Exbabel Translate

RESEARCH PRESENTATION

Evaluation of Real-Time Speech Translation

United Nations General Debate

A comparative study of two Exbabel translation models and live UN interpretation

EXBABEL RESEARCH / NOVEMBER 2026



Exbabel Flagship scored 4 points higher than the UN interpreter in blind review.

91.3

Exbabel Flagship Model

Average meaning score / 100

87.1

Exbabel Translation Model

Average meaning score / 100

87.3

UN interpreter

Average meaning score / 100

On the same 55 clips, blind AI reviewers scored Flagship higher on 34 and the interpreter higher on 21. The Exbabel Translation Model and interpreter had nearly identical averages.

The figures are judgments of meaning on selected transcripts. They do not measure the percentage of correct translations.

Three tests answer three different questions

Recorded speeches	Could the system deliver translated speech from the delegate's microphone? We sent the same 180 recordings to both engines.
Meaning and wording	For 66 clips we compared each output with the interpreter's wording. For 55 we also judged all three versions against the delegate's original meaning.
Written material	We translated UN sentences as text and sent synthetic readings of the same material through the speech systems. We also matched prepared statements to live speeches.

From 180 excerpts to 55 fairer comparisons

180

Original floor clips

14 delegations · 15–23 seconds

66

Matched passages

Both alignment models agreed

55

Scored passages

A further check supported both ends

Interpreters finish a thought after the delegate. Cutting both channels at the same second can leave out part of the message. Most exclusions came from disagreement about where the matching passage began or ended.

Source: UN Transcription Corpus, 80th General Assembly General Debate, September 2025. Thirty original clips per direction.

How the blind reviewers scored meaning

Hide the identities

Three AI reviewers received shuffled candidates. Two rated each translation against the speaker’s transcript, including the UN version.

Score the message

100 meant all recoverable meaning survived. At 75 the main point survived with a local error. At 50 roughly half survived or a serious change affected part of the message.

Flag what mattered

Reviewers separately flagged changes to who acted, numbers, timing, negation, obligations and missing ideas. A faithful paraphrase could earn full credit.

55 clips × 3 translations × 2 reviews = 330 ratings. Reviewers agreed within 10 points on 162 of 165 translations. All shared one model family.

The lead changed from clip to clip

Exbabel Flagship Model versus the UN interpreter



34 clips · Flagship scored higher

21 · Interpreter scored higher

Flagship averaged 4.0 points above the interpreter (95% interval +0.4 to +6.1). The Translation Model averaged 0.2 points below (−4.1 to +2.1). Its result is consistent with a lead in either direction.

Exploratory intervals use 10,000 paired resamples of delegations within language directions. They do not capture shared reviewer bias.

How did the six language directions compare?

Speaker → listener	Clips	UN interpreter	Exbabel Flagship	Exbabel Translation
English → French	6	81.2	86.1	82.7
French → English	14	85.6	90.8	92.0
Spanish → English	7	92.1	99.1	86.6
Arabic → English	7	82.9	91.1	83.4
Russian → English	11	92.7	91.7	87.2
Mandarin → English	10	87.0	89.2	86.0
All directions	55	87.3	91.3	87.1

Mean score out of 100. Six to fourteen clips per row, with one delegation for Mandarin. The overall mean weights each scored clip equally.

“No menores” means just as important

UN interpreter	“There are other just as important challenges.”
Exbabel Flagship Model	“There are other challenges too, no less important.”
Exbabel Translation Model	“There are other challenges, minor ones.”

The first two keep the comparison. “Minor” tells listeners to give the other challenges less weight.

Spain · GA80-ES-8090-es-en · excerpts from the transcripts used for scoring.

Small changes can carry different consequences

Mandarin · a priority	“Put people first” and “advocating the people-centered approach” expressed the same priority. The shorter phrase deserved credit.
Russian · a time span	The interpreter and Translation Model said “the past quarter century.” The Flagship transcript said “Last quarter,” changing 25 years to three months.
Arabic · a financing term	The Translation Model caption kept “concessional financing.” The audio transcript read “consensus.” A bilingual listener needs to check the recording before attributing the error.

Clips CN-6647, BY-10982 and EG-10503. The dossier casebook includes the source, candidate transcripts and further examples.

How closely did the models follow UN wording?

Into English	Matched clips	Flagship COMET	Translation COMET
Spanish	10	0.595	0.639
Arabic	7	0.693	0.569
Russian	11	0.656	0.715
Mandarin	13	0.640	0.640

On Spanish, the Translation Model led on reference-based COMET. Flagship led on the smaller blind meaning test. Both the question and the selected clips changed between tests.

66 clips across six directions. COMET uses source, candidate and reference. chrF++ and BLEU compare character and word sequences.

Clean text was easier in the French reading test

French → English · 5 sentences	Flagship family	Translation family
Written input · COMET	0.915	0.911
Spoken reading · audio transcript COMET	0.687	0.429

The same prepared sentences scored higher as written input. Reading them aloud added a synthetic voice, different engines and speech recognition. This design cannot assign the gap to any one step.

35 items from 28 UN Parallel Corpus development rows. Written branch: 65/65 outputs. Separately, 12 prepared-statement passages yielded 24 text translations.

Which systems delivered audio on the first try?

180 / 180

Flagship first attempts

All sessions confirmed completion

16 / 180

Translation first attempts

176 recordings after retries

30 / 30

English → Arabic

Arabic voice verified after recovery

Across the six directions, median time to the first audio packet ranged from 0.53–0.85 seconds for Flagship and 2.84–3.18 seconds for the Translation Model. A packet can contain silence.

No Translation Model primary session supplied the expected final-turn signal by the deadline. English → Arabic had no three-way meaning score.

What reached the Exbabel worker?

Routing	Seven isolated worker route probes returned speech after recovery, including the explicit Arabic voice route.
Custom instructions	One French clip was run with and without a terminology instruction. Both runs produced audio. One example cannot quantify the benefit of prompting.
Difficult conditions	One Bahamas speech supplied noise, interruptions and an 80-second continuous probe. Listener comprehension, production failover and long-session performance need broader testing.

The 180-clip benchmark ran the underlying engines directly. The deployed audience stream was outside this comparison.

Where to inspect the evidence

Hear the comparison	Listening-Comparison.html pairs the original speaker, interpreter context and model audio. Translation-Choices.pdf documents cases across speech, written and read-aloud material.
Check the scores	Reports-and-workbooks contains the interpreter comparison and source-fidelity workbooks. Methods-and-sources contains the scoring rubric, language matrix and source registers.
Trace the recordings	The UN Transcription Corpus supplied audio. The UN Parallel Corpus supplied written material. The source-based scoring rubric was adapted from IWSLT 2024, Appendix A.2.

UN audio corpus

UN Parallel Corpus

IWSLT scoring method

Meaning survived often. Details still mattered.

What we learned

A good translation could use different words from the interpreter. Flagship scored highest on average in this sample, while the interpreter led on 21 of their 55 comparisons.

What needs a listener

Both judges flagged consequential issues on 12 Flagship, 22 Translation Model and 15 interpreter clips. Audio cuts and transcription can affect any of the three.

What to test next

Have independent bilingual listeners assess full Exbabel audio on fresh speeches. Agree on tolerances for critical errors, listener comprehension and latency before scoring.

Only 55 of 180 clips qualified for the primary comparison. These estimates apply to the selected excerpts and review method.