NOVARA
CONSULTING GROUP

This document was translated from English by machine and may contain mistakes. The English original is at https://www.novaracg.com/2026/09/11/consistently-wrong-apple-discosign-and-sign-language-ai/.

Konsequent falsch: Apple DiscoSign und Gebärdensprach-KI

Heather Grizzle · September 11, 2026

Forscher von Apple und Gallaudet haben eine Gebärdensprach-KI entwickelt, die sich erinnert. Wenn ein System Zustände beibehält, hört ein früher Fehler auf, zufällig zu sein, und wird systematisch.

Inhalt

1. Was das Papier behauptet und was es über sich selbst sagt
2. Konsistenz ist keine Genauigkeit
3. Was die neuen Metriken messen und was sie validiert hat
4. Gloss ist immer noch nicht die Sprache
5. Der Datensatz, der nicht genannt wird
6. Was sich für die Bewertung ändert, wenn das System sich erinnert
7. Die Grenzen dieser Analyse
8. Quellen

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 expli-

cit 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.

Was das Papier behauptet und was es über sich selbst sagt

DiscoSign befasst sich mit drei Phänomenen. Das erste ist die räumliche Koreferenz: In ASL werden Entitäten Orte im Gebärdenraum zugewiesen und später durch Zurückzeigen auf diese referenziert, sodass ein System sich merken muss, dass Alice links und Bob rechts platziert wurde. Das zweite sind Question-Answer Clauses, die Topic-Comment-Strukturen, bei denen ein Gebärdender eine rhetorische Frage stellt und sie sofort beantwortet, was an manchen Stellen im Diskurs angemessen und an anderen unpassend ist. Das dritte ist die Konzept-Gloss-Konsistenz, was bedeutet, dass ein Konzept, das fünfmal in einem Dokument vorkommt, jedes Mal durch dieselbe Gebärde wiedergegeben werden sollte, anstatt zwischen nahezu synonymen Varianten zu wechseln. Jedes wird von einem Modul behandelt, das ein Register führt, es dem Sprachmodell als harte Beschränkung vorgibt und dann die zurückgegebene Ausgabe prüft und Verstöße direkt korrigiert.

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 struc-

ture 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.

Konsistenz ist keine Genauigkeit

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."

Satzbasierte Systeme versagen laut, und lautes Versagen hat eine unterschätzte Eigenschaft: Ein Leser bemerkt es. Inkonsistenz ist als Fehler erkennbar. Ein System, das konsistent falsch liegt, sieht aus wie ein System, das es so meint. Für einen Deaf-Leser, der die Ausgabe empfängt, ohne sie mit der Quelle abzugleichen, ist ein stabil falscher Referent überhaupt nicht offensichtlich ein Fehler; es ist einfach das, was das Dokument aussagt. Dies kehrt die übliche Annahme um, dass bessere interne Kohärenz sicherer sei. Kohärenz erhöht die Kosten eines Fehlers genau in dem Moment, in dem sie dessen Sichtbarkeit verringert, und kein Durchschnitt von Satzbewertungen wird das zeigen.

Was die neuen Metriken messen und was sie validiert hat

Das Papier ist hier sorgfältig, und die Sorgfalt lohnt eine genaue Lektüre, weil sie die Obergrenze dessen setzt, was verantwortungsvoll nachgelagert behauptet werden kann. Die neuen Metriken messen die Konsistenz räumlicher Indizes, die Stabilität der Konzept-Gloss-Zuordnung und die Angemessenheit der Verwendung von Question-Answer Clauses. Zwei der drei messen, ob das System durchgängig dasselbe getan hat, was eine Eigenschaft der Ausgabe für sich genommen ist und kein Maß dafür, ob ein Deaf-Leser die Passage verstanden hat. Zur Validierung der Metriken bewerteten

zwei ASL-fließende Gutachter zweiunddreißig Geschichten mit hundertzweiundfünfzig Sätzen auf einer fünfstufigen Skala. Die räumliche Konsistenz korrelierte mäßig mit ihren Urteilen. Die Konzept-Gloss-Konsistenz korrelierte. Die Angemessenheitsmetrik für Question-Answer Clauses erreichte keine Signifikanz, und das Papier führt dies darauf zurück, dass die Gutachter uneinig waren, was überhaupt als solche Klausel zählt, wobei die gewichtete Übereinstimmung zwischen ihnen als mäßig beschrieben wird.

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 ist immer noch nicht die Sprache

Die Berichterstattung, die dies als Gebärdensprach-KI darstellt, die über die satzweise Übersetzung hinausgeht, greift weiter, als tatsächlich gebaut wurde. DiscoSign erzeugt Gloss, eine schriftliche Bezeichnungsfolge, die für manuelle Gebärden steht, und die Autoren sagen direkt, dass nicht-manuelle Marker nicht einbezogen sind. In ASL liegt ein großer Teil der Diskursstruktur, die das Papier zu bewahren versucht, genau in diesen Kanälen. Themenmarkierung, die Grenze zwischen den Frage- und Antworthälften genau der Strukturen, die das Framework modelliert, die Perspektivwechsel, die anzeigen, aus wessen Sicht gesprochen wird, der Blickkontakt, der einen Verweis mit einem Ort im Raum verknüpft: Diese werden durch Gesicht, Kopf und Körper getragen. Eine Gloss-Zeichenkette kann festhalten, dass ein Referent Index-j ist; sie kann nicht den Blick festhalten, der den Verweis verankert.

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 re-

gistry 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.

Der Datensatz, der nicht genannt wird

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.

Was sich für die Bewertung ändert, wenn das System sich erinnert

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.

Die Grenzen dieser Analyse

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.

Das Nützliche an diesem Papier ist, dass es ein verborgenes Problem sichtbar macht. Satzbasierte Systeme versagten die ganze Zeit beim Diskurs; es gab schlicht kein Instrument, das dies erkennen konnte. Jetzt gibt es den Anfang eines solchen, und er kommt mit einem Befund, der ebenso als Warnung wie als Ergebnis gelesen werden sollte. Ein System konsistent zu machen, macht es nicht korrekt. Es macht es durchgängig korrekt oder durchgängig falsch und entscheidet, bevor jemand ein Wort der Ausgabe liest, welches von beidem es sein wird.

Quellen

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

TIPS

Know something about a sign language AI system that the public should know? Send it to tips@novaracg.com. Every tip is read by me and kept confidential. If you work in the UK, writing to us is not by itself a protected disclosure under the Public Interest Disclosure Act 1998, so before you share employer information, check your options with Protect (protect-advice.org.uk), the free whistleblowing advice line.

Zitierweise

Grizzle, H. M. (2026, September 11). Konsequent falsch: Apple DiscoSign und Gebärdensprach-KI. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003304