NOVARA
CONSULTING GROUP

Consistently Wrong: Apple DiscoSign and Sign Language AI

Heather Grizzle · September 11, 2026

Apple and Gallaudet researchers built sign language AI that remembers. When a system keeps state, one early error stops being noisy and becomes systematic.

Contents

A paper posted to arXiv on 2 September 2026 and accepted to the main conference at EMNLP describes DiscoSign, a framework for translating English text into American Sign Language gloss while keeping track of what has already been said. Its authors are researchers at Apple and at Gallaudet University, and aggregators reported it appearing on Apple′s machine learning research site on 11 September, which is where most readers outside the field will have encountered it. The technical claim is narrow and well stated. Most text-to-sign systems translate one sentence at a time and have no memory between sentences, so a referent placed in one location can be silently reassigned in the next, the same concept can be rendered by three different signs in three consecutive sentences, and the rhetorical structures that organise signed discourse cannot be chosen because the system cannot see the discourse. DiscoSign adds explicit registries that persist across sentences, enforces them as constraints on the model, and then verifies and corrects the output against them.

This is a real advance on a real defect, and the paper is candid about what it does not do. The governance significance lies one step further on. When a system carries state across a document, the unit that can fail is no longer the sentence. Errors stop being independent events and start being properties of the whole passage, and the evaluation practices the field has built, which score sentences and average them, cannot see that. The risk is not that the research overstates itself. It is that the words "discourse-aware" will arrive in procurement documents long before anyone has agreed how to test whether discourse is actually being handled.

What the paper claims, and what it says about itself

DiscoSign addresses three phenomena. The first is spatial coreference: in ASL, entities are assigned locations in the signing space and later referred to by pointing back to them, so a system must remember that Alice was placed to the left and Bob to the right. The second is Question-Answer Clauses, the topic-comment structures in which a signer poses a rhetorical question and immediately answers it, which are appropriate in some discourse positions and awkward in others. The third is concept-gloss consistency, meaning that a concept appearing five times in a document should be rendered by the same sign each time rather than alternating between near-synonyms. Each is handled by a module that maintains a registry, states it to the language model as a hard constraint, and then checks the returned output and corrects violations in place.

The paper reports that this substantially improves spatial consistency and entity tracking against sentence-level baselines while holding single-sentence quality steady, and it introduces its own metrics because, as it says, standard machine translation metrics "fail to capture discourse-level quality." Its limitations section is unusually direct. The framework "addresses manual sign production through glosses but does not incorporate non-manual markers (NMMs) such as facial expressions and prosody, which carry critical grammatical and affective information in sign languages." One of its three datasets "lacks ground truth ASL annotations," so parts of the discourse-level evaluation rest on automatic metrics and back-translation. The rule that decides when a Question-Answer Clause is appropriate "cannot represent the optionality" of a structure that is sometimes a matter of style. The authors also state that they "collaborated with members of the signing community" and include Gallaudet co-authors, which is not the norm in this literature and should be said plainly.

Consistency is not accuracy

The mechanism that makes DiscoSign work is also what makes it a governance problem, and this is the part that does not follow from reading the abstract. A registry enforces that a decision made early is applied throughout. If the decision is right, the whole passage is coherent. If the decision is wrong, the whole passage is wrong the same way. A referent assigned to the wrong location in the second sentence does not produce one bad sentence; it produces a document in which every later reference points confidently to the wrong person. A concept mapped to an inexact sign is not an isolated slip; it is that sign, repeated, with the system's consistency machinery actively preventing the correction that variation might otherwise have introduced. The paper's own description of the pipeline says that "discourse-level errors introduced at the gloss stage propagate through the entire pipeline."

Sentence-level systems fail noisily, and noisy failure has an underrated property: a reader notices it. Inconsistency is legible as a defect. A system that is wrong consistently looks like a system that means it. For a Deaf reader who is receiving the output rather than checking it against the source, a stable wrong referent is not obviously an error at all; it is simply what the document says. This inverts the usual assumption that better internal coherence is safer. Coherence raises the cost of a mistake at the moment it lowers the visibility of one, and no average of per-sentence scores will show it.

What the new metrics measure, and what validated them

The paper is careful here, and the care is worth reading closely because it sets the ceiling on what can responsibly be claimed downstream. The new metrics measure spatial index consistency, concept-gloss mapping stability, and the appropriateness of Question-Answer Clause use. Two of the three measure whether the system did the same thing throughout, which is a property of the output considered on its own terms and not a measure of whether a Deaf reader understood the passage. To validate the metrics, two ASL-fluent evaluators rated thirty-two stories, one hundred and fifty-two sentences, on a five-point scale. Spatial consistency correlated moderately with their judgements. Concept-gloss consistency correlated. The appropriateness metric for Question-Answer Clauses did not reach significance, and the paper attributes this to the evaluators disagreeing about what counts as such a clause at all, with weighted agreement between them described as moderate.

For a research contribution validating a new metric, two expert raters is a reasonable design and the honesty about the negative result is creditable. As a basis for institutional deployment it is nowhere near sufficient, and the paper does not claim otherwise. Three distinctions matter if this work travels into procurement. First, correlating an automatic metric with expert ratings of the same property is not the same as measuring whether Deaf readers comprehend the document, which nobody has yet done at discourse level. Second, "ASL-fluent" is not a synonym for Deaf, and the paper does not report the evaluators' deafness, certification or background. Third, when qualified raters disagree about whether a grammatical structure is even present, an automated score asserting that its use was appropriate is not yet an assurance artifact, whatever number it produces.

Gloss is still not the language

The coverage framing this as sign language AI moving beyond sentence-by-sentence translation is reaching past what was built. DiscoSign produces gloss, a written label sequence standing in for manual signs, and the authors say directly that non-manual markers are not included. In ASL, a great deal of the discourse structure the paper is trying to preserve lives precisely in those channels. Topic marking, the boundary between the question and answer halves of the very structures the framework models, the role shifts that signal whose perspective is being voiced, the eye gaze that ties a reference back to a location in space: these are carried by the face, head and body. A gloss string can record that a referent is index-j; it cannot record the gaze that makes the reference land.

The authors see this and say something useful about it, which is that the state their registries hold is close to the information a downstream visual system would need in order to produce non-manual markers, and that deriving those annotations from the registry is a natural next step. That is a sober description of unfinished work. It is not the same as a system that produces coherent signed discourse, and the distance between the two is where a procurement officer reading a press summary will be misled. A buyer told that a vendor's pipeline is discourse-aware is entitled to ask which stage that describes, and whether anything downstream of the gloss preserves it.

The dataset that is not named

The paper lists three data sources: ASL STEM Wiki, Aesop′s Fables from Project Gutenberg, and, in its own words, "a licensed dataset." The first two are identified and checkable. The third is not named. This is ordinary practice and very likely reflects a commercial licence rather than anything untoward, but it is worth registering as a standing question rather than passing over it, because in signed language work the provenance of a corpus is a question about people. Signed data is recorded from signers whose faces and bodies are the data. Whether those signers were consented for machine translation research, whether they were compensated, and what the licence permits downstream are not answerable when the source is unnamed. A field that increasingly turns on the quality and provenance of its corpora will eventually be asked these questions by regulators and by Deaf organisations, and naming the source is the cheapest possible way to be ready for that.

What changes for evaluation when the system remembers

If discourse-aware systems become the norm, and the direction of travel suggests they will, then anyone evaluating them needs instruments that operate at the level the system now operates at. Six requirements follow from what this paper shows. The first is persistence under length: consistency measured across three sentences says nothing about a system holding a dozen referents across a page of prose, so tests must state document length and referent density rather than reporting an average. The second is propagation: an evaluation should report how far a single early error travels, measured as affected downstream sentences, not diluted across a document as a fractional score. The third is worst case alongside mean, because an institution′s exposure is determined by the passage that failed, not the average passage.

The fourth is recovery: a system that maintains state should be tested on whether it can revise an assignment when later text contradicts an earlier inference, and on what happens when a document legitimately reassigns a referent. The fifth is human correction at the right level, since a reviewer checking sentence by sentence will approve a passage whose error is a property of the whole, which means review procedures written for sentence-level output do not transfer. The sixth is comprehension by Deaf readers, measured on whole documents against a human signed comparator, which remains the only evidence that any of the internal consistency being measured actually reaches the person it is for. NCG′s position, which this paper does not contradict, is

that internal metrics are an input to assurance and never a substitute for validated output, and the evidence required rises with the stakes of the setting in which the output is used.

The limits of this analysis

This reading rests on the arXiv version of the paper as posted, together with public coverage of its appearance on Apple's research site, which was not independently verified. DiscoSign is research, not a product: nothing here describes a deployed system, no vendor is making claims about it, and no NCG instrument has been applied to it. The criticisms above are not criticisms of the authors, who state most of these limitations themselves and did the field a service by publishing metrics that make discourse failure measurable at all. The argument is about what happens next, when a genuine research advance is compressed into a capability claim by people who did not read the limitations section.

The useful thing about this paper is that it makes a hidden problem visible. Sentence-level systems were failing at discourse all along; there was simply no instrument that could see it. Now there is the beginning of one, and it arrives with a finding that should be read as a warning as much as a result. Making a system consistent does not make it correct. It makes it correct or incorrect all the way through, and decides, before anyone reads a word of the output, which of those it will be.

Sources

1. Vasileios Baltatzis, Mert Inan, Connor Gillis, Raja Kushalnagar, Lorna Quandt, Leah Findlater and Colin Lea, "DiscoSign: Discourse-Aware Text to Sign Language Gloss Translation," arXiv:2609.02796, 2 September 2026, accepted at EMNLP 2026 Main Conference. https://arxiv.org/abs/2609.02796
2. DiscoSign, full text (HTML). https://arxiv.org/html/2609.02796
3. AI Global Wire, report of the paper's appearance on Apple Machine Learning Research, 11 September 2026. https://aiglobalwire.com/article/discosign-discourse-aware-text-to-sign-language-gloss-translation-dbf89917-bb51-4793-be6d-29ed6727d3e0

How to cite

Grizzle, H. M. (2026, September 11). Consistently Wrong: Apple DiscoSign and Sign Language AI. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003304