Marktanalyse zur KI-gestützten Untertitelübersetzung

Unabhängige Technologiebewertung

slator

5. März 2026

Inhaltsverzeichnis

Zusammenfassung

1. DeepL Voice ist sowohl bei der Übersetzungsqualität als auch bei der Stabilität der Untertitel führend

DeepL Voice erzielte sowohl bei der menschlichen Bewertung als auch bei der automatisierten Analyse der Stabilität der Untertitel die höchsten Werte. DeepL Voice für Zoom Meetings erzielte einen Qualitäts-Score von **96,4/100**, während DeepL Voice für Teams **96,3/100** erreichte, verglichen mit 87–89 bei den anderen Plattformen. Die DeepL Voice-Produkte erzielten zudem die höchste Stabilität der Untertitel mit Stabilitätswerten von **88,6** für Zoom und **85,8** für Teams.

2. DeepL reduziert schwerwiegende Übersetzungsfehler erheblich

Über alle Sprachkombinationen hinweg senkte DeepL Voice die Rate kritischer oder schwerwiegender Übersetzungsfehler im Vergleich zu anderen bewerteten Plattformen um **durchschnittlich 76 %**.

Bei DeepL Voice waren 79 % der Segmente vollständig korrekt, verglichen mit **42 % bei den anderen Tools**.

Über alle Sprachkombinationen hinweg **reduzierte** DeepL Voice die durchschnittliche Anzahl an Übersetzungsfehlern pro Segment um **66 % gegenüber Microsoft Teams und um 64 % gegenüber Zoom**.

3. Bei der Stabilität der Untertitel zeigen sich deutliche Unterschiede zwischen den Plattformen

Auf allen Plattformen wurde Instabilität von Untertiteln beobachtet – dabei flackern die Übersetzungen auf dem Bildschirm oder werden wiederholt neu geschrieben. Die DeepL Voice-Produkte zeigten jedoch die geringste Instabilität der Untertitel (Neuschreibungen/Flackern).

Über alle Sprachkombinationen hinweg **reduzierte DeepL Voice diese Instabilität bei den Untertiteln um durchschnittlich 37,6 % gegenüber Microsoft Teams und um durchschnittlich 54,7 % gegenüber Zoom**.

4. DeepL Voice wurde von Sprachexperten eindeutig bevorzugt

Nach Abschluss der Blindbewertungen bevorzugten **96 % der Sprachexperten** ein **DeepL Voice-Produkt für die Untertitelübersetzung**.

---

5. Die DeepL Voice-Produkte sind führend bei KI-gestützter Untertitelübersetzung

Zur Zusammenführung der Ergebnisse zur Übersetzungsqualität und zur Stabilität der Untertitel hat Slator die Plattformen anhand von zwei Bewertungsdimensionen verglichen:

1. Übersetzungsqualität, gemessen anhand menschlicher Bewertung
2. Stabilität der Untertitel, gemessen anhand automatisierter Frame-Level-Analyse

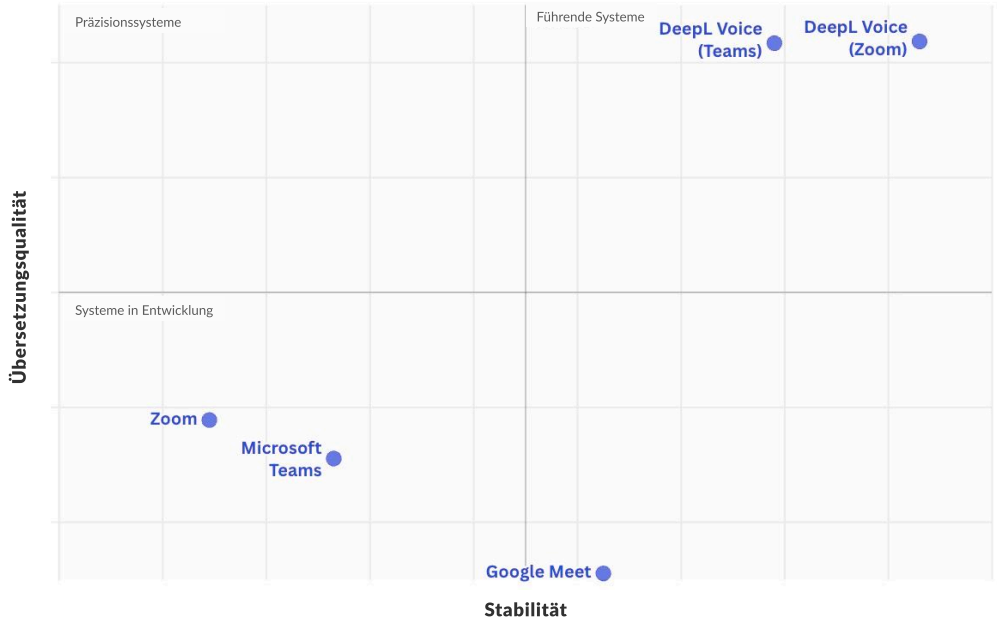
So ergeben sich vier mögliche Systemprofile:

Quadrant	Beschreibung
Führende Systeme	Hohe Übersetzungsqualität und hohe Stabilität der Untertitel
Präzisionssysteme	Hohe Übersetzungsqualität, aber geringere Stabilität der Untertitel
Effizienzorientierte Systeme	Geringere Übersetzungsqualität, aber relativ stabile Untertitel
Systeme in Entwicklung	Niedrige Leistung sowohl bei der Übersetzungsqualität als auch bei der Stabilität der Untertitel

Die Ergebnisse ordnen die DeepL Voice-Produkte im „**Führend**“-Quadranten dieses Bewertungsrahmens ein:

Quadrant für Plattformen zur KI-gestützten Übersetzung von Untertiteln

Vergleichende Bewertung von Systemen zur Untertitelübersetzung in Echtzeit hinsichtlich der Übersetzungsqualität und der Stabilität der Untertitel



Einführung

Übersetzungsfunktionen werden zunehmend in Unternehmenssoftware integriert, und KI-gestützte Untertitel sind inzwischen Bestandteil von Software für mehrsprachige Meetings. Die Übersetzung von Untertiteln in Echtzeit ist mittlerweile auf Plattformen wie Google Meet, Microsoft Teams und Zoom verfügbar, um die sprachübergreifende Kommunikation zu ermöglichen.

Trotz der zunehmenden Verbreitung dieser Systeme gibt es jedoch nur wenige unabhängige Vergleichstests, die ihre Leistungsfähigkeit in realen Besprechungssituationen bewerten.
In der Praxis hängt die Benutzerfreundlichkeit von Live-Untertiteln von zwei entscheidenden Faktoren ab:

- Übersetzungsqualität: Geben die Untertitel die Bedeutung des Gesprochenen korrekt wieder?
- Untertitelstabilität: Werden die Untertitel während der Sprachverarbeitung konsistent auf dem Bildschirm angezeigt, ohne dass sie häufig neu geschrieben werden oder flackern?

Häufige Änderungen von Untertiteln, teilweise neu geschriebene oder schwankende Untertitel können das Verständnis beeinträchtigen, selbst wenn die endgültige Übersetzung korrekt ist. Die gemeinsame Betrachtung von Übersetzungsqualität und Untertitelstabilität liefert daher ein umfassenderes Bild davon, wie Echtzeit-Untertitelungssysteme für Endnutzer abschneiden.

Um diese Lücke zu schließen, führte Slator eine unabhängige Bewertung von KI-gestützten Untertiteln über fünf Plattformen hinweg durch:

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

In der Studie wurden sowohl die Übersetzungsqualität als auch die Stabilität der Untertitel in 14 Sprachkombinationen untersucht, wobei sieben Sprachen ins Englische übersetzt und sieben Sprachen aus dem Englischen übersetzt wurden: Spanisch, Französisch, Deutsch, Italienisch, Portugiesisch, Koreanisch und Japanisch.

Diese Studie bewertet die Untertitel, die den Nutzern tatsächlich auf dem Bildschirm angezeigt werden, und nicht die zugrunde liegenden Transkripte der Spracherkennung oder internen Übersetzungsausgaben. Slator zeichnete Meetings auf, analysierte sie und extrahierte Untertitel direkt aus den auf dem Bildschirm dargestellten Videobildern, sodass bei der Analyse das tatsächlich für den Nutzer sichtbare Untertitelerlebnis gemessen werden konnte.

Achtundzwanzig professionelle Sprachexperten führten eine Blindbewertung der Untertitelübersetzung durch, um die Leistung der Plattformen zu vergleichen. Den Sprachexperten wurde nicht mitgeteilt, dass angepasste DeepL Voice-Systeme in die Bewertung einbezogen wurden, um sicherzustellen, dass die Bewertungen ausschließlich auf der Qualität und der Nutzerfreundlichkeit der Untertitel basierten.

Die Plattformen wurden unter Verwendung der standardmäßigen, voreingestellten Übersetzungsoptionen für Untertitel in Google Meet, Microsoft Teams und Zoom getestet. DeepL Voice für Teams und DeepL Voice für Zoom Meetings wurden unter Verwendung der für Endnutzer verfügbaren produktseitigen Funktionen bewertet, darunter Glossare sowie – im Fall von DeepL Voice für Teams – Funktionen zur Erkennung gesprochener Begriffe, die die Transkription von Eigennamen und Fachbegriffen verbessern.

Die Audioaufnahmen stammten aus Podcast-Aufzeichnungen, in denen zwei Sprecher in einem Gespräch über geschäftsbezogene Themen diskutierten. Jede Aufnahme wurde so bearbeitet, dass pro Sprache etwa 12 Minuten zusammenhängende Rede entstanden.

Durch diesen Ansatz wurde sichergestellt, dass bei der Bewertung natürliche Dialogmuster, fachspezifische Terminologie und realistische Sprachmerkmale erfasst wurden, wie sie für professionelle Besprechungen typisch sind. Die Bewertungsmethodik wurde entwickelt, um das tatsächlich für den Nutzer sichtbare Untertitelerlebnis plattformübergreifend unter vergleichbaren Bedingungen zu messen.

Da sich sprachliche Strukturen von Sprache zu Sprache erheblich unterscheiden, können einige Übersetzungsrichtungen naturgemäß mehr zwischenzeitliche Änderungen der Untertitel erzeugen als andere. Beispielsweise erfordern Sprachen mit abweichender Wortstellung oder Verbendstellung, wie Japanisch oder Koreanisch, unter Umständen, dass Untertitel mit zunehmendem Kontext nachträglich angepasst oder neu geschrieben werden müssen. Aus diesem Grund werden die Ergebnisse sowohl aggregiert als auch für jede einzelne Sprache analysiert.

Zusätzlich zur menschlichen Bewertung entwickelte Slator ein automatisiertes Messverfahren zur Quantifizierung der Stabilität der Untertitel, bei dem Frame-by-Frame-Änderungen der dargestellten Untertitel analysiert wurden.

Das Verfahren sah wie folgt aus: Videoaufzeichnung > Frame-Extraktion (10 Bilder pro Sekunde) > Zuschneiden des Untertitelbereichs > optische Zeichenerkennung (OCR, Tesseract) > Textnormalisierung > Frame-Vergleich > Erkennung von Änderungsereignissen.

Zusammen liefern diese manuellen und automatisierten Analysen eine umfassende Bewertung der Leistung von KI-gestützter Untertitelübersetzung in mehrsprachigen Echtzeit-Besprechungen.

Eine vollständige Beschreibung der Methodik und des Verfahrens finden Sie im Anhang.

Übersetzungsqualität

Slator beauftragte 28 professionelle Sprachexperten mit der Blindbewertung von KI-gestützten Untertiteln über 14 Sprachkombinationen hinweg (7 ins Englische, 7 aus dem Englischen). Weitere Einzelheiten zur Methodik finden Sie im Anhang.

Über alle getesteten Sprachen und Plattformen hinweg **übertraf DeepL Voice die integrierten Untertitelfunktionen durchgängig** sowohl bei der Übersetzungsqualität als auch bei der Reduktion kritischer Übersetzungsfehler.

Die Ergebnisse sind wie folgt:

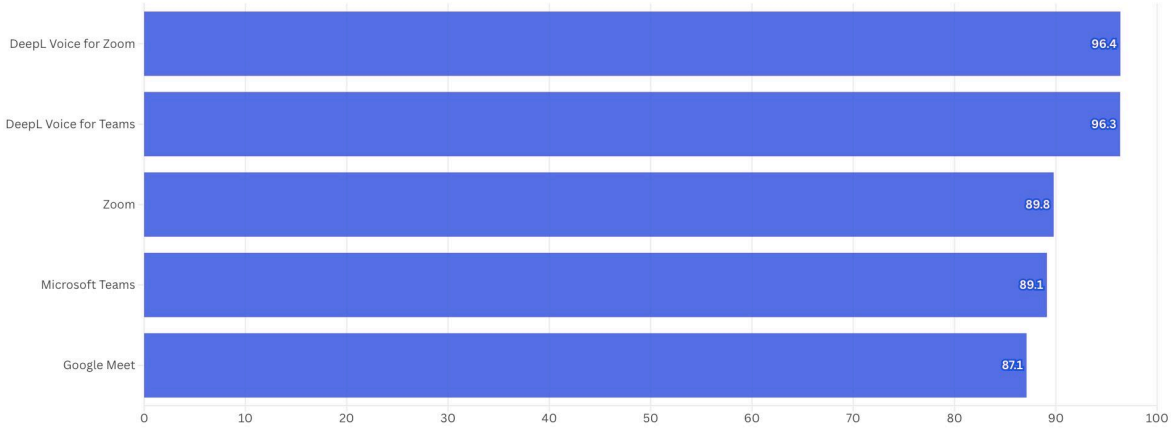
Gesamtleistung

Slator fasste die Ergebnisse in einem einheitlichen Qualitäts-Score von 0 bis 100 zusammen, der die Schwere der Übersetzungsfehler in allen bewerteten Segmenten widerspiegelt. Dadurch konnte Slator diese Frage beantworten: „Wie gut ist die Qualität der KI-gestützten Untertitelübersetzung insgesamt auf allen fünf Plattformen?“

Die Ergebnisse der Bewertung lauten wie folgt:

DeepL Voice für Zoom Meetings erzielte bei der menschlichen Bewertung von KI-gestützter Untertitelübersetzung die höchsten Werte

Qualitäts-Score (/100) für KI-gestützte Untertitel über alle getesteten Sprachkombinationen und Plattformen hinweg



Quelle: Slator • Der Gesamtqualitätswert wurde auf Grundlage von Blindbewertungen KI-gestützter Untertitel durch jeweils zwei muttersprachliche Sprachexperten pro Sprachkombination ermittelt, wobei Übersetzungsqualität und Sprachfluss bewertet wurden. Der Qualitäts-Score spiegelt den durchschnittlichen Schweregrad der Fehler über alle bewerteten Segmente hinweg wider und wird auf eine Skala von 0–100 übertragen. Die getesteten Sprachkombinationen umfassten Übersetzungen aus dem Englischen in Spanisch, Französisch, Deutsch, Italienisch, Portugiesisch, Koreanisch und Japanisch sowie in umgekehrter Richtung.

- DeepL Voice für Zoom Meetings erzielte bei der menschlichen Bewertung von KI-gestützten Untertiteln mit 96,4 von 100 Punkten den höchsten Qualitäts-Score.
- Dicht dahinter folgte DeepL Voice für Teams mit einer Bewertung von 96,3 von 100 Punkten.
- Google Meet erzielte mit 87 von 100 Punkten den niedrigsten Wert.

Bewertung der Fehlerschwere

Slator wertete die oben genannten Qualitäts-Scores weiter aus, um folgende Frage zu beantworten: „Wie häufig wirkt sich die Übersetzungsqualität spürbar auf das Verständnis von auf dem Bildschirm angezeigten Untertiteln durch Endnutzer aus?“ Dies ermöglicht ein besseres Verständnis dafür, wie schwerwiegend Übersetzungsfehler auf den einzelnen Plattformen sind.

Slator erfasste den Anteil der Segmente mit kritischen oder schwerwiegenden Übersetzungsfehlern im Bereich der Genauigkeit. Die Sprachexperten wurden gebeten, die Schwere von Übersetzungsfehlern anhand der Genauigkeit (Fehlübersetzungen, Auslassungen und Ergänzungen) und des Sprachflusses (Stil, Grammatik, Rechtschreibung) zu bewerten. Auf diese Weise konnte Slator die Fehlerquote der KI-gestützten Untertitelübersetzung ermitteln, d. h. den Prozentsatz der Segmente mit kritischen oder schwerwiegenden Genauigkeitsfehlern, die zu Bedeutungsverlust oder Fehlübersetzungen führten.

Dadurch ergibt sich eine differenziertere Einordnung des oben dargestellten Qualitäts-Scores, da dieser Anteil zeigt, wie häufig Untertitel die Bedeutung des Gesprochenen erheblich verfälschen oder verlieren. Ebenso zeigt der Anteil bestandener Segmente, wie oft die jeweilige Plattform die Bedeutung des Gesprochenen korrekt wiedergibt und einen angemessenen Sprachfluss in der Zielsprache aufweist.

Hier sind Beispiele für jede Kategorie (Übersetzungen aus Google Meet):

Beispiel 1 (nicht bestanden):

- Originaltext (Englisch): „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.“
- Untertitel (Spanisch): „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.“
- Fehlerkategorisierung: Fehlübersetzung (schwerwiegend), Stil (schwerwiegend)
- Kommentar des Sprachexperten: Eine wörtliche Übersetzung führt zu semantischen Fehlern, insbesondere bei Formulierungen wie „when you sit in this position“ (wenn Sie in dieser Position sitzen), sowie zu umständlichen Formulierungen wie „stakeholders who have interests“ (Interessengruppen, die Interessen haben). „Merck“ wurde im Spanischen als „Murder“ (Mord) interpretiert.

Beispiel 2 (bestanden mit Mängeln):

- Originaltext (Englisch): „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.“
- Untertitel (Spanisch): „Y si tienes que pagar tanto dinero, todas las cosas que acabo de decir no te harán sentir satisfecho con el sistema.“
- Fehlerkategorisierung: Stil (schwerwiegend)
- Kommentar des Sprachexperten: Die KI-gestützte Untertitelübersetzung wechselte zur informellen Anrede („tú“). Diese Übersetzung ist ungewöhnlich, da im natürlichen Spanisch die neutrale dritte Person verwendet wird („Y si hay que pagar tanto dinero…“).

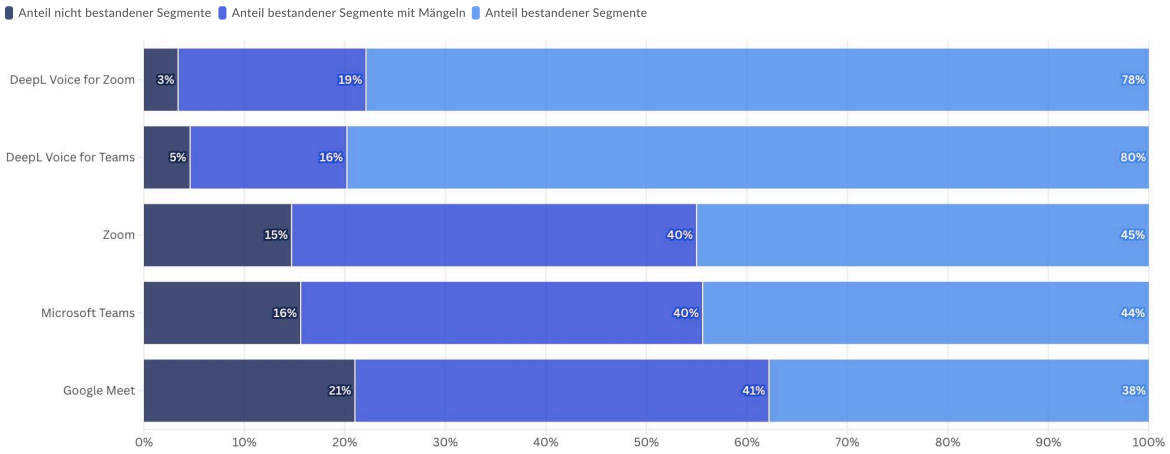
Beispiel 3 (bestanden):

- Originaltext (Englisch): „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."
- Untertitel (Spanisch): „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."
- Fehlerkategorisierung: Kommentar des Sprachexperten: „Die spanische Übersetzung in diesem Segment entspricht dem gesprochenen Englisch."

Die Ergebnisse dieser Analyse lauten wie folgt:

DeepL Voice reduziert kritische Fehler im Durchschnitt um 13 % und erreicht einen durchschnittlichen Anteil bestandener Segmente von 79 %

Anteil nicht bestandener Segmente / Anteil bestandener Segmente mit Einschränkungen / Anteil bestandener Segmente für KI-gestützte Untertitel über alle getesteten Sprachkombinationen und Plattformen hinweg



Quelle: Slator • Die Werte wurden auf Grundlage von Blindbewertungen KI-gestützter Untertitel ermittelt, bei denen jeweils zwei muttersprachliche professionelle Sprachexperten pro Sprachkombination die Genauigkeit und den Sprachfluss beurteilten. Die Werte spiegeln den Schweregrad der Übersetzungsfehler über alle bewerteten Segmente hinweg wider. Die getesteten Sprachkombinationen umfassten Übersetzungen aus dem Englischen in Spanisch, Französisch, Deutsch, Italienisch, Portugiesisch, Koreanisch und Japanisch sowie in umgekehrter Richtung.

- KI-gestützte Übersetzung von Untertiteln bei **DeepL Voice-Produkten** wiesen im Durchschnitt einen Anteil **nicht bestandener Segmente von 4 %** auf, verglichen mit durchschnittlich **17 %** bei allen anderen Tools. **Dies entspricht einer Reduktion kritischer oder schwerwiegender Genauigkeitsfehler um 76 % bei Verwendung von DeepL Voice**.
- Der Anteil **bestandener Segmente** lag bei KI-gestützten Untertiteln mit DeepL Voice im Durchschnitt bei **79 %**, verglichen mit 42 % bei allen anderen Tools. Dies entspricht einem relativen Anstieg **vollständig bestandener Segmente um 88 %** gegenüber anderen Tools.
- Rund 60 % der Segmente bei Google Meet sowie etwa die Hälfte der Segmente bei Zoom und Microsoft Teams enthielten schwerwiegende Fehler oder erhebliche Übersetzungsprobleme.

Leistung nach Sprachen

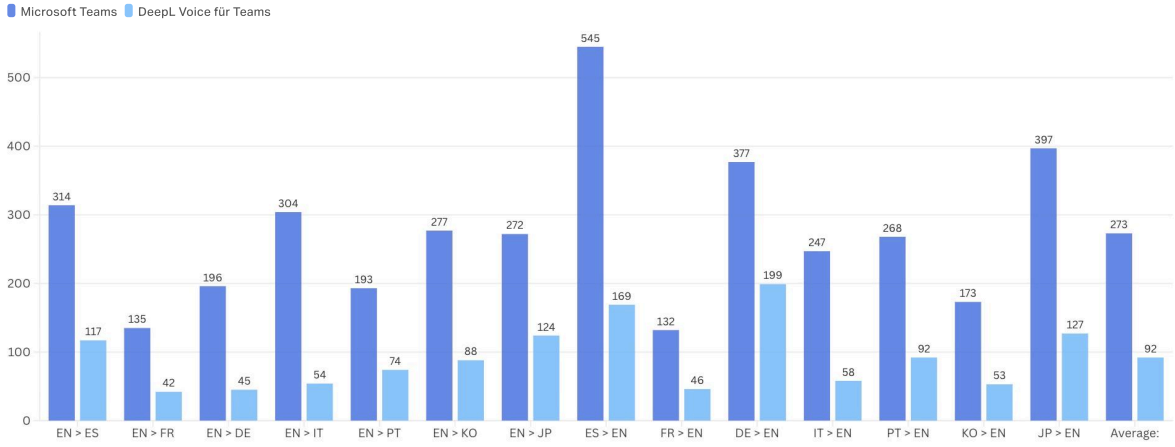
Slator hat sich eingehender mit den sprachlichen Unterschieden zwischen Standardplattformen (Microsoft Teams, Zoom) und angepassten Plattformen (DeepL Voice für Teams und DeepL Voice für Zoom Meetings) befasst, um folgende Frage zu beantworten: „Wie deutlich unterscheidet sich die Leistung von DeepL Voice im Vergleich zu Standardplattformen?"

Über alle Sprachkombinationen hinweg reduzierte DeepL Voice die durchschnittliche Anzahl von Übersetzungsfehlern pro Segment um 66 % gegenüber Microsoft Teams und um 64 % gegenüber Zoom.

Beim Vergleich von Microsoft Teams und DeepL Voice für Microsoft Teams:

Der Einsatz von DeepL Voice für Teams reduziert die durchschnittliche Anzahl von Übersetzungsfehlern pro Segment um 66 %

Vergleich der durchschnittlichen Anzahl von Übersetzungsfehlern pro Segment in KI-gestützten Untertiteln über verschiedene Sprachkombinationen hinweg zwischen Microsoft Teams und DeepL Voice für Teams



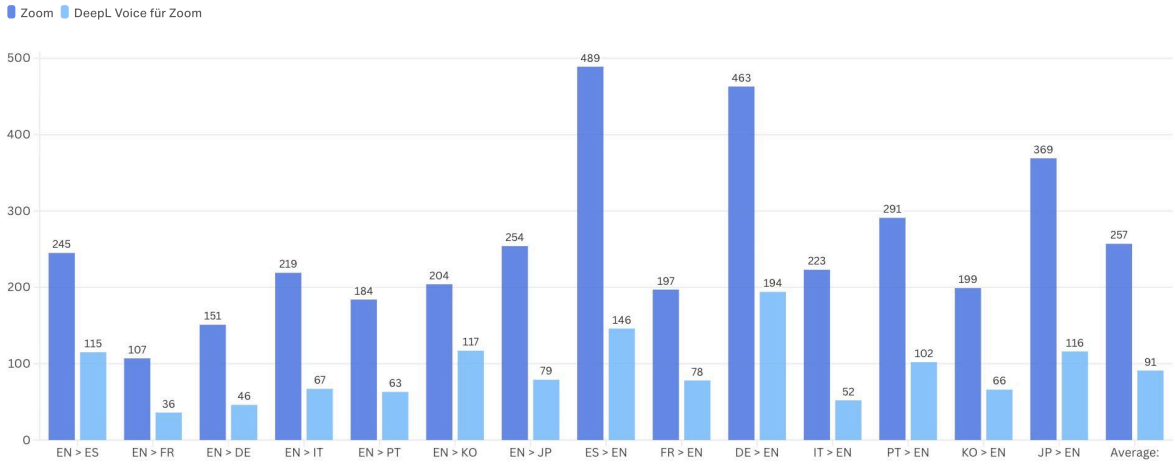
Quelle: Slator • Niedrigere Zahlen stehen für höhere Qualität. Die Untertitel wurden von jeweils zwei professionellen Sprachexperten pro Sprachkombination geprüft und hinsichtlich Genauigkeit und Sprachfluss beurteilt. Im Rahmen der Bewertung wurden Fehler in der Sprachflüssigkeit mit dem Faktor 0,5 gewichtet.

- Die stärksten Rückgänge bei den durchschnittlichen Übersetzungsfehlern pro Segment wurden bei DeepL Voice für Teams in folgenden Sprachpaaren verzeichnet:
  - Englisch ins Italienische (82 %)
  - Englisch ins Deutsche (77 %)
  - Italienisch ins Englische (76 %)
- Nur eine Sprachkombination lag bei der Verwendung von DeepL Voice für Teams unter einer Verbesserung von 50 % (Deutsch ins Englische, 47 %).

Beim Vergleich von Zoom und DeepL Voice für Zoom Meetings:

Der Einsatz von DeepL Voice für Zoom Meetings reduziert die durchschnittliche Anzahl von Übersetzungsfehlern pro Segment um 65 %

Vergleich der durchschnittlichen Anzahl von Übersetzungsfehlern pro Segment in KI-gestützten Untertiteln über Sprachkombinationen hinweg zwischen Zoom und DeepL Voice für Zoom Meetings



Quelle: Slator • Niedrigere Zahlen stehen für höhere Qualität. Die Untertitel wurden von jeweils zwei professionellen Sprachexperten pro Sprachkombination geprüft und hinsichtlich Genauigkeit und Sprachfluss beurteilt. Im Rahmen der Bewertung wurden Fehler in der Sprachflüssigkeit mit dem Faktor 0,5 gewichtet.

- Die **größten Reduktionen** der durchschnittlichen Übersetzungsfehler pro Segment mit **DeepL Voice für Zoom Meetings** wurden in folgenden Sprachkombinationen erzielt:
  - Italienisch ins Englische (77 %)
  - Spanisch ins Englische (70 %)
  - Englisch ins Deutsche (70 %)
- Nur bei einer Sprachkombination lag die Verbesserung bei der Verwendung von DeepL Voice für Zoom Meetings unter 50 % (Englisch ins Koreanische, 43 %).

Bevorzugte Plattformen der Sprachexperten

Nach Abschluss der Qualitätsbewertungen bat Slator alle 28 teilnehmenden Sprachexperten, ihre bevorzugte Plattform für KI-gestützte Untertitel auf Grundlage des Nutzererlebnisses zu bewerten. Diese Sprachexperten führten ihre Analyse im Rahmen einer Blindbewertung durch und wählten ihre bevorzugte Plattform auf Basis von für jede Plattform vergebenen Codenamen aus.

Die Ergebnisse lauten wie folgt:

- **52 %** der Sprachexperten (15) nannten **DeepL Voice für Teams** als ihre bevorzugte Plattform.
- **44 %** der Sprachexperten (12) bevorzugten **DeepL Voice für Zoom Meetings**.
- Insgesamt nannten **96 %** der Sprachexperten (27) entweder DeepL Voice für Teams oder DeepL Voice für Zoom Meetings als ihre **beiden bevorzugten Optionen**.
- Zoom belegte in der Regel den dritten Platz, erhielt jedoch eine Stimme als bevorzugtes Tool. Microsoft Teams und Google Meet erhielten keine Stimmen für die ersten beiden Plätze und belegten durchweg den 4. oder 5. Platz.

Stabilität der Untertitel

Slator hat die Anzahl der Änderungsereignisse über alle Sprachen und Plattformen hinweg automatisch erfasst. Im Folgenden sind Sie drei Beispiele aus den analysierten Audioaufnahmen aufgeführt, die zeigen, wie Untertitel vor ihrer vollständigen Stabilisierung dargestellt werden:

Beispiel 1 (schrittweiser Satzaufbau, Französisch ins Englische):

- 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 (stabil): „Hello Nathalie the topic of the day is the subject of pharmacy groups.“

In diesem Beispiel zeigen die Live-Untertitel Teilübersetzungen an, bevor der Sprecher einen Satz beendet hat. Während weitere Wörter verarbeitet werden, werden die Untertitel schrittweise ergänzt, bis der vollständige Satz stabil ist.

Beispiel 2 (Neuschreibung von Untertiteln, Spanisch ins Englische):

- 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 you can file a case?“
- Frame 4 (stabil): „When you make a claim what is the minimum economic amount so that you can file a case?“

In diesem Beispiel werden die Live-Untertitel aktualisiert, sobald weiterer sprachlicher Kontext verfügbar wird.

Beispiel 3 (Flackern der Untertitel, Koreanisch ins Englische):

- Bild 1: „the fourth quarter ended about 10 days“
- Frame 2: „the fourth quarter is now about 10 days“
- Frame 3 (stabil): „the fourth quarter ended about 10 days ago“

In diesem Beispiel werden die Live-Untertitel bereits angezeigte Wörter vollständig neu geschrieben, bevor sie erneut in eine andere endgültige Fassung übergehen (A>B>A, A>B>C oder ähnliche Muster).

Beispiel 4 (hohe Stabilität, Koreanisch ins Englische):

- 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 diesem Beispiel wurde die endgültige Untertitelübersetzung angezeigt und blieb unverändert.

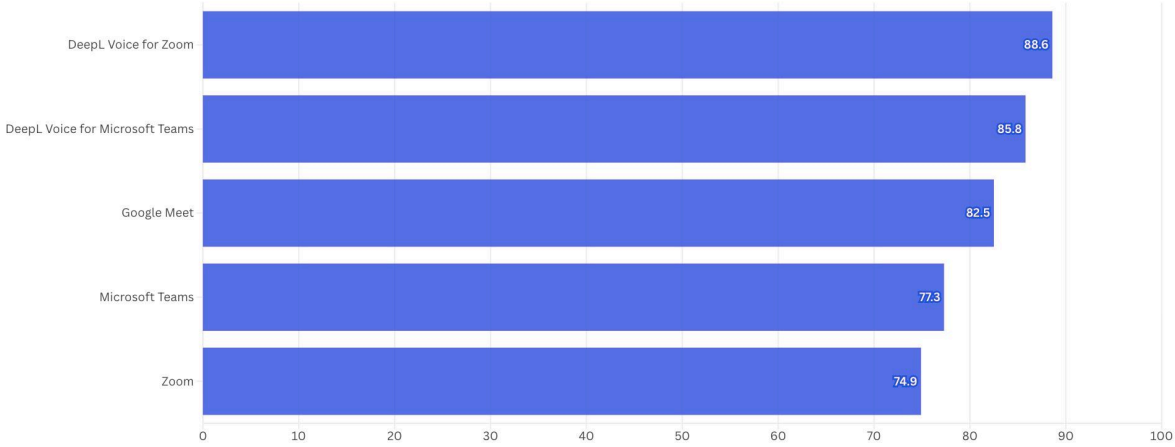
Gesamtleistung

Slator fasste die Ergebnisse in einem einheitlichen Qualitäts-Score von 0–100 zusammen, der die Stabilität der Untertitel plattformübergreifend widerspiegelt. Dadurch konnte Slator diese Frage beantworten: „Wie stark flackern oder ändern sich die Übersetzungen auf den einzelnen Plattformen, bis die Untertitel vollständig stabil sind?“

Die Ergebnisse der Bewertung lauten wie folgt:

DeepL Voice für Zoom Meetings ist die stabilste Plattform für Untertitelübersetzungen

Der durchschnittliche Stabilitätswert (/100) für Untertitel über alle getesteten Sprachkombinationen und Plattformen hinweg



Quelle: Slator • Der durchschnittliche Stabilitätswert wurde ermittelt, indem Frame-by-Frame-Aufzeichnungen der übersetzten Untertitel in jedem Tool analysiert und die Gesamtzahl der Änderungsereignisse über alle Sprachkombinationen hinweg ermittelt wurde. Diese Werte wurden zu einem Stabilitätswert auf einer Skala von 0–100 zusammengefasst.

- **DeepL Voice für Zoom Meetings** erzielte bei einer automatisierten Bewertung der Stabilität der KI-gestützten Untertitelübersetzungen mit **88,6 von 100 Punkten** die **höchste Punktzahl**.
- Dicht dahinter folgte **DeepL Voice für Teams** mit einer Punktzahl von **85,8 von 100**.
- Zoom erzielte mit 74,9 von 100 Punkten das niedrigste Ergebnis.

Leistung nach Sprachen

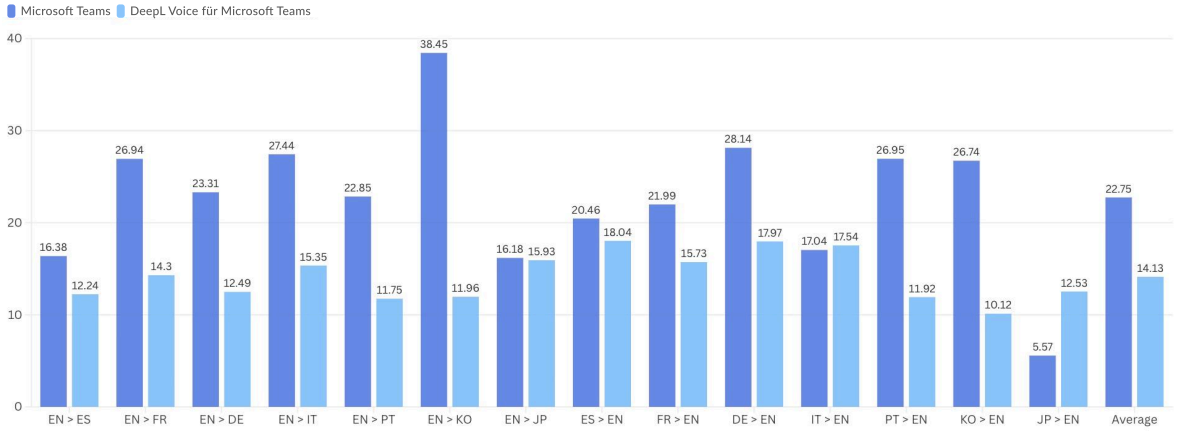
Slator hat sich eingehend mit den sprachlichen Unterschieden zwischen Standardplattformen (Microsoft Teams, Zoom) und angepassten Plattformen (DeepL Voice für Teams und DeepL Voice für Zoom Meetings) befasst, um folgende Frage zu beantworten: „Sind die DeepL Voice-Produkte im Vergleich zu Standardlösungen auf sprachlicher Ebene stabiler oder weniger stabil?“

Über alle Sprachkombinationen hinweg **reduzierte DeepL Voice die Instabilität der Untertitel** im Durchschnitt um **37,6 %** gegenüber **Microsoft Teams** und um **54,7 %** gegenüber **Zoom**.

Beim Vergleich von Microsoft Teams und DeepL Voice für Microsoft Teams:

Der Einsatz von DeepL Voice für Teams verbessert die Stabilität um durchschnittlich 38 %

Ein Vergleich der Instabilität bei KI-übersetzten Untertiteln in verschiedenen Sprachkombinationen zwischen Microsoft Teams und DeepL Voice für Teams



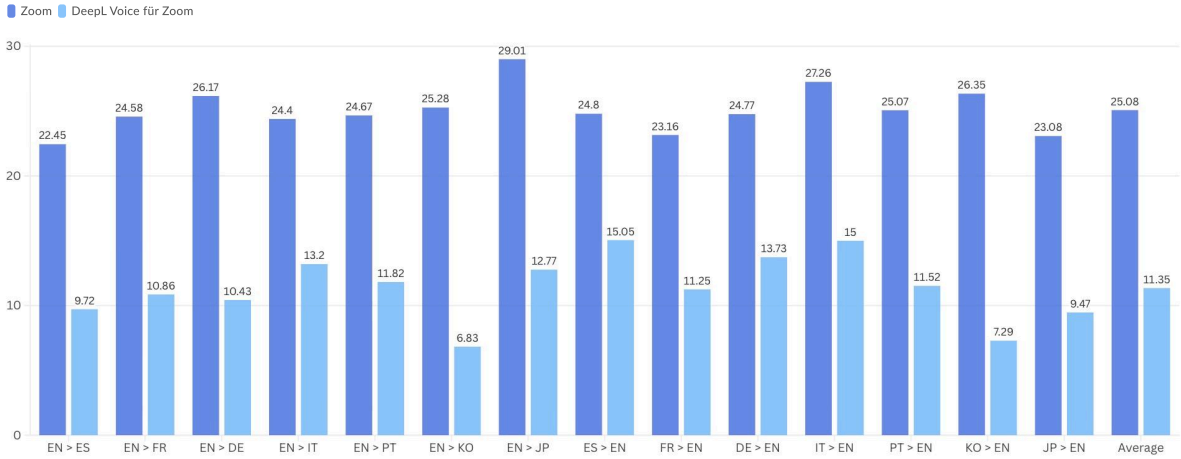
Quelle: Slator • Die Instabilität ist definiert als der prozentuale Anteil der Frames, in denen sich ein angezeigter Untertitel ändert. Dies macht direkt sichtbar, wie oft Nutzer Aktualisierungen oder Unterbrechungen bei den Untertiteln erleben. Die Zahlen wurden automatisch mit Python berechnet.

- Die **größten Verbesserungen** hinsichtlich der Stabilität der Untertitel bei **DeepL Voice für Teams** waren in folgenden Kombinationen zu verzeichnen:
  - Englisch ins Koreanische (+69 %)
  - Koreanisch ins Englische (+62 %)
  - Portugiesisch ins Englische (+56 %)
- Nur bei zwei Sprachkombinationen waren die KI-gestützten Untertitel in Microsoft Teams stabiler – Italienisch ins Englische (3 % stabiler in Microsoft Teams) und Japanisch ins Englische (125 % stabiler in Microsoft Teams).

Beim Vergleich von Zoom und DeepL Voice für Zoom Meetings:

Der Einsatz von DeepL Voice für Zoom Meetings verbessert die Stabilität um durchschnittlich 55 %

Ein Vergleich der Instabilität (Neuschreibungen/Flackern) bei KI-gestützten Untertiteln über verschiedene Sprachkombinationen hinweg zwischen Zoom und DeepL Voice für Zoom Meetings



Quelle: Slator • Die Instabilität ist definiert als der prozentuale Anteil der Frames, in denen sich ein angezeigter Untertitel ändert. Dies macht direkt sichtbar, wie oft Nutzer Aktualisierungen oder Unterbrechungen bei den Untertiteln erleben. Die Zahlen wurden automatisch mit Python berechnet.

- Die **größten Verbesserungen** hinsichtlich der Stabilität der Untertitel bei **DeepL Voice für Zoom Meetings** waren bei folgenden Kombinationen zu verzeichnen:
  - Englisch ins Koreanische (73 %)
  - Koreanisch ins Englische (72 %)
  - Englisch ins Deutsche (60 %)

- Für keine Sprachkombination zeigte DeepL Voice für Zoom Meetings im Vergleich zu Zoom eine geringere Stabilität.

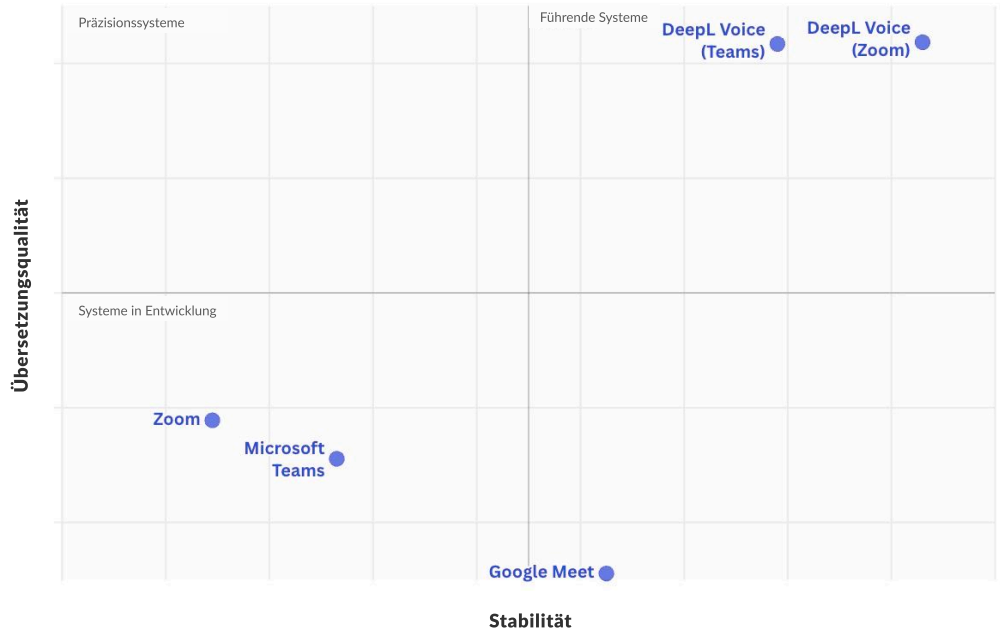
Schlussfolgerung

Gesamtergebnisse

Hier ist die Gesamtwertung aller fünf Plattformen für KI-Übersetzungen von Untertiteln:

Quadrant für Plattformen zur KI-gestützten Übersetzung von Untertiteln

Vergleichende Bewertung von Systemen zur Untertitelübersetzung in Echtzeit hinsichtlich der Übersetzungsqualität und der Stabilität der Untertitel



DeepL Voice für Teams und **DeepL Voice für Zoom Meetings** erzielten sowohl bei der Übersetzungsqualität als auch bei der Stabilität hohe Werte und positionieren sich damit als **führende Lösungen für KI-übersetzte Untertitel**. Echtzeit-Untertitelungssysteme stehen möglicherweise vor einem Zielkonflikt zwischen Übersetzungsgenauigkeit und Stabilität der Untertitel, da sie einerseits hochpräzise Übersetzungen liefern und andererseits die Verzögerung bei der Darstellung dieser Übersetzungen gering halten müssen. Die Ergebnisse dieser Studie deuten jedoch darauf hin, dass DeepL Voice beide Aspekte gleichzeitig verbessert.

Sowohl **Microsoft Teams** als auch **Zoom** werden als **Systeme in der Entwicklung** eingestuft, da sie sowohl bei der Darstellung von Untertiteln als auch bei der Übersetzungsqualität unterdurchschnittliche Ergebnisse erzielen. Microsoft Teams schnitt hinsichtlich der Stabilität geringfügig besser ab, während Zoom bei der Übersetzungsqualität etwas besser abschnitt. Beide Tools müssten jedoch in beiden Kategorien verbessert werden, um ihr derzeitiges Angebot zu optimieren.

Google Meet war zwar im Vergleich zu anderen Tools auf dem Markt relativ stabil, erzielte jedoch den niedrigsten

Qualitäts-Score, wie von Slators Team aus Sprachexperten ermittelt. Damit positioniert sich Google Meet im Gesamtranking als „Effizienzorientiertes System“ und könnte sich zum „Führenden System“ entwickeln, wenn die Plattform die Übersetzungsqualität in allen Sprachkombinationen verbessert und gleichzeitig die Darstellung der Untertitel beibehält oder verbessert.

Auswirkungen für Unternehmen

Slator zieht auf Grundlage dieser Bewertung folgende Schlussfolgerungen für Unternehmen:

1. **Die Qualität der KI-Untertitel verbessert sich zwar, ist jedoch weiterhin uneinheitlich zwischen den Plattformen.**

Die Ergebnisse zeigen deutliche Unterschiede in der Übersetzungsqualität zwischen gängigen Meeting-Plattformen. Unternehmen, die auf integrierte Untertitelübersetzung setzen, können je nach verwendetem Tool erhebliche Unterschiede in der Übersetzungsqualität feststellen.

2. **Die Stabilität der Untertitel wirkt sich unmittelbar auf die Benutzerfreundlichkeit aus.**

Selbst wenn die Übersetzungen korrekt sind, können häufige Neuschreibungen (Flackern) das Verständnis beeinträchtigen. Die Messung der Untertitelstabilität dient als sinnvoller Proxy für die tatsächliche Nutzererfahrung in mehrsprachigen Besprechungen.

3. **Spezialisierte Plattformen mit Anpassungsmöglichkeiten können die Übersetzungsleistung verbessern.**

Die Ergebnisse deuten darauf hin, dass der Einsatz spezialisierter KI-Übersetzungstools sowie die Verwendung fachspezifischer Glossare und die Ergänzung durch Transkriptionen die Übersetzungsleistung bei KI-Untertiteln in Echtzeit erheblich verbessern können.

Anhang

Methodik

Beschaffung von Audio-Samples

Slator hat Podcast-Aufnahmen in allen relevanten Sprachen zusammengestellt. Es wurden Podcast-Aufnahmen ausgewählt, da diese in allen Sprachen eine gleichbleibende Audioqualität bieten und somit einen kontrollierten Vergleich der Übersetzungsleistung ermöglichen, anstatt Unterschiede durch Hintergrundgeräusche oder Mikrofonqualität zu verursachen.

In jedem Podcast führten zwei Sprecher ein wechselseitiges Gespräch über geschäftliche, regulierte Themen (Finanzen, Wissenschaft, Recht), wobei die Hintergrundgeräusche minimal bis gar nicht vorhanden waren. Dieser Ansatz diente dazu, die Übersetzungsqualität von Fachbegriffen zu messen und zu bewerten, wie die Plattformen natürliche Gespräche zwischen Muttersprachlern der Ausgangssprache verarbeiten.

Jeder Podcast wurde so bearbeitet, dass ein isolierter Ausschnitt des Ausgangsmaterials mit einer Länge von etwa 12 Minuten pro Sprache entstand. Diese Audiodatei in der Ausgangssprache wurde mit einem automatisierten Transkriptionstool verarbeitet, um ein Transkript zu erstellen.

Technische Einrichtung

Slator richtete ein virtuelles Audiokabel ein, um die Audiodatei an jede Plattform (Google Meet, Microsoft Teams, Zoom) weiterzuleiten. Dadurch wurde sichergestellt, dass während der Aufnahme keine zusätzlichen oder unbeabsichtigten Hintergrundgeräusche hinzugefügt wurden. Slator erstellte mithilfe einer Screenshot-Funktion eine MP4-Videodatei der übersetzten Untertitel auf jeder Plattform, die die Live-Untertitel genau so zeigt, wie sie für den Nutzer erscheinen.

Für Google Meet, Microsoft Teams und Zoom wurden die standardmäßigen, vorkonfigurierten Einstellungen für die Untertitelübersetzung verwendet. Speziell bei Google Meet wurde die Übersetzung vom Englischen ins Italienische als Beta-Funktion gekennzeichnet.

Beim Testen von DeepL Voice in Microsoft Teams und Zoom verwendete Slator ein Glossar mit Fachbegriffen und wählte in den Formalitätseinstellungen von DeepL den informellen Sprachstil aus, da die Podcast-Gespräche in einem informellen Ton geführt wurden.

Darüber hinaus aktivierte Slator die Funktion „Gesprochene Begriffe“ von DeepL Voice, die ausschließlich in Microsoft Teams verfügbar ist. Diese Funktion ist eine einsprachige Liste von gesprochenen Begriffen (d. h. Abkürzungen oder Eigennamen), die die Fähigkeit des Tools verbessert, einsprachige Begriffe im Rahmen der Speech-to-Text-Transkription zu erkennen.

Slator hat alle Aufzeichnungen über eine automatisierte Verarbeitungs-Engine (wie im Abschnitt „Automatisierte Qualitätsbewertungen“ beschrieben) verarbeitet und diese Aufzeichnungen an Sprachexperten weitergeleitet.

Auswahl der Sprachexperten

Slator beauftragte für jede Sprachkombination zwei Sprachexperten und stellte sicher, dass diese Muttersprachler oder nahezu muttersprachlich in der Zielsprache waren. Die ausgewählten Sprachexperten wurden auf Grundlage ihrer fachlichen Expertise in Bezug auf die Audioinhalte gezielt ausgewählt und verfügen über entsprechende Qualifikationen.

Sie hatten vor allem Erfahrung in den Bereichen Medienlokalisierung, Übersetzung, Dolmetschen und Qualitätsbewertung (einschließlich KI-Training und KI-Bewertung). Diese Kombination aus Fachkenntnissen erforderte es, die Audiodatei anzuhören und gleichzeitig die Qualität der Untertitel auf dem Bildschirm zu beurteilen.

Von Menschen durchgeführte Qualitätsbewertungen

Slator erstellte für jede Audiodatei in der Ausgangssprachei eine Scorecard zur Qualitätsbewertung. Nachfolgend ein Beispiel:

	A	B	C	D	E	F	G	H	I	J	K
1	Baseline Parameters				Final Caption Quality						
2	Tool Number	Segment ID	Start Time	Original Transcript	Mistranslation	Omission	Addition	Style	Grammar	Spelling	Evaluator Comments
3	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.							
4	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.							
5	1	3	[00:00:21.720]	That actually gets to the short term versus the long term issue.							

Die Scorecard enthält:

- Basisparameter:
  - Tool-Nummer: Dabei handelt es sich um eine anonymisierte Referenz für jedes Tool bzw. jede Plattform, um zu verhindern, dass die Linguisten den Namen der von ihnen bewerteten Plattform erfahren oder erkennen können, welche Plattform angepasst wurde und welche nicht. Den Sprachexperten wurde nicht mitgeteilt, dass diese Bewertung auch angepasste Modelle von DeepL umfasste.
  - Segment ID: Dies ist die Segment-ID des Transkripts der Audiodatei in der Ausgangssprache. Dadurch konnte Slator die Bewertung in einzelne Textabschnitte unterteilen, was eine klarere Analyse durch mehrere Sprachexperten innerhalb derselben Sprachkombination sowie über verschiedene Sprachen und Plattformen hinweg ermöglichte.
  - Startzeit: Dies basiert auf dem Originaltranskript und ermöglichte es den Linguisten, bei Bedarf gezielt zu bestimmten Abschnitten in der Original-Audiodatei zu navigieren.
  - Originaltranskript: Dies ist das Originaltranskript der Ausgangssprache. Die Sprachexperten wurden darauf hingewiesen, dass dieser Text Fehler enthalten könnte. Slator hat bewusst darauf verzichtet, Fehler im

Transkript zu bereinigen, um zu verhindern, dass die Sprachexperten die übersetzten Untertitel mit dem schriftlichen Transkript abgleichen, und stattdessen sicherzustellen, dass sie sich auf das Anhören der Audiodatei konzentrieren. Aus denselben Gründen stellte Slator auch kein schriftliches Transkript der übersetzten Untertitel zur Verfügung.

- Qualität der letztlich angezeigten Untertitel
  - Kategorien der Übersetzungsgenauigkeit:
    - Fehlübersetzung: Slator definierte diese Kategorie folgendermaßen: „Die Bedeutung des übersetzten Untertitels ist falsch oder irreführend." Die möglichen Einstufungen wurden wie folgt definiert:
      - Keine – Kein erkennbares Problem
      - Geringfügig – Die Bedeutung ist weitgehend korrekt, lediglich Nuancen gehen verloren
      - Schwerwiegend – Die Bedeutung ist teilweise verfälscht
      - Kritisch – Die Bedeutung ist inhaltlich falsch
    - Auslassung: Slator definierte diese Kategorie folgendermaßen: „Im übersetzten Untertitel fehlen wichtige Informationen." Die möglichen Einstufungen wurden wie folgt definiert:
      - Keine – Kein erkennbares Problem
      - Geringfügig – Die Bedeutung ist weitgehend korrekt, lediglich Nuancen gehen verloren
      - Schwerwiegend – Die Bedeutung ist teilweise verfälscht
      - Kritisch – Die Bedeutung ist inhaltlich falsch. Die Linguisten wurden angewiesen, Segmente als „kritisch" einzustufen, wenn die Plattform das Segment nicht übersetzte oder die Ausgangssprache im übersetzten Untertitel anzeigte
    - Hinzufügung: Slator definierte diese Kategorie folgendermaßen: „Dem übersetzten Untertitel wurden Informationen hinzugefügt, die im Audiomaterial nicht enthalten sind." Die möglichen Einstufungen wurden wie folgt definiert:
      - Keine – Kein erkennbares Problem
      - Geringfügig – Die Bedeutung ist weitgehend korrekt, lediglich Nuancen gehen verloren
      - Schwerwiegend – Die Bedeutung ist teilweise verfälscht
      - Kritisch – Die Bedeutung ist inhaltlich falsch
  - Kategorien der sprachlichen Qualität:
    - Stil: Slator definierte diese Kategorie folgendermaßen: „Der Stil oder Tonfall der Untertitel wirkt für Live-Untertitel ungeeignet. Beispiele: Die Untertitel sind zu ausführlich, zu formell oder zu informell, eher wie geschriebene Sprache, usw." Die möglichen Einstufungen wurden wie folgt definiert:
      - Keine – Kein erkennbares Problem
      - Geringfügig – Wahrnehmbar, aber gut lesbar
      - Schwerwiegend – Lenkt ab oder beeinträchtigt die Lesbarkeit
    - Grammatik: Slator definierte diese Kategorie folgendermaßen: „Der Satz verstößt gegen die grammatischen

Regeln der Zielsprache: Beispiele: Kongruenzfehler, falsche Zeitform oder fehlerhafte Konjugation.“ Die möglichen Einstufungen wurden wie folgt definiert:

- Keine – Kein erkennbares Problem
- Geringfügig – Wahrnehmbar, aber gut lesbar
- Schwerwiegend – Lenkt ab oder beeinträchtigt die Lesbarkeit

- Rechtschreibung: Slator definierte diese Kategorie folgendermaßen: „Tippfehler oder Rechtschreibfehler, die das Lesen beeinträchtigen. Beispiele: Rechtschreibfehler, falsche diakritische Zeichen, fehlerhafte Groß- und Kleinschreibung.“ Die möglichen Einstufungen wurden wie folgt definiert:
  - Keine – Kein erkennbares Problem
  - Geringfügig – Wahrnehmbar, aber gut lesbar
  - Schwerwiegend – Lenkt ab oder beeinträchtigt die Lesbarkeit

- Kommentare der Sprachexperten: Slator bat die Linguisten, ihre Bewertungen auf Englisch zu begründen. Dadurch konnte Slator sprach- und plattformübergreifende Gegenprüfungen durchführen, um sicherzustellen, dass die Anweisungen korrekt befolgt wurden und die Bewertungen einheitlich sind. Dies veranlasste die Linguisten zudem dazu, ihre Bewertungen bewusst zu vorzunehmen und die Konsistenz während der Evaluierung zu erhöhen. Zudem erhielten wir auf diese Weise qualitatives Feedback zu den übersetzten Untertiteln über verschiedene Sprachen und Plattformen hinweg.

Die Sprachexperten wurden angewiesen, die Scorecard zur Qualitätsbewertung ausschließlich auf der Grundlage der finalen bzw. stabilisierten übersetzten Untertitel auszufüllen.

Darüber hinaus bat Slator die Linguisten, die Plattformen nach Benutzerfreundlichkeit/Präferenz sowie nach Qualität zu bewerten. Slator nutzte diese Bewertungen, um daraus abzuleiten, welche Plattform von Sprachexperten bevorzugt wird und welche Plattform die schlechteste wahrgenommene Benutzerfreundlichkeit bietet.

Bewertung der menschlichen Evaluierungen

Datenbereinigung und Gewichtung

Da jede Sprachkombination von zwei unabhängigen Sprachexperten bewertet wurde, überprüfte Slator die Bewertungen auf Konsistenz und Abweichungen, bevor die Ergebnisse zusammengefasst wurden. Die Bewertungen wurden über alle Evaluierenden hinweg gemittelt, um durchschnittliche Segmentwerte pro Plattform und Sprachkombination zu berechnen. Zwar sind bei der menschlichen Bewertung geringfügige Abweichungen bei der Einstufung des Schweregrads zu erwarten, doch bestätigte die Gegenprüfung ein hohes Maß an Übereinstimmung bei der Fehlerklassifikation über alle Evaluierenden hinweg.

Slator wandelte die Qualitätsbewertungen in numerische Werte um (Keine [0 Punkte], Geringfügig [1 Punkt], Schwerwiegend [3 Punkte], Kritisch [5 Punkte]).

Die Kategorien zur Übersetzungsgenauigkeit (d. h. Fehlübersetzungen, Auslassungsfehler und Hinzufügungsfehler) wurden zu einem Gesamtwert für die Genauigkeit zusammengefasst.

Die Kategorien zur sprachlichen Qualität (d. h. Stilfehler, Recht- und Grammatikfehler), erhielten ebenfalls einen Gesamtwert für die sprachliche Qualität.

Diese beiden Werte – Übersetzungsgenauigkeit und sprachliche Qualität – wurden anschließend zu einem Gesamtwert pro Segment kombiniert. Die Bewertungen verschiedener Linguisten für dieselbe Sprache wurden zusammengefasst und gemittelt, um Durchschnittswerte für die Kategorien Genauigkeit, sprachliche Qualität und den Gesamtwert pro Segment zu ermitteln.

Im Gesamtwert pro Segment wurde die sprachliche Qualität mit einem Gewichtungsfaktor von 0,5 berücksichtigt. Dadurch wurde verhindert, dass ein Segment aufgrund eines einzigen schwerwiegenden Rechtschreib-, Grammatik- oder Stilfehlers als „nicht bestanden“ bewertet wird. Gleichzeitig wurde der Gesamtwert zugunsten von Genauigkeitsaspekten gewichtet, da Fehler wie Fehlübersetzungen oder Auslassungen von Sprachexperten als gravierender für die Gesamtqualität wahrgenommen werden.

Datenkategorisierung

Der Gesamtwert pro Segment zeigte zudem an, ob ein Segment als „Bestanden“, „Bestanden mit Mängeln“ oder „Nicht bestanden“ eingestuft wurde. Dies ermöglichte es Slator, die Gesamtquote nicht bestandener Segmente über Plattformen und Sprachen hinweg zu berechnen.

Aufgrund der oben beschriebenen Gewichtung zugunsten der Genauigkeit wird ein Segment automatisch als „nicht bestanden“ klassifiziert, wenn eine der folgenden Bedingungen erfüllt ist:

- Ein Segment, bei dem es um die Genauigkeit geht (Fehlerübersetzung, Auslassung oder Hinzufügung), wird als „kritisch“ gekennzeichnet;
- ein Segment, bei dem es um die Genauigkeit geht, wies zwei oder mehr „schwerwiegende“ Fehler auf (z. B. eine schwerwiegende Auslassung und eine schwerwiegende Hinzufügung);
- eine Fehlerübersetzung wird als „schwerwiegend“ (oder gemäß Punkt 1 als „kritisch“) gekennzeichnet

Ein Abschnitt wird als „Bestanden mit Mängeln“ eingestuft, wenn die folgenden Bedingungen erfüllt sind:

- Ein Segment mit sprachlichen Fehlern (Rechtschreibung, Grammatik, Stil) wird als „schwerwiegend“ eingestuft, oder
- ein Segment mit Genauigkeitsfehlern wird als „schwerwiegend“ eingestuft (ohne eine Einstufung als „nicht bestanden“ auszulösen), oder
- es liegen insgesamt mehr als drei geringfügige Fehler über alle Kategorien hinweg vor

Wenn lediglich ein oder zwei geringfügige Fehler vorliegen, gilt das Segment als bestanden.

Datenbewertung

Die Gesamtwerte pro Segment wurden für jede Plattform und Sprachkombination berechnet und ermöglichten einen direkten Vergleich zwischen diesen.

Die gewichteten Gesamtwerte pro Segment sind wie folgt (je niedriger der Wert, desto besser das Ergebnis):

Sprach-kombination	Google Meet	Microsoft Teams	Zoom	DeepL Voice für Teams	DeepL Voice für 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

Die Anteile für „Nicht bestanden“ bzw. „Bestanden“ wurden berechnet, indem die Gesamtzahl der Segmente in jeder Kategorie („Nicht bestanden“, „Bestanden mit Mängeln“, „Bestanden“) ermittelt wurde, wie folgt:

Tool	Anteil nicht bestandener Segmente	Anteil bestandener Segmente mit Mängeln	Anteil bestandener Segmente
Google Meet	307	603	555
Microsoft Teams	229	586	650
Zoom	216	590	659
DeepL Voice für Teams	67	228	1170
DeepL Voice für Zoom	50	274	1141

Diese Zahlen wurden durch die Gesamtzahl der Segmente aller Sprachen (1.465) geteilt, um die Anteile als Gesamtprozentsätze zu generieren:

Tool	Anteil nicht bestandener Segmente	Anteil bestandener Segmente mit Mängeln	Bewertung Anteil bestandener Segmente
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 für Teams	4,6 %	15,6 %	79,9 %
DeepL Voice für Zoom	3,4 %	18,7 %	77,9 %

Datenverarbeitung

Um einen klaren Vergleich zwischen Tools und Sprachrichtungen zu ermöglichen, wurden die Gesamtwerte pro Segment in einen normierten Qualitäts-Score auf einer Skala von 0–100 überführt. Für jedes Tool wurde zunächst die durchschnittliche Strafpunktzahl pro Segment berechnet, definiert als die Gesamtwerte pro Segment über alle bewerteten Segmente (über alle Sprachkombinationen und beide Richtungen hinweg), geteilt durch die Gesamtzahl der bewerteten Segmente.

Die Fehlerpunkte wurden gemäß dem vordefinierten Bewertungsrahmen vergeben (Kategorien für Übersetzungsgenauigkeit und sprachliche Qualität mit Gewichtung nach Schweregrad). Das bedeutet, dass die theoretisch maximal mögliche Strafpunktzahl pro Segment 24 Punkte beträgt – 15 aus den Genauigkeitskategorien und 9 aus den Kategorien zur sprachlichen Qualität. Der normierte Wert wurde anschließend wie folgt berechnet: 100 × (1 – durchschnittliche Strafpunktzahl pro Segment / 24).

Nach dieser Berechnung entspricht ein Wert von 100 einem System ohne beobachtete Fehler. Ein Wert von null entspricht dem theoretischen Extremfall, bei dem jedes Segment die maximal mögliche Strafpunktzahl erreicht.

Die Daten sind wie folgt:

Tool	Gewichteter Gesamtwert	Gesamtzahl der Segmente	Durchschnittliche Strafpunktzahl	Normierter Wert von 0 bis 100
Google Meet	4534	1456	3,09	87,10537543
Microsoft Teams	3828	1456	2,61	89,11333902
Zoom	3592	1456	2,45	89,78313424
DeepL Voice für Teams	1286	1456	0,88	96,34172355
DeepL Voice für Zoom	1277	1456	0,87	96,36732082

Die ermittelten Tool-Werte (87–96) spiegeln die Leistung im Verhältnis zu dieser maximalen theoretischen Fehlergrenze wider und gewährleisten so die Vergleichbarkeit zwischen verschiedenen Tools, Sprachen und künftigen Bewertungen.

Automatisierte Qualitätsbewertungen

Wir haben eine automatisierte Messverfahren entwickelt, das quantifiziert, wie stabil oder instabil Live-Untertitel im Zeitverlauf auf dem Bildschirm in Google Meet, Microsoft Teams, Zoom, DeepL Voice für Microsoft Teams und DeepL Voice für Zoom Meetings sowie in mehreren Sprachen mithilfe sprachspezifischer OCR erscheinen.

Kurz gesagt: Das System verarbeitet alle aufgezeichneten Meeting-Videos (je eines pro Sprache und Plattform), extrahiert aus jedem Frame den Untertitelbereich und führt darauf OCR aus, um

den Text zu rekonstruieren, den ein Nutzer in diesem Moment gesehen hätte, und berechnet anschließend Metriken zur Untertitelstabilität sowie einen zusammenfassenden Bericht.

Ziel ist es, die tatsächlich für Nutzer sichtbare Untertitelerfahrung zu messen – nicht das zugrunde liegende ASR-Transkript. Wenn Untertitel flackern, sich neu schreiben oder schwanken, erfasst Slator dies direkt aus den gerenderten Videobildern.

Konkret extrahierte Slator für jede aufgezeichnete Datei etwa 10 Frames pro Sekunde, speicherte diese als Bilddateien und ermöglichte so eine Frame-by-Frame-Analyse dessen, was auf dem Bildschirm über verschiedene Sprachen und Plattformen hinweg angezeigt wurde – einschließlich schneller Neuschreibungen, die im Bruchteil einer Sekunde erfolgen.

Für jeden extrahierten Frame schnitt das System von Slator ein Rechteck aus, das nur den Bereich der Untertitel-Einblendung enthielt, um irrelevante Bildschirminhalte zu reduzieren.

Slator nutzte Tesseract – eine OCR-Engine, die jede Untertitelsprache erkennt –, die auf jedes zugeschnittene Untertitelbild angewendet wurde, um den sichtbaren Untertiteltext zu extrahieren und einen Datensatz pro Frame zu speichern.

Die Rohausgabe der OCR kann aufgrund von Abständen, Unicode-Varianten und unsichtbaren Zeichen variieren. Vor dem Vergleich der Frames normalisiert das System den Text kontrolliert, sodass die Vergleiche tatsächliche sichtbare Änderungen widerspiegeln und nicht durch OCR-Rauschen verfälscht werden.

Die Normalisierung umfasst:

- Unicode-Normalisierung
- Entfernen von Zeitstempeln
- Entfernen von Sprechernamen
- Entfernen von UI-Text (z. B. „Übersetzung als Untertitel ist aktiviert“)
- Zusammenführung mehrfacher Leerzeichen
- Entfernung führender und nachgestellter Leerzeichen
- Entfernung von Zeichen ohne Zeichenbreite
- Sprachgruppenspezifische Verarbeitung (lateinische vs. ostasiatische Schriftsysteme)

Dies trägt dazu bei, dass „Änderungen“ möglichst tatsächliche, für Nutzer sichtbare Neuschreibungen widerspiegeln und gleichzeitig inhaltlich relevante Unterschiede erhalten bleiben.

Das System liest die Ergebnisse der optischen Zeichenerkennung Frame für Frame und vergleicht die Texte im Zeitverlauf, um festzustellen, wann Untertitel aktualisiert werden, wie oft sie geändert werden und ob sie flackern oder schwanken.

Das Messverfahren erstellt eine Zusammenfassung in Form einer CSV-Datei mit Schlüsselkennzahlen pro Video, d. h. der Anzahl aktiver Frames und der Anzahl von Änderungsereignissen (siehe „Bewertung automatisierter Analysen“). Diese Ergebnisse sind so konzipiert, dass ein direkter Vergleich zwischen verschiedenen Plattformen, Sprachen und Testbedingungen (DeepL vs. ohne DeepL) möglich ist.

Bewertung automatisierter Prüfungen

Um die Stabilität von Untertiteln objektiv zu messen, erfasste Slator, wie häufig sich Untertitel von einem Frame zum nächsten verändern. Jede sichtbare Veränderung, einschließlich Wortergänzungen, Neu-Segmentierungen oder kurze Schwankungen bzw. Flackern, wurde als Änderungsereignis gezählt. Konkret wurden folgende Änderungsereignisse erfasst:

- Hinzufügen oder Löschen von Zeichen
- Vervollständigung von Wörtern
- Geringfügige Ersetzungen
- Anpassungen der Zeichensetzung
- Formatierungsänderungen
- Hin- und Herwechseln (Oszillation)
- Textinstabilität
- Wahrnehmbar störendes Flackern

Der Anteil der Frames ohne sichtbare Änderung wurde in einen Stabilitätswert von 0 bis 100 umgerechnet, wobei 100 vollständig stabile Untertitel (keine Änderungen zwischen Frames) bedeutet und niedrigere Werte auf häufigere Aktualisierungen auf dem Bildschirm hinweisen.

Parallel dazu berechnete Slator die Instabilität der Untertitel, definiert als der prozentuale Anteil der Frames, in denen sich der angezeigte Untertitel verändert hat. Während der Stabilitätswert den Anteil visuell stabiler Frames angibt, bildet die Untertitelinstabilität direkt ab, wie häufig Nutzer Aktualisierungen oder Unterbrechungen der Untertitel erleben.

Zusammen bieten der Stabilitätswert und die Untertitelinstabilität ein transparentes, verhaltensbasiertes Rahmenwerk für den plattformübergreifenden Vergleich der Untertitel-Leistung. Sie messen nicht die sprachliche Genauigkeit, die Übersetzungsqualität oder die grammatische bzw. semantische Korrektheit (dies wurde separat durch eine menschliche Bewertung evaluiert).

Rohdaten zur Stabilität

Die Rohdaten lauten wie folgt:

Rangliste der plattformübergreifenden Stabilität

Auf Grundlage dieser Daten konnte Slator die prozentuale Zu- bzw. Abnahme bei der Nutzung von DeepL Voice im Vergleich zur gleichen Plattform ohne DeepL Voice ermitteln.

Tool / Plattform	Durchschnittlicher Stabilitätswert (0–100)	Instabilität der Untertitel
DeepL Voice für Zoom	88,6 %	11,35 %
DeepL Voice für Microsoft Teams	85,8 %	14,20 %
Google Meet	82,5 %	17,50 %
Microsoft Teams	77,3 %	22,75 %
Zoom	74,9 %	25,08 %

Die prozentualen Verbesserungen wurden auf Grundlage der Verringerung der Untertitelinstabilität zwischen DeepL Voice-Produkten und der entsprechenden nativen Plattform berechnet.

Sprachübergreifende Rangliste der Stabilität

Nach Sprachen gegliedert ergeben sich folgende durchschnittliche Stabilitätswerte:

Sprach-kombination	DeepL Voice für Zoom	DeepL Voice für 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	86,8	87,5	77,5	76,7	73,8
EN > IT	88,2	84,7	85,4	72,6	75,6
EN > PT	93,2	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
Durchschnitt	88,6	85,9	82,5	77,3	74,9

Die Untertitelinstabilität – d. h. der prozentuale Anteil der Frames, in denen sich der angezeigte Untertitel verändert hat – sowie die Unterschiede zwischen DeepL Voice-Produkten und Standardlösungen sind wie folgt:

Sprach-kombination	DeepL Voice für Zoom	DeepL Voice für Microsoft Teams	Google Meet	Microsoft Teams	Zoom	Unterschiede zwischen DeepL Voice für Teams und Teams	Unterschiede zwischen DeepL Voice für Zoom und 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 %
Durchschnitt	11,35 %	14,13 %	17,50 %	22,75 %	25,08 %	38 %	55 %

In den letzten beiden Spalten geben positive Prozentwerte die Verbesserung bei der Nutzung von DeepL Voice-Produkten im Vergleich zur Standardlösung an. Negative Prozentwerte weisen darauf hin, dass die Standardlösung im Vergleich zu DeepL Voice-Produkten eine höhere Stabilität aufweist.

Gesamter Datensatz

Nachfolgend ist der vollständige Datensatz für alle Tools und Sprachen dargestellt.

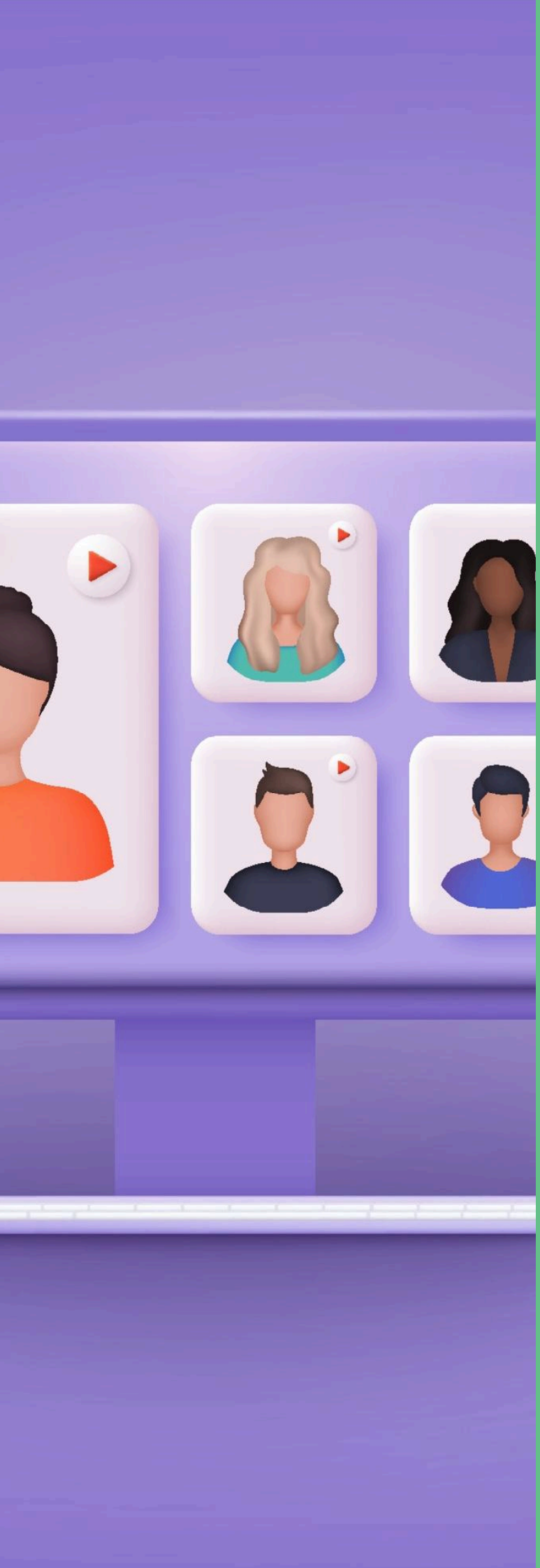
Tool / Plattform	Sprachkombination	Aktive Frames	Frames pro Sekunde	Gesamtzahl der Änderungsereignisse
DeepL Voice für Teams	DE > EN	8987	11	1615
DeepL Voice für Teams	EN > DE	7212	11	901
DeepL Voice für Teams	EN > ES	7167	11	877
DeepL Voice für Teams	EN > FR	7099	10	1015
DeepL Voice für Teams	EN > IT	7005	10	1075
DeepL Voice für Teams	EN > JP	6957	10	1108
DeepL Voice für Teams	EN > KO	7021	10	840
DeepL Voice für Teams	EN > PT	7249	11	852
DeepL Voice für Teams	ES > EN	7057	10	1273
DeepL Voice für Teams	FR > EN	8123	11	1278
DeepL Voice für Teams	IT > EN	7214	10	1265
DeepL Voice für Teams	JP > EN	5921	10	742
DeepL Voice für Teams	KO > EN	5285	10	535
DeepL Voice für Teams	PT > EN	6360	10	758
DeepL Voice für Zoom Meetings	DE > EN	8766	11	1204
DeepL Voice für Zoom Meetings	EN > DE	7112	10	742
DeepL Voice für Zoom Meetings	EN > ES	7246	11	704
DeepL Voice für Zoom Meetings	EN > FR	7192	11	781
DeepL Voice für Zoom Meetings	EN > IT	7157	11	945
DeepL Voice für Zoom Meetings	EN > JP	7241	11	925
DeepL Voice für Zoom Meetings	EN > KO	7272	11	497
DeepL Voice für Zoom Meetings	EN > PT	7239	11	856

DeepL Voice für Zoom Meetings	ES > EN	7268	11	1094
DeepL Voice für Zoom Meetings	FR > EN	7923	10	891
DeepL Voice für Zoom Meetings	IT > EN	7299	10	1095
DeepL Voice für Zoom Meetings	JP > EN	6117	11	579
DeepL Voice für Zoom Meetings	KO > EN	5380	11	392
DeepL Voice für Zoom Meetings	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 Teamss	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

Ableitung der Schlussfolgerungen

Die in diesem Bericht dargestellten Schlussfolgerungen sind unabhängig von DeepL oder anderen Drittanbietern. Slator hat im Rahmen dieser Analyse keine Plattform gezielt oder bewusst bevorzugt und sich um eine vollständig unabhängige und neutrale Auswertung bemüht. Slator hat die Methodik eigenständig entwickelt und die vollständige redaktionelle Kontrolle über Analyse und Ergebnisse behalten.



Über Slator

Slator ist die führende Quelle für Forschungsergebnisse und Marktanalysen in den Bereichen Übersetzung, Lokalisierung, Dolmetschen und KI-Sprachtechnologie. Die Beratungssparte von Slator ist ein zuverlässiger Partner für Kunden, die M&A-Dienstleistungen und unabhängige Analysen in Anspruch nehmen möchten. Slator unterhält Niederlassungen in Zürich (Hauptsitz) und London und verfügt über Analysten in Asien, Europa und den USA.

Projektteam

FLORIAN FAES
Geschäftsführer
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