Exbabel translation research

> **Exbabel model configurations.** Exbabel Flagship Model denotes the OpenAI-backed speech engine. Exbabel Translation Model denotes the Gemini-backed speech engine. Exbabel uses custom prompting in its translation system. The 180-clip comparison measured the underlying engines directly. The isolated worker probes tested Exbabel routing separately; one French clip compared the worker with and without a terminology instruction. The study did not estimate a general improvement from custom prompting.

Diplomatic speech translation feasibility

Exbabel Events / Evidence report / 20 November 2026

Seven live engineering runs returned audio. The primary UN speech pilot has zero eligible segments, so diplomatic accuracy and readiness remain undetermined.

Scope	Executed	Whattheevidenceestablishes
UNspeechpilot	0of180segments	NomeasuredperformanceonauthenticUNspeech .
Engineeringstimulus	1sentence 7attempts ,	Bothprovidersandthelocalproductionworkeremitted
		target -language audio .
Publicresources	206corpusrecords	Metadatasupportscandidatediscovery Audiotransferrequires .
		separateclearanceforthisuse .

Research question

Can Exbabel preserve meaning, names, numbers and obligations while interpreting diplomatic speeches in real time? The engineering runs establish access, audio capture and observable stream behavior. Answering the research question requires authenticated speeches, aligned references and bilingual adjudication.

Decision for the stakeholder presentation

Present the recorded engineering comparison and the sampling constraints. Reserve any claims about interpretation quality, professional-interpreter equivalence or production readiness for the scored pilot.

Sources and six-language coverage

The General Debate inventory contains 192 records from session 80. The meeting inventory contains 14 records with document symbols, channel links and official-record fields. Metadata identifies a candidate channel. A bilingual listener must confirm the original speaker and the segment boundary.

Declaredsource	Debaterecords	Prioritytarget	Eligiblesegments
English	100	French	0
French	19	English	0
Spanish	12	English	0
Arabic	13	English	0
Russian	2	English	0
Mandarin/zh	0	English	0

The remaining 46 debate records carry other, missing or ambiguous language labels. Country identity cannot establish spoken language. Find additional verified Mandarin and Russian speakers before any language-level comparison.

Source-use decisions

The dataset card distinguishes UN audio terms from the metadata license. The public grant and Web TV notice do not establish permission for commercial vendor transfer or stakeholder playback. All recordings and human interpretation channels therefore remain outside this execution.

The text workstream contains 150 bilingual alignments sampled from five UNPC training configurations. The official fixed test archive returned HTTP 403. The glossary contains 117 designations from 12 UNTERM entries with record-specific status. Both resources support development. Neither provides a time-aligned speech reference.

Observed engineering results

One original 8.066-second English eSpeak NG stimulus described a meeting time and a sound check. Identical source audio was streamed at normal speed. Exbabel Flagship Model received 24 kHz PCM16 and Exbabel Translation Model 16 kHz PCM16. The worker received 24 kHz frames. All captured WAVs contain 24 kHz mono PCM.

Condition	Direction	Caption(s)	Audio(s)	Termination
Exbabel Flagship Model	ENtoFR	130 .	069 .	Finalclose
Exbabel Translation Model	ENtoFR	270 .	273 .	Draintimeout
Exbabeldefault	ENtoFR	152 .	054 .	Clientstopped
Exbabel Flagship Model	ENtoAR	130 .	063 .	Finalclose
Exbabel Translation Model	ENtoAR	275 .	279 .	Draintimeout
Exbabeldefault	ENtoAR	160 .	069 .	Clientstopped
ExbabelforcedGemini	ENtoAR	287 .	289 .	Clientstopped

Times measure the first caption delta and first audio packet from the first input send. A packet can carry silence. These observations have one stimulus per condition. They do not estimate meaningful phrase latency, p95 service latency or a comparative speed advantage.

Account behavior and completion

Exbabel Flagship Model returned a final session close in both directions. Its published output list omits Arabic, yet independent ASR detected Arabic in the standalone and routed recordings. Both Exbabel Translation Model baseline sessions timed out awaiting a final turn. Their partial evidence remains available. The local worker was stopped after a fixed drain window in all three runs.

The local worker uses repository routing, provider setup, audio feeding and output events. The forced Exbabel Translation Model route is an override. Deployed sessions, listener delivery, legacy STT/TTS failover and glossary behavior require separate tests.

Meaning, audio and statistical limits

Independent acoustic check

A frozen faster-whisper-base checkpoint decoded each generated waveform with automatic language detection. It identified French or Arabic as requested in all seven. The French Exbabel Flagship Model decode reproduced the meeting time. Exbabel Translation Model’s French decode ended mid-word while its native caption contained the full ending. A listener must determine whether this reflects the waveform or recognition error.

Outcome	Denominator	Interpretation
NativeBLEU chrF++andCOMET ,	0alignedspeechreferences	Undefinedfortheprimarypilot .
vs humantranslation .		
Meaning criticalerrors terminology , , ,	0adjudicatedsegmentsor	Noaccuracyestimateorconfidenceinterval .
namesandnumbers	instances	
Intelligibility naturalness , ,	0independentlistener	Pendingbilinguallistening .
comprehensionandattribution	ratings	
ASR vs native-caption chrF++ .	7engineeringoutputs	Diagnosticagreementonly Individualvalues .
		areinresults summarycsv _ . .

Planned comparison

Freeze 30 segments in each priority direction with seed 20261120. Separate clean and adverse recordings, held-out speakers and glossary conditions. Two blinded reviewers rate each item independently. A third adjudicates divergences while preserving both originals. Report Wilson intervals for binary outcomes and paired, speaker-clustered comparisons on identical utterances.

Run bounds, backlog, drift, interruptions, induced noise, reconnects and phrase-level timing require aligned observations or controlled impairment tests. The current study has no professional-interpreter comparator or approved acceptance threshold.

Release decisions and evidence index

1 Audio rights and source identity

All 180 planned segments await clearance. Secure provider-transfer and playback rights, then verify floor routing, speaker, target interpretation and exact cut points.

2 Exbabel Translation Model stream completion

Both standalone attempts produced audio after streaming. Neither met the terminal-event requirement. Validate a continuous-stream drain rule against timed content and test the final phrase explicitly.

3 Arabic routing contract

The default Exbabel route selected Exbabel Flagship Model. This account returned Arabic speech beyond the published list. Add a monitored capability check and exercise failover independently of a forced provider override.

4 Audio and captions

Independent ASR disagreed with native captions on the engineering sentence. Ask blinded listeners to check final words, numbers and the reported meeting time before attributing errors to translation.

5 Review and acceptance

Recruit two qualified reviewers per direction and an adjudicator. Approve criteria, isolate held-out speakers, then freeze the scored sample manifest.

Traceability

results_raw.jsonl links every engineering attempt to captured media, metadata and event timestamps. methodology_and_provenance.json records corpus and design hashes. The full 30-direction matrix, package lock, test logs, independent ASR outputs, source approvals and reviewer forms accompany this report.

Primary sources

UN Transcription Corpus / UN Web TV licensing / UN Parallel Corpus and disclaimer / UNTERM / Exbabel Flagship Model supported languages / Google Live Translation / OpenSTBench

Written corpus attribution: United Nations. Ziemski, Junczys-Dowmunt and Pouliquen (2016), The United Nations Parallel Corpus. See the linked corpus disclaimer. This report is an Exbabel feasibility study.