EVALUATING SPEECH-TO-SPEECH TRANSLATION FOR DUBBING: CHALLENGES AND NEW METRICS

Fred Bane, Celia Soler Uguet, Llorenç Suau, João Torres, and Alan Vivares
TransPerfect AI

AGENDA

Translation of speech vs. text

Dubbing and Voice Over (V.O.)

New developments in speech translation

Existing evaluation methods

What *should* we be evaluating?

Pilot evaluation results

TEXT VS SPEECH TRANSLATION

Differences in the translation of text and speech.

WORKING WITH SPEECH VS. TEXT

Text	Speech
Discrete input	Continuous input
Singular signal	Mixed signals
Time-independent	Time-dependent
Linguistic evaluation	Linguistic + Voice evaluation
Less information overall	More information overall
Compact representation	Larger representation

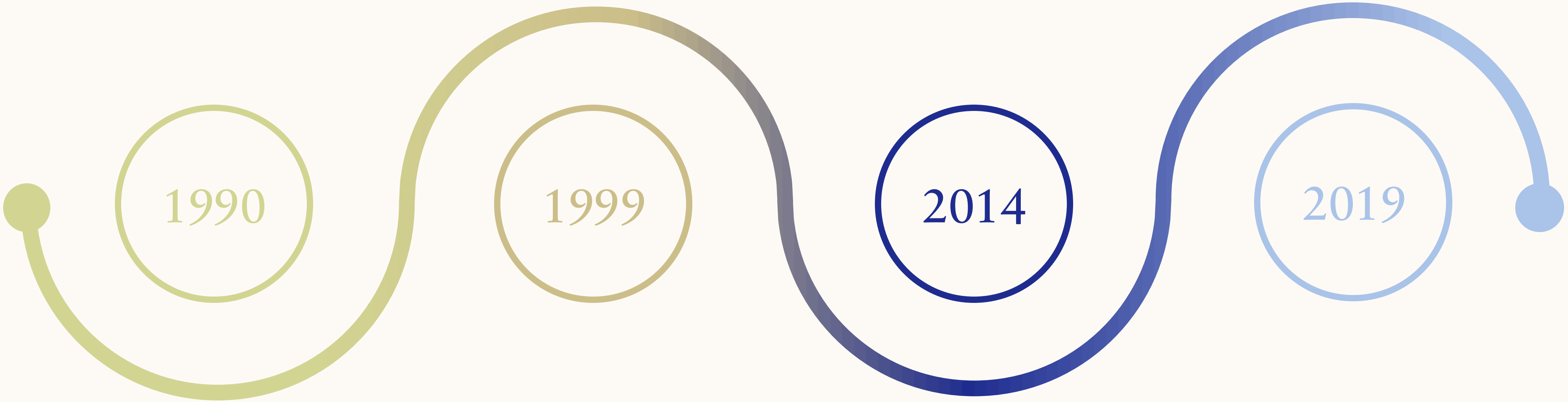
DUBBING AND VOICE-OVER

Key differences with other applications of speech translation

- **Timing:** Must fit in the same time span as the original
- **Synchronization:** In dubbing, synchronization of the voice with the lip movements is critical
- **Emotional expressivity:** In dubbing, matching the emotional content of the voice to the situation is critical
- **Fidelity:** Natural speech content that does not break immersivity is more important than maintaining fidelity
- **Character appropriateness:** The voice, speech content, and expressivity must be appropriate for the character

SPEECH-SPEECH TRANSLATION IS ENTERING A NEW ERA

THE DEVELOPMENT OF AI-DRIVEN TRANSLATION



EARLY DEVELOPMENT

1990s: The concept of machine translation (MT) began to gain traction, with early models focusing primarily on text-based translations.

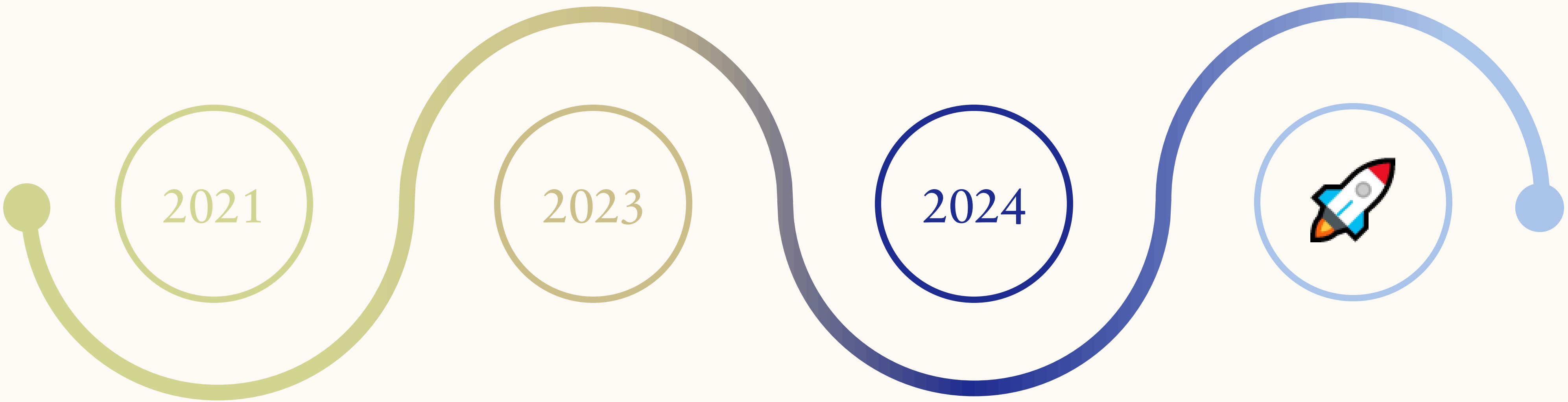
1999: Early S2S translation system introduced by the C-STAR-2 Consortium. By 2003, similar systems were developed for handheld devices.

THE RISE OF NEURAL NETWORKS

2014: Microsoft introduced (cascade-based) speech translation in Skype. Around the same time, Google launched Neural Machine Translation (NMT),

2019: Google introduced Translatotron, the first end-to-end model that directly translated speech from one language to another, bypassing text altogether.

THE DEVELOPMENT OF AI-DRIVEN TRANSLATION



RECENT ADVANCEMENTS

2021: Meta introduced SeamlessM4T, a multilingual and multimodal model capable of both text-to-text and speech-to-speech translation.

2023: Translatotron, Meta's SeamlessM4T and others continued to evolve, covering more languages, and improving emotional expressivity

FUTURE TRENDS

The future of AI in translation is expected to see further advancements in real-time translation capabilities across multiple modalities, particularly in enhancing the translation of low resource languages and incorporating non-verbal communication cues. LLMs have started to roll out voice capabilities, but audio is still separate from visual input.

EXISTING EVALUATION METHODS

TRANSLATION EVALUATION IS STILL TEXT-BASED

ASR-BLEU: Transcribing the speech using ASR and calculating BLEU, a text-based measure of similarity

- Dependent on the quality of the ASR system
- Not robust to dialectal variations or non-standardized orthographies
- Falls short in low-resource languages

BLEU has long been considered a poor metric for text translation, it is even less adequate for speech

VOICE QUALITY IS EVALUATED MANUALLY

Most major papers use Mean Opinion Score (MOS) as the only way of measuring voice quality.

The Seamless Expressive paper is a welcome exception: automated tools for sentence-level prosody similarity, and a rhythm evaluation toolkit

ALTERNATIVE EVALUATION METHODS

What *should* we be evaluating?

VOICE

- **Intelligibility**
- **Voice quality**
  - Articulation, fluency, projection
- **Appropriateness**
  - Suitability for character, cultural appropriateness
- **Expressivity**
  - Emotional content, consistency with context
- **Timing** (***task specific***)
  - Duration, lip synchronization

LINGUISTIC

- **Accuracy**
  - Mistranslation, over/under-translation, addition, omission, untranslated
- **Style**
  - Organizational, language register, consistency
- **Terminology**
  - Wrong term, consistency
- **Linguistic Conventions**
  - Grammar, word form, part of speech, tense, agreement, word order
- **Locale Conventions**

VOICE - QUALITY

❖ Articulation

Phoneme Error Rate (PER):

- Quantifies the accuracy of phoneme production by comparing expected vs. actual phonemes. This is useful for identifying pronunciation issues.

Mel cepstral distortion

Formant Analysis:

- Analyzes the resonant frequencies (formants) of the vocal tract, particularly crucial for vowel sounds. Deviations from expected formant values can indicate articulation issues.

❖ Projection

Similarity of amplitude envelope features (inspired by Cummings et al. 1999)

VOICE - QUALITY

❖ Fluency

Perplexity of vocal path through frequency-time space

- Transform the voice into frequency-time space, fit Bezier curves to the resulting path, calculate perplexity compared with a dataset of natural speech

Rhythmic analysis

- Speech rate (Librosa, AutoPCP), pauses (Praat, pydub, Rhtyhm Toolkit)

F0 contour and amplitude envelope (Cummins et al., 1999)

VOICE - INTELLIGIBILITY

Perplexity of audio -> phoneme decoder

VOICE - APPROPRIATENESS

Mel frequency cepstral coefficient similarity

Cosine distance embedding vectors (x-vectors, PnG NAT TTS model in Nobuyuki et al., 2022)

Automated MOS prediction (MOSnet in Lo et al. 2019)

Classifier trained to predict if the voice is the same

VOICE - EXPRESSIVITY

Prosody similarity (AutoPCP)

Emotion detection systems

LINGUISTIC – ACCURACY

Encoder embedding similarity (BLASER - Bilingual and Language-Agnostic Speech Evaluation by Retrieval)

Round-trip phoneme F1

- Back-translating the output translation into the source language and comparing the two audios represented as sequences of phonemes

Round-trip BLASER

- BLASER may capture semantic features, in a way that COMET can augment chrF1/BLEU scores for text translation

LINGUISTIC – STYLE

???

LINGUISTIC – TERMINOLOGY

???????

LINGUISTIC/LOCALE CONVENTIONS

???????????

There is still a long way to go ☺

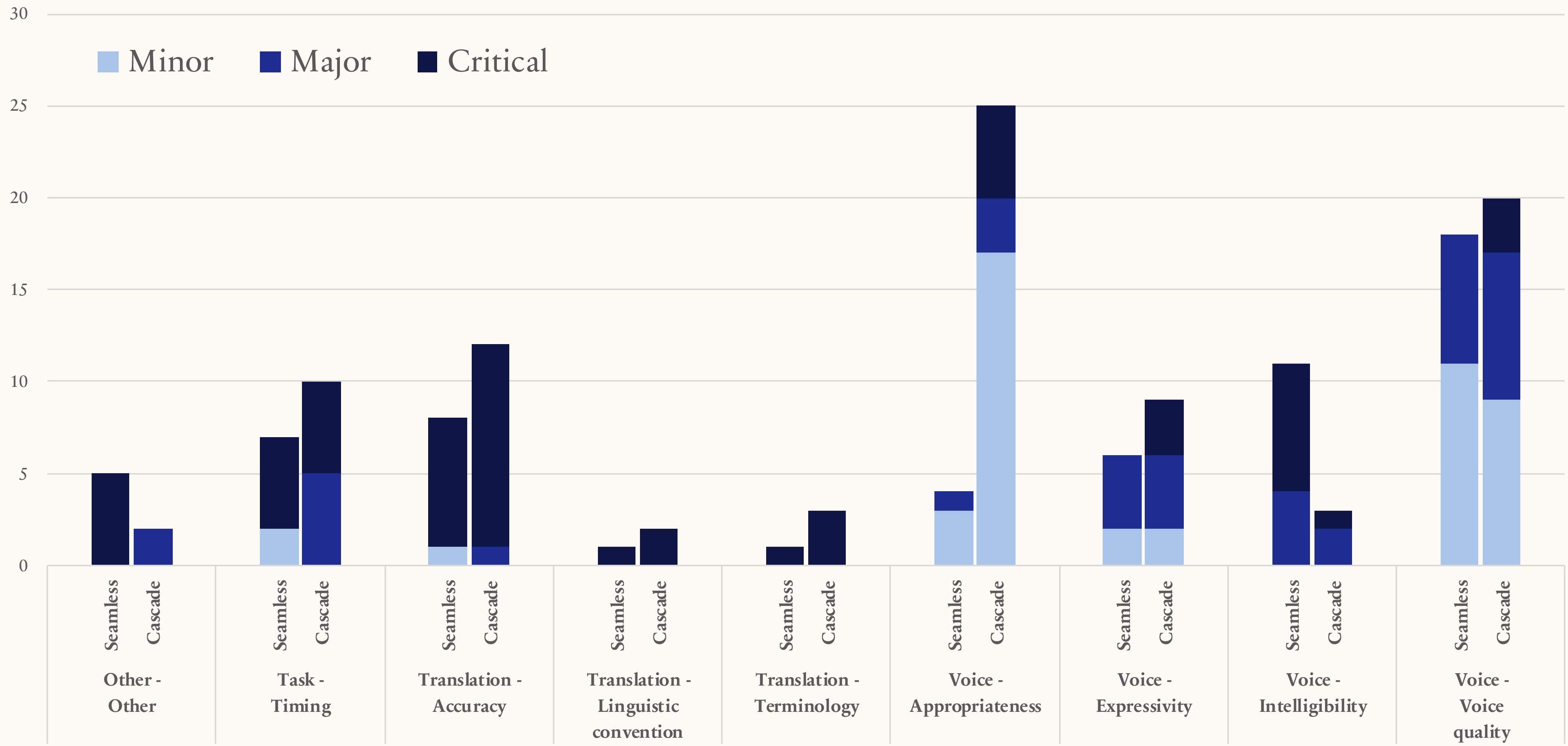
PILOT EVALUATION RESULTS

Results from a small-scale pilot, reviewed manually with the error taxonomy shown previously

PILOT SETUP

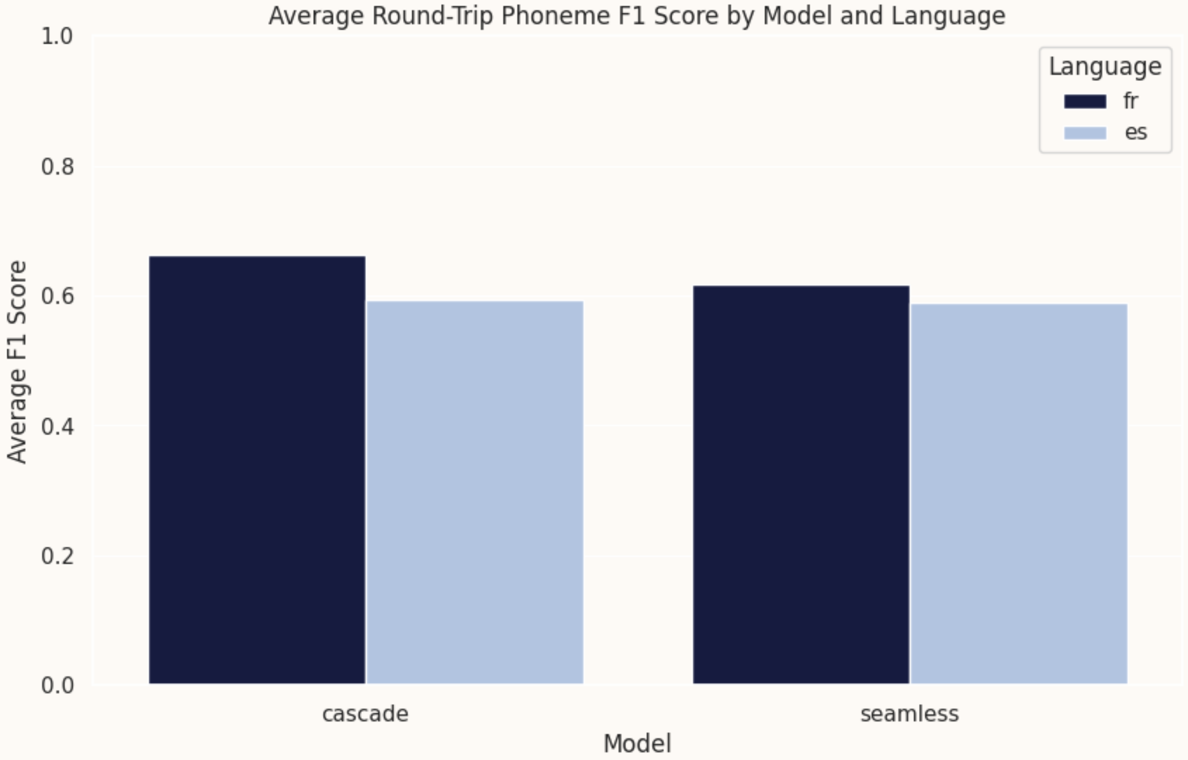
- Short clips from movies, web series, and documentaries, showcasing a variety of expressive conditions;
- We first separated speech signals from background noise in the audio track;
- Then we translated each vocal track into FR and ES using Seamless Expressive and a cascade approach (whisper ➔ internally trained MT models ➔ internally developed TTS models);
- Next, translated audio was reinserted into the background noise at the corresponding time using the time codes of the speech signals;
- Reviewers worked on the DataForce platform to annotate errors, indicating the type, severity, start time, and end time of each error.

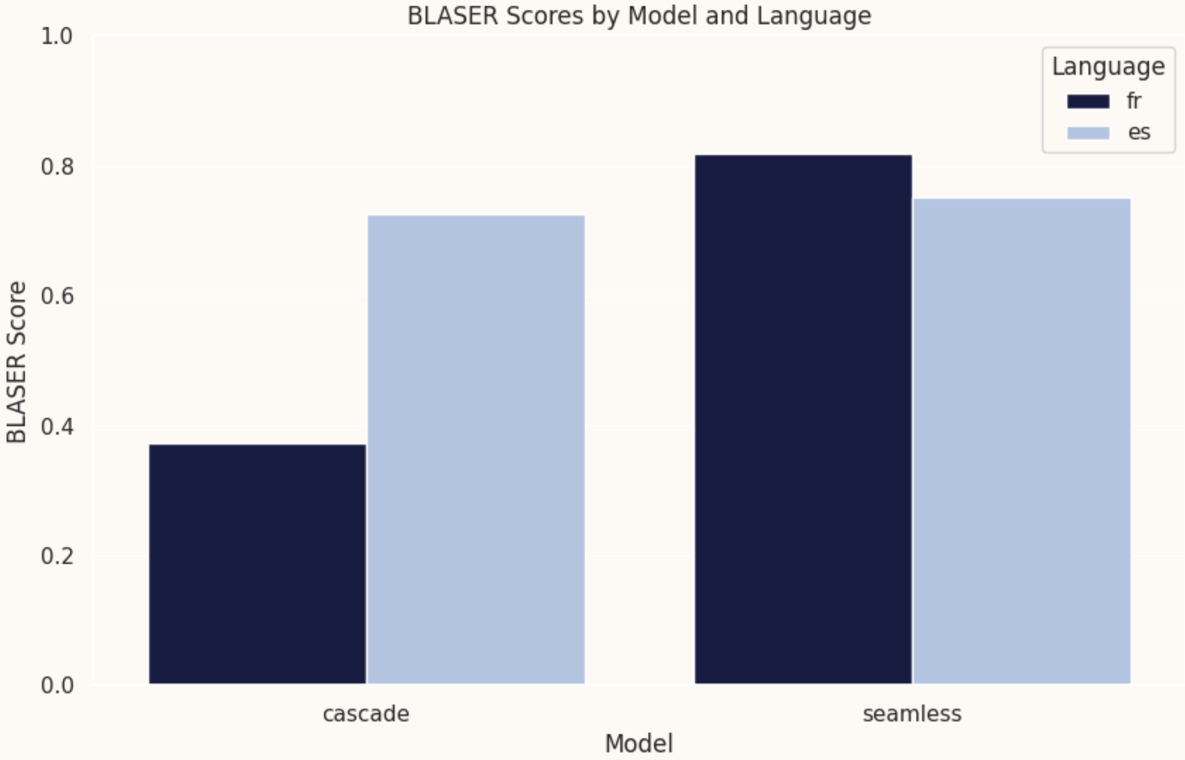
ERRORS BY MODEL, TYPE, SEVERITY



PILOT RESULTS VS BLASER

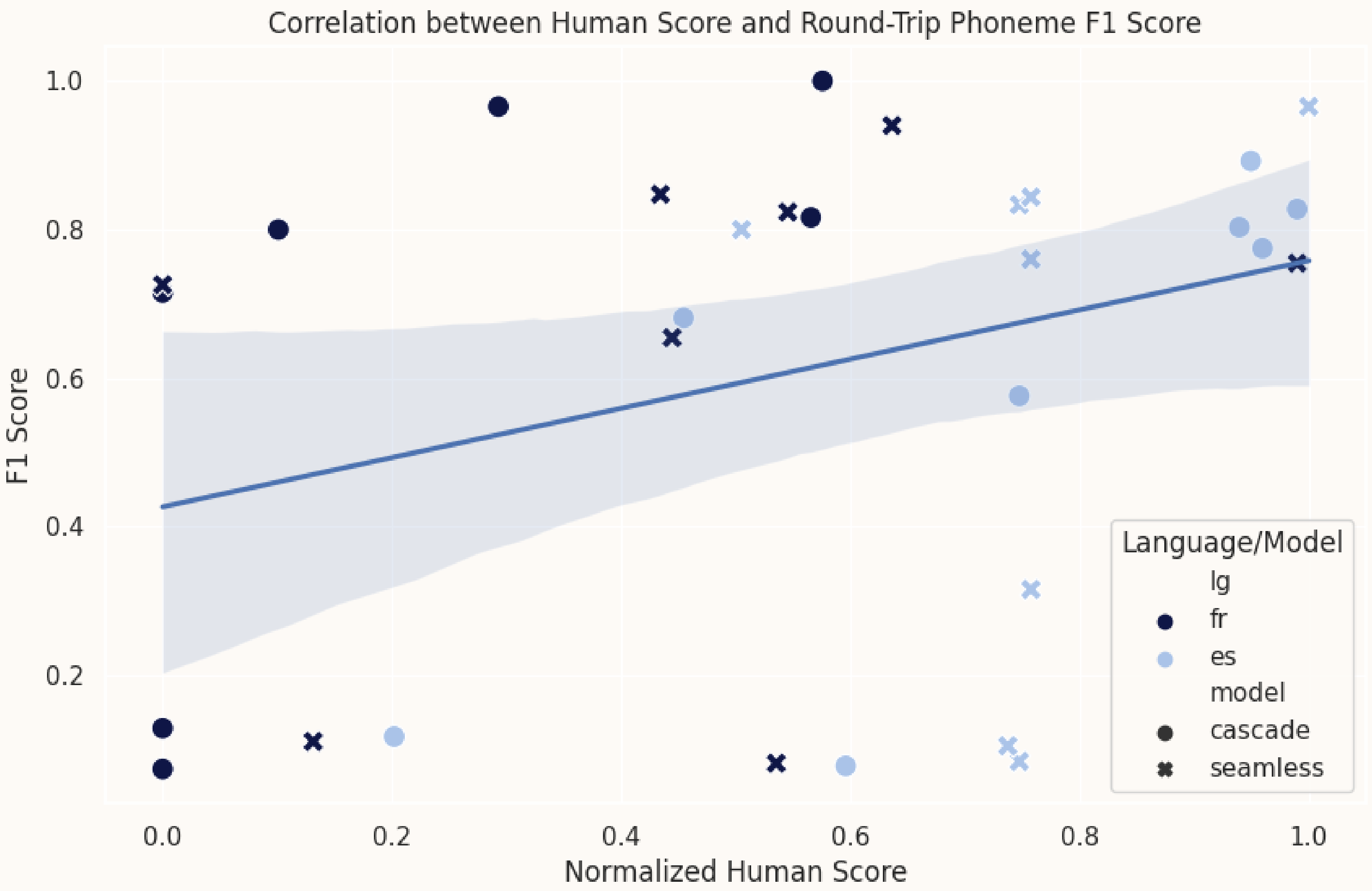
- ❖ Round-Trip Phoneme F1 scores exhibit a similar trend to BLASER. However, French translations using the Cascade model received much lower BLASER scores, possibly due to differences in vocal rather than linguistic characteristics of the translations





CORRELATION BETWEEN HUMAN EVALUATION AND PILOT SCORES

- We normalized the Human Evaluation Scores to a scale between 0 and 1, with 1 representing a perfect, error-free translation. This normalization allowed us to benchmark the Round-Trip Phoneme F1 Score against human judgment;
- Although the positive slope indicates that higher human scores generally align with better F1 scores, the correlation is not statistically significant.



THANK YOU

Fred Bane
fbane@translations.com
github.com/TransperfectAI/amta2024_S2SEvaluation

TRANSPERFECT