in which no visible change occurred was converted into a 0-100 Stability Score, where 100 represents completely stable captions (no frame-to-frame changes) and lower scores indicate more frequent on-screen updates.

In parallel, we calculated caption churn, defined as the percentage of frames in which the displayed caption changed. While the Stability Score highlights the proportion of visually stable frames, churn directly quantifies how often users experience caption updates or disruptions.

Together, the Stability Score and churn provide a transparent and behavior-based framework for comparing caption performance across platforms. They do not measure linguistic accuracy, translation quality, or grammar or semantic correctness (this was evaluated separately through human linguistic review).

Raw Stability Data

The raw data is as follows:

Cross-Tool Stability Rankings

This data enabled Slator to conclude the % increase or decrease of using DeepL Voice compared to the same platform without DeepL Voice.

Tool	0-100 Mean Stability Score	Churn
DeepL Voice for Zoom	88.6	11.35%
DeepL Voice for Microsoft Teams	85.8	14.20%
Google Meet	82.5	17.50%
Microsoft Teams	77.3	22.75%
Zoom	74.9	25.08%

Percentage improvements were calculated based on reductions in caption churn between DeepL Voice products and the corresponding native platform.

Cross-Language Stability Rankings

By language, the mean stability score results are as follows:

Language Combination	DeepL Voice for Zoom	DeepL Voice for Microsoft Teams	Google Meet	Microsoft Teams	Zoom
EN > ES	90.3	87.8	87.5	83.6	77.6
EN > FR	89.1	85.7	70.5	73.1	75.4
EN > DE	89.6	87.5	77.5	76.7	73.8
EN > IT	86.8	84.7	85.4	72.6	75.6
EN > PT	88.2	88.2	91.4	77.2	75.3
EN > KO	93.2	88.0	80.2	61.5	74.7
EN > JP	87.2	84.1	67.2	83.8	71.0
ES > EN	84.9	82.0	88.7	79.5	75.2
FR > EN	88.8	84.3	89.4	78.0	76.8
DE > EN	86.3	82.0	75.5	71.9	75.2
IT > EN	85.0	82.5	88.4	83.0	72.7
PT > EN	88.5	88.1	78.6	73.1	74.9
KO > EN	92.7	89.9	87.9	73.3	73.7
JP > EN	90.5	87.5	86.7	94.4	76.9
Average	*88.6*	*85.9*	*82.5*	*77.3*	*74.9*

The churn — i.e., the percentage of frames in which the displayed caption changed, with differences between DeepL Voice products and off-the-shelf products — are as follows:

Language Combination	DeepL Voice for Zoom	DeepL Voice for Microsoft Teams	Google Meet	Microsoft Teams	Zoom	Difference DeepL Voice for Teams vs. Teams	Difference DeepL Voice for Zoom vs. Zoom
EN > ES	9.72%	12.24%	12.53%	16.38%	22.45%	25%	57%
EN > FR	10.86%	14.30%	29.45%	26.94%	24.58%	47%	56%
EN > DE	10.43%	12.49%	22.54%	23.31%	26.17%	46%	60%
EN > IT	13.20%	15.35%	14.63%	27.44%	24.40%	44%	46%

EN > PT	11.82%	11.75%	8.57%	22.85%	24.67%	49%	52%
EN > KO	6.83%	11.96%	19.76%	38.45%	25.28%	69%	73%
EN > JP	12.77%	15.93%	32.78%	16.18%	29.01%	2%	56%
ES > EN	15.05%	18.04%	11.32%	20.46%	24.80%	12%	39%
FR > EN	11.25%	15.73%	10.57%	21.99%	23.16%	28%	51%
DE > EN	13.73%	17.97%	24.53%	28.14%	24.77%	36%	45%
IT > EN	15.00%	17.54%	11.58%	17.04%	27.26%	-3%	45%
PT > EN	11.52%	11.92%	21.39%	26.95%	25.07%	56%	54%
KO > EN	7.29%	10.12%	12.11%	26.74%	26.35%	62%	72%
JP > EN	9.47%	12.53%	13.30%	5.57%	23.08%	-125%	59%
Average	*11.35%*	*14.13%*	*17.50%*	*22.75%*	*25.08%*	*38%*	*55%*

In the final two columns, positive percentages represent the improvement in using DeepL Voice products compared to the out-of-the-box platform. Negative percentages represent more stability in the out-of-the-box platform compared to DeepL Voice products.

Full Dataset

Below is the full dataset for all tools and languages.

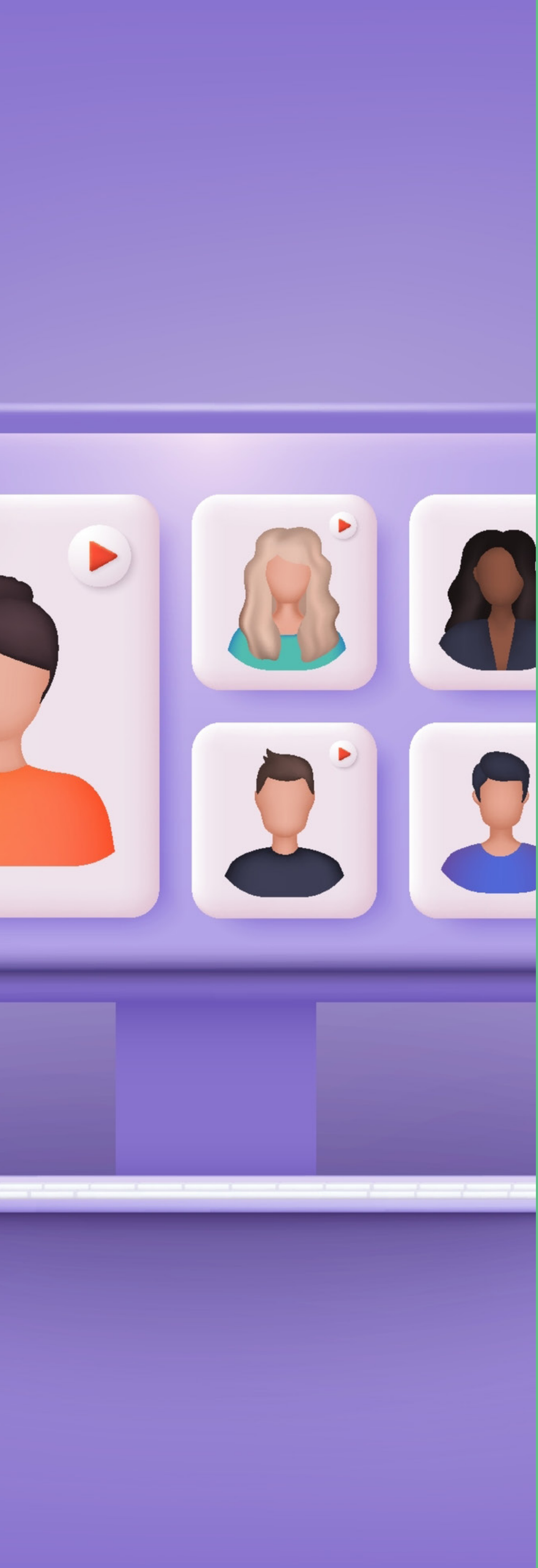
Tool	Language Combination	Active Frames	Frames per Second	Total Change Events
DeepL Voice for Teams	DE > EN	8987	11	1615
DeepL Voice for Teams	EN > DE	7212	11	901
DeepL Voice for Teams	EN > ES	7167	11	877
DeepL Voice for Teams	EN > FR	7099	10	1015
DeepL Voice for Teams	EN > IT	7005	10	1075
DeepL Voice for Teams	EN > JP	6957	10	1108
DeepL Voice for Teams	EN > KO	7021	10	840
DeepL Voice for Teams	EN > PT	7249	11	852
DeepL Voice for Teams	ES > EN	7057	10	1273
DeepL Voice for Teams	FR > EN	8123	11	1278
DeepL Voice for Teams	IT > EN	7214	10	1265
DeepL Voice for Teams	JP > EN	5921	10	742
DeepL Voice for Teams	KO > EN	5285	10	535
DeepL Voice for Teams	PT > EN	6360	10	758
DeepL Voice for Zoom	DE > EN	8766	11	1204
DeepL Voice for Zoom	EN > DE	7112	10	742
DeepL Voice for Zoom	EN > ES	7246	11	704
DeepL Voice for Zoom	EN > FR	7192	11	781
DeepL Voice for Zoom	EN > IT	7157	11	945
DeepL Voice for Zoom	EN > JP	7241	11	925
DeepL Voice for Zoom	EN > KO	7272	11	497
DeepL Voice for Zoom	EN > PT	7239	11	856

DeepL Voice for Zoom	ES > EN	7268	11	1094
DeepL Voice for Zoom	FR > EN	7923	10	891
DeepL Voice for Zoom	IT > EN	7299	10	1095
DeepL Voice for Zoom	JP > EN	6117	11	579
DeepL Voice for Zoom	KO > EN	5380	11	392
DeepL Voice for Zoom	PT > EN	6258	10	721
Google Meet	DE > EN	8512	10	2088
Google Meet	EN > DE	6185	9	1394
Google Meet	EN > ES	6712	10	841
Google Meet	EN > FR	6183	9	1821
Google Meet	EN > IT	7141	11	1045
Google Meet	EN > JP	6977	10	2287
Google Meet	EN > KO	7139	11	1411
Google Meet	EN > PT	7097	10	608
Google Meet	ES > EN	7237	11	819
Google Meet	FR > EN	7783	10	823
Google Meet	IT > EN	6917	9	801
Google Meet	JP > EN	6022	11	801
Google Meet	KO > EN	5045	10	611
Google Meet	PT > EN	6260	10	1339
Microsoft Teams	DE > EN	8443	10	2376
Microsoft Teams	EN > DE	6957	10	1622
Microsoft Teams	EN > ES	6793	10	1113
Microsoft Teams	EN > FR	7035	10	1895
Microsoft Teams	EN > IT	6796	10	1865
Microsoft Teams	EN > JP	6731	10	1089
Microsoft Teams	EN > KO	7053	10	2712
Microsoft Teams	EN > PT	7170	11	1638
Microsoft Teams	ES > EN	7141	10	1461
Microsoft Teams	FR > EN	7748	10	1704
Microsoft Teams	IT > EN	7117	10	1213
Microsoft Teams	JP > EN	5923	10	330
Microsoft Teams	KO > EN	5071	10	1356
Microsoft Teams	PT > EN	6160	10	1660
Zoom	DE > EN	7528	9	1865
Zoom	EN > DE	6780	10	1774
Zoom	EN > ES	7176	11	1611
Zoom	EN > FR	7023	10	1726
Zoom	EN > IT	7099	10	1732
Zoom	EN > JP	7066	10	2050
Zoom	EN > KO	6787	10	1716
Zoom	EN > PT	6805	10	1679
Zoom	ES > EN	6257	9	1552
Zoom	FR > EN	6903	9	1599
Zoom	IT > EN	6328	8	1725

Zoom	JP > EN	5207	9	1202
Zoom	KO > EN	4619	9	1217
Zoom	PT > EN	5504	9	1380

Drawing Final Conclusions

The final conclusions produced in this report are independent of DeepL or any other third-party. Slator did not artificially nor intentionally favor any one platform as part of this analysis and endeavored to provide completely independent and neutral analysis. Slator designed the methodology independently and retained full editorial control over the analysis and findings.



About Slator

Slator is the leading source of research and market intelligence for translation, localization, interpreting, and language AI. Slator's Advisory practice is a trusted partner to clients looking for M&A services and independent analysis. Slator has offices in Zurich (HQ) and London, and Analysts based in Asia, Europe, and the US.

Project Team

FLORIAN FAES
Managing Director
Slator
E: florian@slator.com

ALEX EDWARDS
Head of Consulting
Slator
E: alex@slator.com

ROCIO TXABARRIAGA
Senior Research Analyst
Slator
E: rocio@slator.com