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/.

# Sistemáticamente equivocado: Apple DiscoSign y la IA de lengua de signos

Heather Grizzle · September 11, 2026

*Investigadores de Apple y Gallaudet crearon una IA de lengua de señas con memoria. Cuando un sistema mantiene un estado, un error inicial deja de ser ruido puntual y se convierte en algo sistemático.*

**Contenidos**

1. Lo que afirma el artículo, y lo que dice de sí mismo
2. La consistencia no es exactitud
3. Qué miden las nuevas métricas, y qué las validó
4. La glosa todavía no es la lengua
5. El conjunto de datos que no se nombra
6. Qué cambia en la evaluación cuando el sistema recuerda
7. Los límites de este análisis
8. Fuentes

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.

## Lo que afirma el artículo, y lo que dice de sí mismo

DiscoSign aborda tres fenómenos. El primero es la correferencia espacial: en ASL, a las entidades se les asignan ubicaciones en el espacio de signado y después se hace referencia a ellas señalando de nuevo hacia esos lugares, de modo que un sistema debe recordar que Alice fue ubicada a la izquierda y Bob a la derecha. El segundo son las Cláusulas de Pregunta-Respuesta, las estructuras de tópico-comentario en las que quien signa formula una pregunta retórica y la responde de inmediato, apropiadas en algunas posiciones del discurso e incómodas en otras. El tercero es la coherencia entre concepto y glosa, es decir, que un concepto que aparece cinco veces en un documento debería representarse mediante el mismo signo cada vez, en lugar de alternar entre cuasi-sinónimos. Cada uno se gestiona mediante un módulo que mantiene un registro, lo plantea al modelo de lenguaje como una restricción estricta y, después, comprueba la salida devuelta y corrige las infracciones in situ.

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.

## La consistencia no es exactitud

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."

Los sistemas a nivel de frase fallan de forma ruidosa, y el fallo ruidoso tiene una propiedad infravalorada: el lector lo nota. La inconsistencia se percibe como un defecto. Un sistema que se equivoca de manera consistente parece un sistema que lo dice en serio. Para un lector Sordo que recibe la salida en lugar de compararla con la fuente, un referente erróneo pero estable no resulta en absoluto un error evidente; es sencillamente lo que dice el documento. Esto invierte el supuesto habitual de que una mayor coherencia interna es más segura. La coherencia eleva el coste de un error en el mismo momento en que reduce su visibilidad, y ningún promedio de puntuaciones por frase lo mostrará.

## Qué miden las nuevas métricas, y qué las validó

El artículo actúa aquí con cautela, y merece la pena leer esa cautela con atención porque establece el límite de lo que se puede afirmar razonablemente en fases posteriores. Las nuevas métricas miden la consistencia del índice espacial, la estabilidad del mapeo entre concepto y glosa, y la idoneidad del uso de las Cláusulas de Pregunta-Respuesta. Dos de las tres miden si el sistema hizo lo mismo de principio a fin, lo cual es una propiedad de la salida considerada en sus propios términos y no una medida de si un lector Sordo comprendió el pasaje. Para validar las métricas, dos evaluadores

fluidos en ASL puntuaron treinta y dos historias, ciento cincuenta y dos frases, en una escala de cinco puntos. La consistencia espacial correlacionó moderadamente con sus valoraciones. La consistencia entre concepto y glosa también correlacionó. La métrica de idoneidad para las Cláusulas de Pregunta-Respuesta no alcanzó significación, y el artículo lo atribuye a que los evaluadores no coincidían sobre qué cuenta siquiera como tal cláusula, describiéndose el acuerdo ponderado entre ellos como moderado.

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.

## La glosa todavía no es la lengua

La cobertura que presenta esto como una IA de lengua de signos que supera la traducción frase por frase va más allá de lo que realmente se ha construido. DiscoSign produce glosa, una secuencia de etiquetas escritas que representa los signos manuales, y los autores afirman directamente que los marcadores no manuales no están incluidos. En ASL, buena parte de la estructura del discurso que el artículo intenta preservar reside precisamente en esos canales. La marcación de tópico, el límite entre las mitades de pregunta y respuesta de las propias estructuras que el marco modela, los cambios de rol que indican de quién es la perspectiva que se está expresando, la mirada que vincula una referencia a una ubicación en el espacio: todo esto lo portan el rostro, la cabeza y el cuerpo. Una cadena de glosa puede registrar que un referente es index-j; no puede registrar la mirada que hace que la referencia dé en el blanco.

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.

## El conjunto de datos que no se nombra

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.

## Qué cambia en la evaluación cuando el sistema recuerda

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.

## Los límites de este análisis

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.

Lo útil de este artículo es que hace visible un problema oculto. Los sistemas a nivel de frase venían fallando en el discurso todo este tiempo; simplemente no existía un instrumento capaz de verlo. Ahora existe el comienzo de uno, y llega con un hallazgo que debería leerse tanto como advertencia cual resultado. Hacer que un sistema sea consistente no lo hace correcto. Lo hace correcto o incorrecto de principio a fin, y decide, antes de que nadie lea una sola palabra de la salida, cuál de las dos cosas será.

## Fuentes

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.

---

## Cómo citar

Grizzle, H. M. (2026, September 11). Sistemáticamente equivocado: Apple DiscoSign y la IA de lengua de signos. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003304