Beyond BLEU: Linguistic Invisibility and Interactional Repair Sequence in End-to-End Sign Language Translation

Zirui Wang and **Mayumi Bono**

SOKENDAI(The Graduate University for Advanced Studies), National Institute of Informatics, Japan

Overview

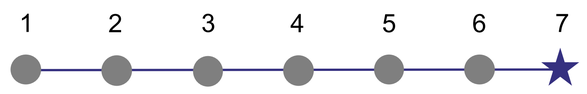
Background

- Recent end-to-end sign language translation (SLT) systems have achieved strong benchmark performance on standard datasets.
- However, benchmark success does not necessarily indicate whether models preserve the linguistic structures through which meaning is negotiated in real signed interaction.
- Many current systems are optimized mainly for sequence-level output similarity.

Problem： Current evaluation metrics such as BLEU score may reward correct-looking outputs while overlooking interactionally decisive signals.

Approach： We investigate this issue through a diagnostic repair sequence from a **Japanese Sign Language (JSL) conversational corpus**.
By combining interactional analysis with kinematic measurements, we examine whether crucial multi-channel signals are likely to remain visible under current end-to-end modeling paradigms.

Real interaction from a Japanese Sign Language (JSL) corpus



Trial 7:
Manual form: returns to trial 1
Mouthing: ma-n-ga → a-ni-me

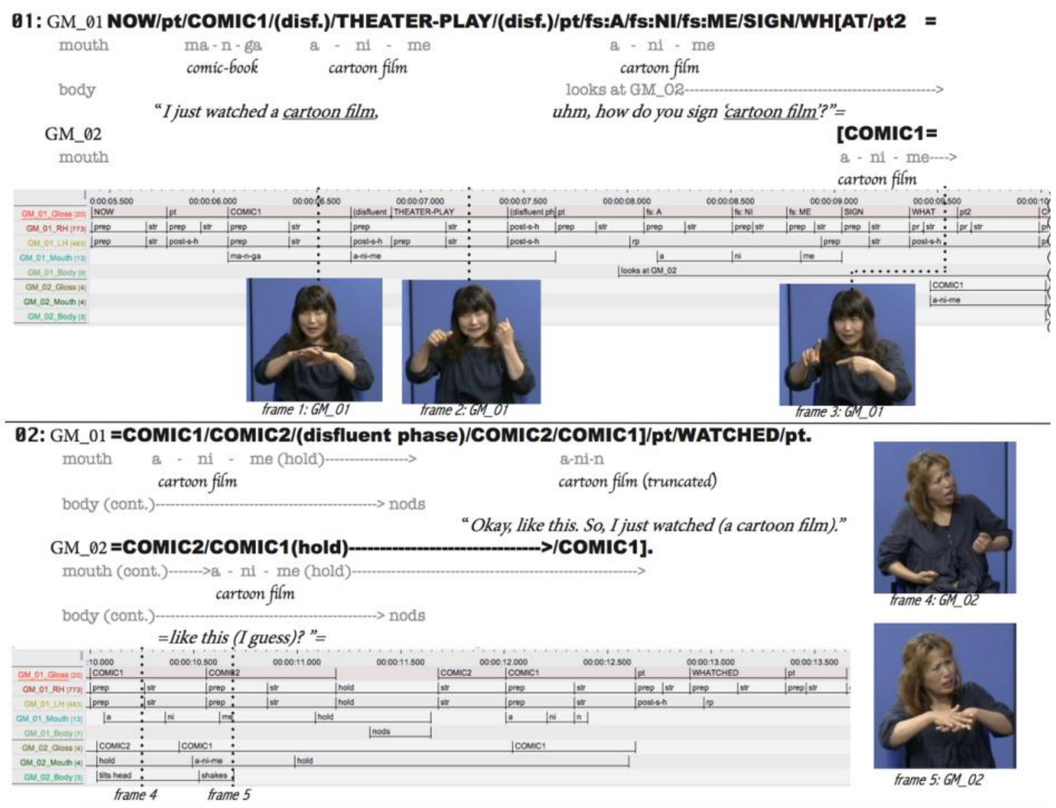
Repair Sequence over 7 Trials
The repair sequence was resolved in Trial 7.

Contribution

- Introduce the concept of linguistic invisibility in SLT
- Show **manual–mouth decoupling** through motion analysis
- Present a real JSL repair sequence as a diagnostic probe
- Argue for evaluation beyond BLEU score alone

Diagnostic Repair Sequence Analysis

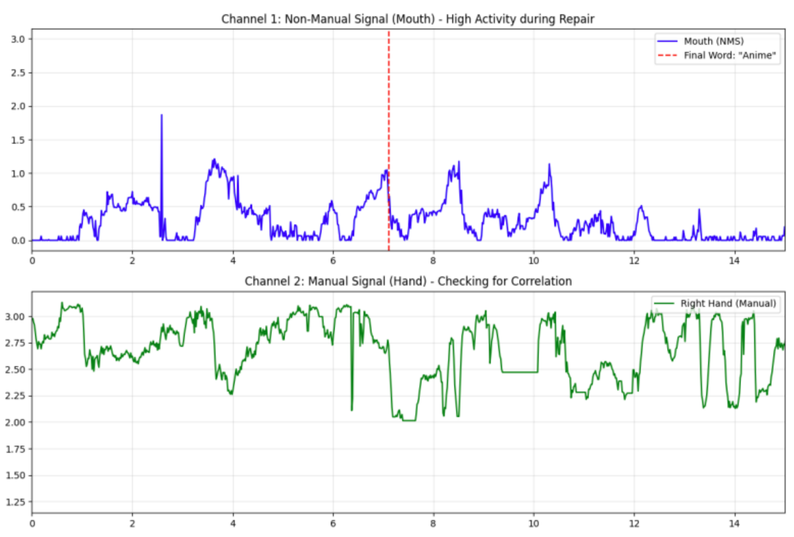
What Changes in Trial 7?



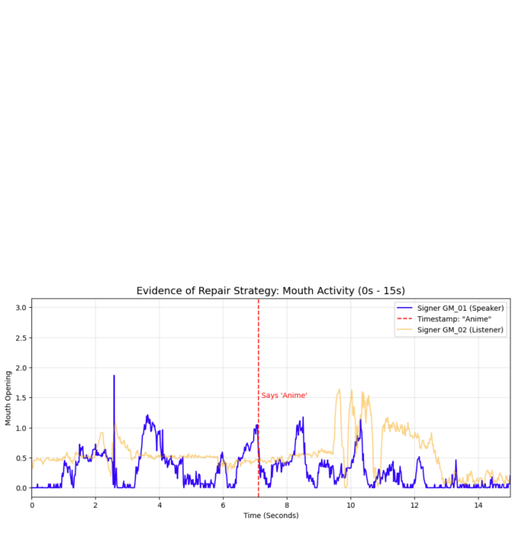
"Repair Sequence" from a Japanese Sign Language (JSL) corpus

- In the final trial, the signer returns to a previously used manual form while altering the mouthing from ma-n-ga to a-ni-me.
- This indicates that successful repair is achieved through subtle cross-channel contrast rather than a large manual change.

Kinematic Evidence of Manual–Mouth Decoupling



Kinematic decoupling between manual and non-manual signals



Interactional coordination and signal validation

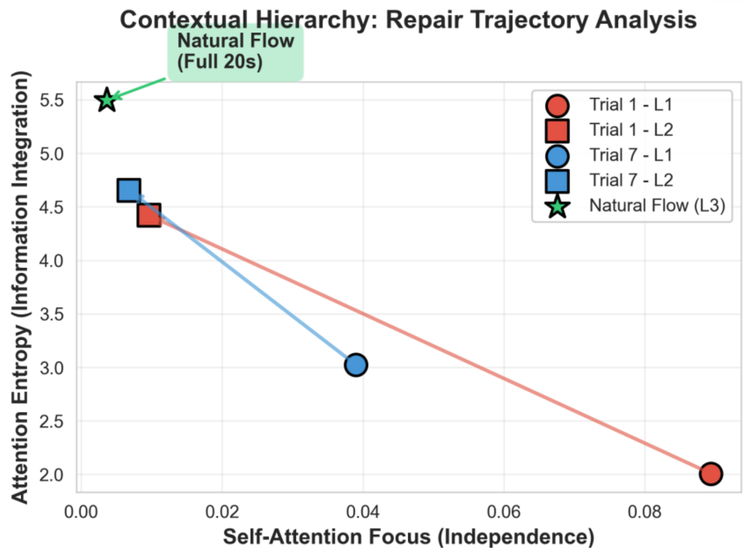
- Hand movement remains broadly active, while mouth activity peaks near the final trial.
- A closely aligned change in the co-participant's activity follows the final mouth peak.
- Successful repair may depend on subtle cross-channel contrasts beyond global motion similarity.

Why This Matters for Current SLT

- Large hand motion may dominate visual encoding
- Subtle mouth contrasts may be attenuated
- Final text accuracy may miss communicative mechanisms

Implications for Future SLT

Additional Computational Evidence



H–F trajectory across contextual levels

Successful repair appears closer to discourse-level context.

Why Current Metrics Are Insufficient

- High benchmark scores mainly reflect output similarity.
- They may not reveal whether models preserve interactionally decisive contrasts.

What Future Systems Need

- Multi-channel evaluation
- Repair-sequence benchmarks
- Sensitivity to non-manual contrasts
- Metrics beyond BLEU score alone

Our Furure Work：

- We plan to extend the analysis beyond repair sequences to smooth narrative phases.
- We also aim to test the framework on other datasets and task-based interactions (e.g., a cooking-based corpus).
- We will explore SLT models that better capture interactionally relevant NMS signals.

Final Takeaway: Translation quality should include structural visibility.