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/.

Costantemente sbagliato: Apple DiscoSign e l'IA per la lingua dei segni

Heather Grizzle · September 11, 2026

Ricercatori di Apple e Gallaudet hanno costruito un'IA per la lingua dei segni dotata di memoria. Quando un sistema mantiene uno stato, un errore precoce smette di essere rumoroso e diventa sistematico.

Contenuti

1. Cosa afferma il paper e cosa dichiara di sé stesso
2. La coerenza non è accuratezza
3. Cosa misurano le nuove metriche e cosa le ha validate
4. Il gloss non è ancora la lingua
5. Il dataset che non viene nominato
6. Cosa cambia nella valutazione quando il sistema ricorda
7. I limiti di questa analisi
8. Fonti

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.

Cosa afferma il paper e cosa dichiara di sé stesso

DiscoSign affronta tre fenomeni. Il primo è la coreferenza spaziale: in ASL, alle entità vengono assegnate posizioni nello spazio segnico e in seguito ci si riferisce a esse indicando quelle posizioni, quindi un sistema deve ricordare che Alice è stata collocata a sinistra e Bob a destra. Il secondo sono le Question-Answer Clauses, le strutture topic-comment in cui chi segna pone una domanda retorica e vi risponde immediatamente, appropriate in alcune posizioni discorsive e goffe in altre. Il terzo è la coerenza concetto-gloss, ossia il fatto che un concetto che compare cinque volte in un documento dovrebbe essere reso con lo stesso segno ogni volta, invece di alternare tra quasi-sinonimi. Ciascuno di questi aspetti è gestito da un modulo che mantiene un registro, lo impone al modello linguistico come vincolo rigido, e poi verifica l'output restituito correggendo sul posto le violazioni.

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.

La coerenza non è accuratezza

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."

I sistemi a livello di frase falliscono in modo rumoroso, e il fallimento rumoroso ha una proprietà sottovalutata: il lettore lo nota. L'incoerenza è leggibile come difetto. Un sistema che sbaglia in modo coerente sembra un sistema che intende ciò che dice. Per un lettore Sordo che riceve l'output senza confrontarlo con la fonte, un referente stabilmente sbagliato non è affatto un errore evidente; è semplicemente ciò che il documento dice. Questo capovolge il presupposto abituale secondo cui una migliore coerenza interna sia più sicura. La coerenza innalza il costo di un errore proprio nel momento in cui ne riduce la visibilità, e nessuna media di punteggi per frase potrà mostrarlo.

Cosa misurano le nuove metriche e cosa le ha validate

Il paper è attento in questo punto, e l'attenzione merita una lettura ravvicinata perché stabilisce il limite massimo di ciò che si può responsabilmente affermare a valle. Le nuove metriche misurano la coerenza degli indici spaziali, la stabilità della mappatura concetto-gloss e l'appropriatezza dell'uso delle Question-Answer Clauses. Due delle tre misurano se il sistema abbia fatto la stessa cosa per tutto il testo, il che è una proprietà dell'output considerato di per sé e non una misura di se un lettore Sordo abbia compreso il passaggio. Per validare le metriche, due valutatori fluenti in ASL hanno valutato trentadue storie, centocinquantadue frasi, su una scala a cinque punti. La

coerenza spaziale ha correlato moderatamente con i loro giudizi. La coerenza concetto-gloss ha correlato. La metrica di appropriatezza per le Question-Answer Clauses non ha raggiunto la significatività, e il paper lo attribuisce al disaccordo tra i valutatori su cosa conti come tale clausola, con un accordo ponderato tra loro descritto come moderato.

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.

Il gloss non è ancora la lingua

La copertura mediatica che presenta questo come l'IA per la lingua dei segni che va oltre la traduzione frase per frase va oltre ciò che è stato effettivamente costruito. DiscoSign produce gloss, una sequenza di etichette scritte che sta al posto dei segni manuali, e gli autori affermano direttamente che i marcatori non manuali non sono inclusi. In ASL, gran parte della struttura discorsiva che il paper cerca di preservare risiede proprio in quei canali. La marcatura del topic, il confine tra le metà domanda e risposta proprio delle strutture che il framework modella, i cambi di ruolo che segnalano di chi sia il punto di vista espresso, lo sguardo che collega un riferimento a una posizione nello spazio: tutto questo è veicolato dal viso, dalla testa e dal corpo. Una stringa di gloss può registrare che un referente è index-j; non può registrare lo sguardo che fa atterrare il riferimento.

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 bet-



ween 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.

Il dataset che non viene nominato

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.

Cosa cambia nella valutazione quando il sistema ricorda

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 writ-

ten 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.

I limiti di questa analisi

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.

La cosa utile di questo paper è che rende visibile un problema nascosto. I sistemi a livello di frase fallivano nel discorso da sempre; semplicemente non esisteva uno strumento in grado di vederlo. Ora ne esiste l'inizio, e arriva con una scoperta che va letta tanto come un avvertimento quanto come un risultato. Rendere un sistema coerente non lo rende corretto. Lo rende corretto o scorretto fino in fondo, e decide, prima ancora che qualcuno legga una sola parola dell'output, quale delle due cose sarà.

Fonti

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-29ed6727-d3e0

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.

Come citare

Grizzle, H. M. (2026, September 11). Costantemente sbagliato: Apple DiscoSign e l'IA per la lingua dei segni. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003304