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/.

Toujours dans l'erreur : Apple DiscoSign et l'IA en langue des signes

Heather Grizzle · 11 septembre 2026

Des chercheurs d'Apple et de Gallaudet ont construit une IA en langue des signes dotée de mémoire. Lorsqu'un système conserve un état, une erreur précoce cesse d'être bruyante et devient systématique.

Sommaire

1. Ce que l'article affirme, et ce qu'il dit de lui-même
2. La cohérence n'est pas l'exactitude
3. Ce que mesurent les nouvelles métriques, et ce qui les a validées
4. Le gloss n'est toujours pas la langue
5. Le jeu de données qui n'est pas nommé
6. Ce qui change pour l'évaluation lorsque le système se souvient
7. Les limites de cette analyse
8. Sources

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.

Ce que l'article affirme, et ce qu'il dit de lui-même

DiscoSign traite trois phénomènes. Le premier est la coréférence spatiale : en ASL, des entités se voient assigner des emplacements dans l'espace de signation et sont ensuite désignées par un pointage vers ces emplacements, de sorte qu'un système doit se souvenir qu'Alice a été placée à gauche et Bob à droite. Le second concerne les clauses question-réponse, ces structures thème-propos dans lesquelles un signeur pose une question rhétorique et y répond immédiatement, appropriées à certaines positions du discours et maladroites à d'autres. Le troisième est la cohérence concept-gloss, c'est-à-dire qu'un concept apparaissant cinq fois dans un document devrait être rendu par le même signe à chaque fois plutôt que d'alterner entre quasi-synonymes. Chacun est pris en charge par un module qui maintient un registre, l'énonce au modèle de langage comme une contrainte stricte, puis vérifie la sortie renvoyée et corrige les violations en 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 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 cohérence n'est pas l'exactitude

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."

Les systèmes au niveau de la phrase échouent bruyamment, et l'échec bruyant possède une propriété sous-estimée : le lecteur le remarque. L'incohérence est lisible comme un défaut. Un système qui se trompe de manière cohérente ressemble à un système qui le fait exprès. Pour un lecteur Sourd qui reçoit la sortie sans la vérifier par rapport à la source, un référent erroné mais stable n'est pas manifestement une erreur du tout ; c'est simplement ce que dit le document. Cela renverse l'hypothèse habituelle selon laquelle une meilleure cohérence interne serait plus sûre. La cohérence augmente le coût d'une erreur au moment même où elle en réduit la visibilité, et aucune moyenne de scores par phrase ne le révélera.

Ce que mesurent les nouvelles métriques, et ce qui les a validées

L'article est prudent ici, et cette prudence mérite d'être lue attentivement car elle fixe le plafond de ce qui peut raisonnablement être affirmé en aval. Les nouvelles métriques mesurent la cohérence des index spatiaux, la stabilité de l'association concept-gloss, et la pertinence de l'utilisation des clauses question-réponse. Deux des trois mesurent si le système a fait la même chose tout du long, ce qui est une propriété de la sortie considérée en elle-même et non une mesure de la compréhension du passage

par un lecteur Sourd. Pour valider les métriques, deux évaluateurs maîtrisant l'ASL ont noté trente-deux récits, cent cinquante-deux phrases, sur une échelle de cinq points. La cohérence spatiale a montré une corrélation modérée avec leurs jugements. La cohérence concept-gloss a montré une corrélation. La métrique de pertinence pour les clauses question-réponse n'a pas atteint la signification statistique, et l'article attribue cela au fait que les évaluateurs n'étaient pas d'accord sur ce qui constitue une telle clause, l'accord pondéré entre eux étant décrit comme modéré.

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.

Le gloss n'est toujours pas la langue

La couverture médiatique qui présente cela comme l'IA en langue des signes dépassant la traduction phrase par phrase va au-delà de ce qui a été réellement construit. DiscoSign produit du gloss, une séquence d'étiquettes écrites tenant lieu de signes manuels, et les auteurs disent directement que les marqueurs non manuels ne sont pas inclus. En ASL, une grande partie de la structure du discours que l'article cherche à préserver réside précisément dans ces canaux. Le marquage du thème, la frontière entre les moitiés question et réponse des structures mêmes que le cadre modélise, les changements de rôle qui signalent de quel point de vue on parle, le regard qui relie une référence à un emplacement dans l'espace : tout cela est porté par le visage, la tête et le corps. Une chaîne de gloss peut enregistrer qu'un référent est index-j ; elle ne peut pas enregistrer le regard qui fait atterrir cette référence.

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.

Le jeu de données qui n'est pas nommé

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.

Ce qui change pour l'évaluation lorsque le système se souvient

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.

Les limites de cette 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.

Ce qui est utile dans cet article, c'est qu'il rend visible un problème caché. Les systèmes au niveau de la phrase échouaient depuis toujours face au discours ; il n'existait tout simplement aucun instrument capable de le voir. Il y en a maintenant l'ébauche d'un, et il arrive avec une conclusion qui devrait être lue autant comme un avertissement que comme un résultat. Rendre un système cohérent ne le rend pas correct. Cela le rend correct ou incorrect de bout en bout, et décide, avant même que quiconque ait lu un seul mot de la sortie, lequel des deux ce sera.

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

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.

Comment citer

Grizzle, H. M. (2026, September 11). Toujours dans l'erreur : Apple DiscoSign et l'IA en langue des signes. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003304