YouTube Automatic Captions Against Creator Captions: Word Accuracy in Eight Languages, and Speaker Attribution and Timing Compared with a Commercial Transcription Service

Transcribe.so

Research report, version 1, 2026. Captions and transcripts collected 2026-10-26.
https://transcribe.so/papers/youtube-auto-captions-2026-10.pdf

Abstract

We measured YouTube's automatic speech-recognition captions against the caption track each creator uploaded for the same public video, in two studies run on 2026-10-26. Study 1 scored word accuracy on 38 captioned talks, lectures and educational videos in 8 languages. Across the 7 word-scored languages (33 videos), the pooled word error rate (WER) was 14.2% (95% CI 10.7% to 18.5%), ranging from 3.0% in English to 29.0% in Korean; Japanese, scored by character, was 13.5%. Re-uploads of YouTube's own captions were the largest source of unusable references: 27 creator tracks were near-copies of the automatic track and were dropped. Study 2, pre-registered before any run of the comparison system, compared YouTube with transcribe.so, a commercial service sold by the authors, on 30 multi-speaker videos with speaker-marked creator captions and on the Study 1 talks. YouTube's automatic captions carry no speaker labels. transcribe.so attributed 91.9% of reference words to the correct speaker (95% CI 88.7% to 94.7%) on 10 videos with named speakers, and detected creator-marked speaker changes with an F1 of 43.4% (95% CI 32.7% to 53.8%) at a 1 s window. Timing against creator timestamps was close to a tie: transcribe.so was closer on 56.2% of caption lines and YouTube on 40.7%, with a median paired difference of 0.026 s. By disclosure policy, transcribe.so's own word error rates are reported only as bands. Per-video results and video lists are published.

1 Introduction

YouTube generates automatic captions for many uploaded videos, and those captions are widely reused: as subtitles, as a source of searchable text, and as raw material for summaries, notes and translations. Their accuracy is rarely measured in public with a stated method, per-video data and more than one language. This report measures it against the most widely available independent reference on the platform itself, the caption track a creator uploads for the same video.

The report combines two studies. Study 1 asks how far YouTube's automatic track is from the creator's track, word by word, in English, Spanish, French, German, Portuguese, Japanese, Korean and Hindi. Study 2 asks two questions a word error rate cannot answer, who said each word and when it was said, and compares YouTube's captions with transcribe.so on both. Both studies were first published as blog posts on transcribe.so (`https://transcribe.so/blog/youtube-auto-caption-accuracy-2026` and `https://transcribe.so/blog/youtube-captions-speakers-timing-2026`); this report consolidates their method and results in one citable document, with the same figures.

The authors sell transcribe.so, a transcription product that competes with captions as a source of transcripts. The Conflict of interest section sets out that conflict of interest and the design choices made because of it. Study 1 reports YouTube's captions only. Study 2 reports both systems, with transcribe.so's word error rates published as bands rather than exact figures.

2 Related work

Word error rate, the number of substitutions, deletions and insertions in a minimum-edit alignment divided by the number of reference words, is the standard accuracy measure for speech recognition; the alignment is a Levenshtein edit distance [1]. For languages written without spaces between words, character error rate (CER) is the usual substitute. Concatenated minimum-permutation WER (cpWER), which scores words and speaker labels together, was introduced for the CHiME-6 multi-speaker challenge [4]. Speaker diarization is commonly scored with diarization error rate, which needs time-stamped speaker segments in the reference; creator captions do not provide those, so Study 2 uses change-point detection and word-level attribution instead. One-to-one assignments (speaker changes to marked changes, hypothesis speakers to reference speakers) use the Hungarian method [2], and confidence intervals use the percentile bootstrap [3].

We are not aware of a published multi-language measurement of YouTube's automatic captions against creator captions that releases per-video data. We did not conduct a systematic literature search, and this section should not be read as one.

3 Data

3.1 *Corpus A: captioned talks (Study 1)*

We searched YouTube in each of the 8 languages for talks, lectures, news explainers and educational videos, using YouTube's own "Subtitles/CC" and four-to-twenty-minute search filters, and kept public videos up to fifteen minutes long, with at most two videos per channel per language. No cookies or sign-in were used, and age-restricted or members-only videos were not fetched. A first discovery pass with plain keyword search found too few captioned videos and was discarded before any scoring.

A video qualified when YouTube reported its language as the target language; it had a creator-uploaded caption track in that language that was not labelled as a translation; its text was actually in that language (the language check: for Hindi, Japanese and Korean, at least 50% of its letters in the language's script; for the Latin-script languages, more of the target language's common function words than of any other study language's); and YouTube offered an automatic track in the original language. Tracks were fetched on 2026-10-26 with yt-dlp version 2026.10.21. The automatic track was rebuilt from YouTube's word-level caption events, which avoids the duplicated rolling lines of YouTube's VTT export. 98 candidates qualified. After the drop rules (Method, Drop rules), each language kept its first 5 surviving videos in search order, and 38 videos were scored. The flow by language is in the table below; the video list is published (see Data and code availability).

Table 1. Corpus A flow by language. Candidates: videos that met the search and caption-track checks. Surplus: candidates that survived the drop rules but were beyond the first five in search order.

Language	Candidates	Not in language	Rule 1	Rule 2	Rule 4	Surplus	Selected	Rule 3	Scored
English	8	0	0	0	0	3	5	0	5
Spanish	8	0	0	1	0	2	5	0	5
French	8	0	0	0	2	1	5	1	4
German	8	0	0	0	0	3	5	0	5
Portuguese	8	0	1	1	0	1	5	0	5
Japanese	14	0	1	0	4	4	5	0	5
Korean	8	0	0	0	2	1	5	0	5
Hindi	36	9	1	1	19	1	5	1	4
Total	**98**	**9**	**3**	**3**	**27**	**16**	**40**	**2**	**38**

3.2 *Corpus B: multi-speaker videos (Study 2)*

Corpus B holds public interviews, panels, news and talk shows between 4 and 20 minutes long, with a creator caption track in the spoken language that marks speakers, and a YouTube automatic track. We searched in 8 languages with YouTube's captions filter and probed 975 videos; 368 had the required tracks, and most of those did not mark speakers. A video needed at least 5 marked speaker changes in its first 10 minutes, with at most 2 videos per channel; Study 1's rules 1, 2 and 4 and its language check were applied with the same thresholds. Languages with fewer than 3 qualifying videos were excluded before selection: Spanish (2 qualifying), Japanese (0 qualifying), Hindi (0 qualifying). Selection was capped at 30 videos, taking videos with named speakers first. The 30 selected videos are English (9), French (7), German (3), Portuguese (3), Korean (8); 10 of them name their speakers.

Table 2. Corpus B screening by language. Of the probed videos, 368 had the required caption tracks and were screened; the remainder failed a track or language check at probe time.

Language	Probed	No speaker markers	Too few changes	Unusable convention	Drop rules	Language below minimum	Surplus	Selected
English	236	66	3	4	2	0	37	9
Spanish	140	31	1	3	3	2	0	0
French	140	40	1	0	1	0	0	7
German	140	41	0	0	0	0	0	3
Portuguese	140	37	0	0	0	0	3	3
Japanese	60	16	0	0	0	0	0	0
Korean	60	16	0	0	0	0	0	8
Hindi	59	31	0	0	0	0	0	0
Total	**975**	**278**	**5**	**7**	**6**	**2**	**40**	**30**

Each video uses exactly one marker convention, chosen by reading its captions and recorded in the video list: ">>" at the start of a line (2 videos; ">>>", which marks a new story, does not count), a leading dash (18), or a "NAME:" prefix of one to three words that recurs in the video (10). Bracketed non-speech tags never count as a speaker. 2 videos that mixed conventions and 5 with rolling, repeated caption lines were excluded at screening. Study 2 also reuses the 38 corpus A videos, for timing only.

3.3 *Systems*

YouTube: the automatic speech-recognition caption track with its word-level offsets, as served on 2026-10-26. transcribe.so: 68 jobs (all corpus A and corpus B videos) submitted on 2026-10-26 through the public API on a Pro-plan test account, with no settings changed; every job was served by the plan's default pipeline (label "standard"). All jobs completed. transcribe.so outputs were also used in Study 1, but only as an outlier detector for rule 3 (Method, Drop rules), never as a reference.

4 Method

4.1 *Reference and normalization*

The creator-uploaded track is the reference throughout. This makes every result a measure of distance from what the creator published, not from a verified verbatim transcript. Every text was normalized the same way before scoring (normalizer "yt-caption-norm-1"): Unicode compatibility folding, lowercasing, curly quotes folded, punctuation removed, bracketed non-speech tags such as [Music] or (laughs) removed, a short fixed list of filler words removed from both sides, and for Hindi, zero-width joiners and nukta and chandrabindu variants folded. Numbers were not normalized: a number written in digits in one track and in words in the other counts as an error. All parenthesized text is removed, including real parentheticals, and Japanese fillers are removed wherever they occur; both side effects apply to every track equally.

4.2 *Drop rules (Study 1)*

Rules 1 to 4 and their thresholds were written into the results file at 2026-10-26 04:53:02 UTC, before final scoring, and were not tuned afterwards.

- **Language check.** The creator track's text must be in the spoken language. This check was added after a first screening pass had already computed error rates, when Hindi tracks scoring above one hundred percent turned out to be English translations filed under the Hindi code. It removes unusable references but it lowers the Hindi figure, so it is flagged as a post hoc addition. Dropped: 9.
- **Rule 1.** The creator track is far too short for the video: fewer reference units per minute than 40% of a typical speech rate (English 150 words, Spanish 160 words, French 150 words, German 130 words, Portuguese 150 words, Japanese 300 characters, Korean 100 words, Hindi 120 words per minute). Dropped: 3.
- **Rule 2.** The creator and automatic tracks differ too much in length: a length ratio outside 0.7 to 1.3. Dropped: 3.
- **Rule 4.** The creator track is a re-upload of YouTube's own automatic track: under 2% different from it. Dropped: 27, of which 19 were Hindi. Scoring these would grade YouTube against itself.
- **Rule 3.** A three-way outlier: YouTube's track and transcribe.so's transcript are within 15% of each other while both are more than 40% away from the creator track. The creator track is then the likely outlier, for example a heavily edited caption. Rule 3 was applied after selection, to the 40 evaluable videos, and dropped 2 (1 French and 1 Hindi).

Rule 3 can fire only when YouTube's error is high, so it can only lower YouTube's measured error. Every Study 1 result is therefore also reported without rule 3.

4.3 *Word and character error rates*

We computed WER per video (CER for Japanese, and CER additionally for Korean and Hindi, whose word boundaries make WER harsh on small spacing or spelling differences). The headline is the micro-average, total edits over total reference units; the macro-average weights each video equally. Pooled figures cover the word-scored languages; Japanese is reported on its own. In Study 2, the same metric is computed on the first 10 minutes of each corpus B video.

4.4 *Speaker changes and attribution (Study 2)*

A reference speaker change is a caption line carrying the video's marker. A transcribe.so change is the start of a segment whose speaker differs from the previous one. Marked and detected changes are paired one to one by the Hungarian method within windows of 0.5, 1 and 2 seconds (1 s is the headline), maximizing matches and then minimizing total time difference. Creator timestamps are often offset by a constant, so each video's reference times are first shifted by that video's median offset between the creator's line starts and transcribe.so's word starts; unshifted results are reported too, and the 10 videos whose offset exceeded 2 s are flagged in the data. Precision, recall and F1 are pooled over videos.

Before freezing the rules we checked the extractor. We sampled 100 extracted changes and 100 unmarked caption boundaries, removed every marker, and judged from two lines of context on each side whether the speaker changed, without knowing which group a boundary came from. This was a text-only check, not a listening test, and it was judged by the same person who wrote the extraction rules: it was blinded but not independent. 82.9% of the decidable extracted changes read as real changes (18 were too ambiguous to call), and 11.7% of the decidable unmarked boundaries also read as changes. Scaled to the 955 marked changes and 5,516 unmarked boundaries in the window, this puts the references' estimated recall at 55.1%. Measured precision of any system against these references is therefore a lower bound.

Table 3. Blinded text-only judgements of sampled caption boundaries, markers removed.

Sample	Change	No change	Unsure	Total
All marked changes	68	14	18	100
Marked, NAME: prefix	24	2	6	32
Marked, Leading dash	39	12	10	61
Marked, >> prefix	5	0	2	7
Unmarked boundaries	11	83	6	100

Attribution is measured on videos with "NAME:" labels. Both transcripts are aligned word by word to the reference, each transcribe.so speaker is matched one to one to a reference speaker by word overlap, and we count the reference words given the right speaker; reference words the transcript missed count as wrong. A transcript with no speaker labels would count as a single speaker. cpWER is computed as in [4], with unmatched speakers counted in full.

4.5 *Timing (Study 2)*

For each caption line we take the first word each system matched exactly and subtract the line's creator start time. Both systems are compared against the same reference on the same lines: we report the share of lines where each is closer, the median of the paired difference, and a de-biased spread, the ninetieth percentile of each offset around that video's median offset. Hindi is excluded from timing and Japanese is aligned by character, both fixed in advance. A video whose creator timestamps matched YouTube's word timings within 50 ms on more than 60% of lines (auto-synced captions) is excluded, because it would grade YouTube against itself.

4.6 *Statistics and pre-registration*

All 95% intervals are percentile bootstrap intervals over 2,000 resamples of whole videos within each language, with a fixed seed (20261026). They reflect how much a result depends on which videos were sampled, not uncertainty in the reference.

Study 2's rules, extractor check and corpus B selection were committed to the authors' repository on 2026-10-26, before transcribe.so was run on any study video, in commit `10a150cf744c4ccde1e0d5fb56d14bceb89744b6`. One amendment followed on 2026-10-26, in commit `3e91ff09d6a32b4421687e5e790a102785fb7176`; it changed the output list only: "Strategic review adopted by the owner: keep the v1 post verbatim as the neutral YouTube-only asset and publish v2 as a new companion post (/blog/youtube-captions-speakers-timing-2026) linked both ways, instead of retitling and expanding the v1 post. No change to metrics, thresholds, collars, extractor rules or the corpus B selection." No metric, threshold, window or video selection changed. The repository is not public, so the hashes are cited as text. Study 1 was not pre-registered in the same way: its thresholds were fixed before final scoring, but the language check was added after a screening pass, as described above.

5 Results: word accuracy (Study 1)

On captions served on 2026-10-26, YouTube's automatic captions had a pooled WER of 14.2% (95% CI 10.7% to 18.5%) across 33 videos in 7 word-scored languages, against 40,827 reference words. The macro-average was 15.4%. Without rule 3 the pooled WER was 15.9% (95% CI 11.9% to 20.4%, 35 videos).

Table 4. YouTube automatic captions against creator captions, by language (captions fetched 2026-10-26). Reference units are words, or characters for Japanese.

Language	Metric	Videos	Reference units	Micro-average (95% CI)	Macro-average (95% CI)	Without rule 3
English	WER	5	7,842	3.0% (2.4% to 3.7%)	3.0% (2.4% to 3.6%)	same
Spanish	WER	5	6,841	8.7% (4.4% to 13.0%)	8.8% (4.9% to 12.7%)	same
French	WER	4	6,597	12.9% (8.7% to 19.4%)	14.2% (8.5% to 20.0%)	16.9% (5 videos)
German	WER	5	6,620	18.0% (13.2% to 22.6%)	17.1% (12.2% to 22.8%)	same
Portuguese	WER	5	4,513	16.8% (9.2% to 25.2%)	15.8% (8.8% to 23.9%)	same
Japanese	CER	5	12,903	13.5% (7.9% to 19.0%)	13.0% (8.2% to 18.0%)	same
Korean	WER	5	4,060	29.0% (17.1% to 44.0%)	30.8% (18.2% to 44.0%)	same
Hindi	WER	4	4,354	22.9% (2.0% to 49.2%)	18.3% (2.0% to 46.1%)	29.6% (5 videos)
Pooled, 7 word-scored languages	**WER**	**33**	**40,827**	**14.2%** (10.7% to 18.5%)	**15.4%** (11.7% to 19.9%)	**15.9% (35 videos)**

The spread between languages is wide. English was 3.0% (95% CI 2.4% to 3.7%), while Korean was 29.0% by words and Hindi 22.9%. Scored by character, Korean was 11.6% and Hindi 17.0%, so much of their word-level error is in spacing and spelling rather than in whole words. With at most 5 videos per language, every per-language figure is indicative; the pooled figure and its interval carry the main result.

Table 5. Character error rate for the two word-scored languages where word boundaries make WER harsh.

Language	Videos	Reference characters	CER micro (95% CI)	CER macro (95% CI)	Without rule 3
Hindi	4	16,185	17.0% (0.9% to 37.0%)	12.7% (0.9% to 34.9%)	22.8% (5 videos)
Korean	5	11,469	11.6% (4.6% to 22.7%)	12.6% (4.9% to 23.0%)	same

9 of the 38 scored videos had an error rate under 5%. Some of these may be lightly corrected re-uploads of YouTube's own captions that rule 4 cannot catch, which would make YouTube look more accurate than it is.

6 Results: speakers and timing (Study 2)

6.1 *Speaker labels*

YouTube's automatic captions are a stream of words with timestamps and no speaker field, so every speaker metric for YouTube is "not available" rather than zero. This is a property of the format, not a measurement. transcribe.so attached a speaker label to 100% of the words it returned on the 30 corpus B videos; 0 transcripts came back without labels. A label being present does not mean it is right, which the next two measures address.

6.2 *Speaker-change detection*

At a 1 s window, transcribe.so's speaker changes matched the 955 creator-marked changes with an F1 of 43.4% (95% CI 32.7% to 53.8%): 41.9% of its changes landed on a marked change, and it found 45.0% of the marked changes. The F1 was 55.9% with a wider window and 32.0% with a narrower one, and 39.0% without the per-video offset shift at the headline window.

Table 6. transcribe.so speaker-change detection against creator-marked changes, 1 s window, reference shifted by each video's median offset. YouTube: not available (the caption format carries no speaker labels).

Language	Videos	Marked changes	Detected changes	Precision (95% CI)	Recall (95% CI)	F1 (95% CI)
English	9	165	170	51.2% (32.4% to 72.5%)	52.7% (35.7% to 70.2%)	51.9% (33.3% to 69.6%)
French	7	275	300	47.3% (26.9% to 76.5%)	51.6% (30.9% to 68.2%)	49.4% (30.8% to 70.5%)
German	3	62	133	36.8% (7.4% to 61.5%)	79.0% (50.0% to 88.9%)	50.3% (12.9% to 72.7%)
Portuguese	3	121	45	17.8% (8.3% to 55.6%)	6.6% (0.0% to 60.0%)	9.6% (0.0% to 52.6%)
Korean	8	332	379	38.0% (13.4% to 62.5%)	43.4% (25.3% to 52.1%)	40.5% (17.7% to 55.2%)
All languages	**30**	**955**	**1,027**	**41.9%** (29.4% to 55.2%)	**45.0%** (33.9% to 56.2%)	**43.4%** (32.7% to 53.8%)

Table 7. Sensitivity of pooled speaker-change detection (30 videos) to the matching window and to the per-video offset shift.

Window	Shifted precision	Shifted recall	Shifted F1	Unshifted precision	Unshifted recall	Unshifted F1
0.5 s	30.9%	33.2%	32.0%	27.0%	29.0%	28.0%
1 s	41.9%	45.0%	43.4%	37.6%	40.4%	39.0%
2 s	53.9%	58.0%	55.9%	48.0%	51.6%	49.8%

This is the weakest result for transcribe.so, and the reference pulls it down in three ways: creators leave many changes unmarked (estimated recall 55.1%), short interjections are often unmarked, and one Portuguese video with many dash-marked changes, in which transcribe.so heard a single speaker throughout, accounts for most of Portuguese's low score. Even so, roughly half of the changes a careful captioner marked were not found within the headline window, and some detected changes were real errors.

6.3 *Who said what*

On the 10 videos whose captions name speakers, 9 English and 1 German, transcribe.so attributed 91.9% of reference words to the correct speaker (95% CI 88.7% to 94.7%). Counting only the reference words the transcript aligned to, the share was 94.2%. cpWER, which includes every word error, was in the band "15% or higher"; it cannot be lower than the word error rate itself.

Table 8. Who said what on the videos whose captions name speakers. Reference words missing from the transcript count as misattributed. YouTube: not available (the caption format carries no speaker labels). A single-video row has no interval.

Language	Videos	Words attributed to the right speaker (95% CI)	cpWER (band)
English	9	93.0% (89.6% to 96.1%)	15% or higher
German	1	77.0%	15% or higher
All	**10**	**91.9%** (88.7% to 94.7%)	**15% or higher**

6.4 *Timing*

Across 56 videos, transcribe.so's first matched word of a caption line was closer to the creator's timestamp on 56.2% of lines and YouTube's on 40.7%; ties make up the rest. The median paired difference was 0.026 s (95% CI 0.000 to 0.042 s). On the talks it was close to an even split (50.2% against 46.8%); on the multi-speaker videos transcribe.so was closer on 61.4% of lines while YouTube had the tighter tail (1.62 s against 1.75 s), with overlapping intervals. We read timing as a draw with a small edge each way. 9 videos with auto-synced creator timestamps were excluded as fixed in advance; 1 video met both exclusion criteria, so the exclusion rows do not sum.

Table 9. Timing of the first exactly matched word of each caption line against the creator timestamp. Paired difference: |YouTube offset| minus |transcribe.so offset| per line, positive when transcribe.so is closer. Spread: 90th percentile of each offset around its video's median offset.

Measure	Corpus A (talks)	Corpus B (multi-speaker)	Pooled
Videos considered	38	30	68
Excluded: auto-synced reference	6	3	9
Excluded: Hindi or no word timing	4	0	4
Videos compared	29	27	56
Paired caption lines	3,730	4,440	8,170
Lines where transcribe.so is closer	50.2%	61.4%	56.2%
Lines where YouTube is closer	46.8%	35.6%	40.7%
Median paired difference (95% CI)	0.001 s (-0.015 to 0.037 s)	0.040 s (0.008 to 0.061 s)	0.026 s (0.000 to 0.042 s)
Spread p90, YouTube (95% CI)	0.53 s (0.43 to 0.65 s)	1.62 s (1.21 to 2.09 s)	1.15 s (0.88 to 1.43 s)
Spread p90, transcribe.so (95% CI)	0.53 s (0.43 to 0.64 s)	1.75 s (1.36 to 2.36 s)	1.28 s (0.97 to 1.59 s)
Raw median |offset|, YouTube	0.159 s	1.280 s	0.302 s
Raw median |offset|, transcribe.so	0.125 s	1.275 s	0.272 s

6.5 *Word error rate on multi-speaker videos*

On corpus B, YouTube's pooled WER was 30.2% (95% CI 24.8% to 35.6%). transcribe.so's pooled WER was in the band "15% or higher". YouTube's figure falls inside that band, so these results do not show which system made fewer word errors on these videos. Several English references are live-stenography captions with typos and omissions, which raises every error rate measured against them.

Table 10. Word error rate on corpus B, first 10 minutes of each video. transcribe.so is shown only as a band (under 5%, 5-15%, 15% or higher) by policy.

Language	Metric	Videos	Reference words	YouTube (95% CI)	transcribe.so (band)
English	WER	9	10,347	30.6% (20.7% to 42.4%)	15% or higher
French	WER	7	8,709	23.2% (13.0% to 34.3%)	15% or higher
German	WER	3	3,277	39.1% (32.1% to 46.4%)	15% or higher
Portuguese	WER	3	4,539	17.6% (5.6% to 28.3%)	5-15%
Korean	WER	8	5,900	44.6% (35.0% to 58.3%)	15% or higher
Pooled	**WER**	**30**	**32,772**	**30.2%** (24.8% to 35.6%)	**15% or higher**

7 Limitations

- **The reference is not ground truth.** Creator captions are sometimes tidied, paraphrased or abridged, and those edits count as errors for the system being scored. In Study 2, creators mark speaker changes inconsistently, so speaker-change scores are a comparison with a careful but incomplete human captioner.
- **Small samples.** Study 1 has at most 5 videos per language and Study 2's attribution result rests on 10 videos, most in English and several of them interview programmes labelled "HOST:" and "GUEST:". Per-language figures are indicative.
- **Selection.** Every video had creator-uploaded captions, which skews toward channels that invest in captioning: talks, lectures, explainers and broadcast programmes. Casual uploads, music, gaming and noisy field recordings were not measured and are likely harder. Corpus B's speaker-marked captions are mostly news, public-affairs and television content. Spanish, Japanese and Hindi had too few speaker-marked videos for Study 2.
- **Drop rules shape the result.** Rule 4 removes exact re-uploads of YouTube's track but not lightly corrected ones. Rule 3 uses transcribe.so as an outlier detector and can only lower YouTube's error; both figures are reported. The language check was added after a screening pass.
- **Numbers and parentheticals.** Numbers are not normalized, and all parenthesized text is removed, which affects both systems equally but changes absolute error rates.
- **Coarse timing.** Creator timestamps mark the start of a line, sometimes seconds after the speech in live captions. Timing results say which system is closer to the same reference, not how close either is to the moment a word was spoken.
- **Banded figures.** transcribe.so's WER and cpWER are published only as bands, so the word-accuracy comparison between the two systems is coarse by design.
- **One point in time.** Both systems change. These are results for captions and transcripts produced on 2026-10-26.

8 Conflict of interest

This report was written, and both studies were designed, run and analysed, by Transcribe.so, which sells transcribe.so, a transcription product that competes with YouTube's captions as a source of transcripts. No external party funded, reviewed or audited the work. The design choices made because of this interest are:

- Study 1 reports YouTube's captions only. transcribe.so was used solely as an outlier detector (rule 3), and Study 1 is reported with and without that rule.
- Study 2's rules, extractor check and video selection were committed before transcribe.so was run on any study video (see Statistics and pre-registration), and the single amendment changed only how results were published.
- The extractor check was blinded to whether a boundary was marked, but it was performed by the same person who wrote the extraction rules, so it was not independent.
- transcribe.so's word error rate and cpWER are published as bands (under 5%, 5-15%, 15% or higher), because exact word-accuracy digits for our pipeline can identify the speech-recognition vendors behind it. Speaker-label, speaker-change, attribution and timing figures for transcribe.so are exact. YouTube's figures are exact throughout.
- The weakest result for transcribe.so (speaker-change detection) is reported in full, alongside the reference problems that depress it.

9 Data and code availability

The following files are published on transcribe.so. Caption text is not republished; the video lists give YouTube IDs so the tracks can be fetched again. transcribe.so's word error rates appear in these files as bands only.

- Study 1, per-language results (CSV): `https://transcribe.so/data/youtube-auto-caption-accuracy-2026-10.csv`

- Study 1, per-video results, thresholds and drop counts (JSON): `https://transcribe.so/data/youtube-auto-caption-accuracy-2026-10-results.json`
- Study 1, scored videos (JSON): `https://transcribe.so/data/youtube-auto-caption-accuracy-2026-10-videos.json`
- Study 2, per-language and per-metric results (CSV): `https://transcribe.so/data/youtube-caption-study-2026-10-v2.csv`
- Study 2, pre-registered rules, extractor check, per-video scores and amendment log (JSON): `https://transcribe.so/data/youtube-caption-study-2026-10-v2-results.json`
- Study 2, corpus B videos with marker conventions (JSON): `https://transcribe.so/data/youtube-caption-study-2026-10-v2-videos-b.json`

The discovery, fetching and scoring code is in the authors' repository, which is not public at the time of writing. Every result figure and count in this report was generated by a script from the results files above, not typed by hand.

References

1. Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8), 707-710.
2. Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2), 83-97.
3. Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1), 1-26.
4. Watanabe, S., Mandel, M., Barker, J., Vincent, E., et al. (2020). CHiME-6 Challenge: tackling multispeaker speech recognition for unsegmented recordings. *Proceedings of the 6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*.

How to cite

Transcribe.so (2026). *YouTube Automatic Captions Against Creator Captions: Word Accuracy in Eight Languages, and Speaker Attribution and Timing Compared with a Commercial Transcription Service.* Research report. `https://transcribe.so/papers/youtube-auto-captions-2026-10.pdf`

```
@techreport{transcribeso2026youtubecaptions,
  author      = {{Transcribe.so}},
  title       = {YouTube Automatic Captions Against Creator Captions: Word Accuracy in Eight Languages, and
Speaker Attribution and Timing Compared with a Commercial Transcription Service},
  institution = {Transcribe.so},
  year        = {2026},
  url         = {https://transcribe.so/papers/youtube-auto-captions-2026-10.pdf}
}
```