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/.

Consistentemente Errado: Apple DiscoSign e IA para Língua de Sinais

Heather Grizzle · 11 de setembro de 2026

Investigadores da Apple e da Gallaudet construíram uma IA para língua de sinais que tem memória. Quando um sistema mantém estado, um erro inicial deixa de ser ruidoso e passa a ser sistemático.

Conteúdo

1. O que o artigo afirma, e o que diz sobre si mesmo
2. Consistência não é precisão
3. O que as novas métricas medem, e o que as validou
4. Gloss ainda não é a língua
5. O conjunto de dados que não é nomeado
6. O que muda na avaliação quando o sistema tem memória
7. Os limites desta análise
8. Fontes

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.

O que o artigo afirma, e o que diz sobre si mesmo

O DiscoSign aborda três fenómenos. O primeiro é a correferência espacial: em ASL, as entidades recebem localizações no espaço de sinalização e são posteriormente referidas apontando de volta para elas, pelo que um sistema tem de se lembrar que Alice foi colocada à esquerda e Bob à direita. O segundo são as Cláusulas de Pergunta-Resposta, as estruturas de tópico-comentário em que quem sinaliza coloca uma pergunta retórica e responde-lhe de imediato, o que é apropriado em algumas posições do discurso e estranho noutras. O terceiro é a consistência conceito-gloss, o que significa que um conceito que aparece cinco vezes num documento deveria ser representado pelo mesmo sinal de cada vez, em vez de alternar entre quase-sinónimos. Cada um é tratado por um módulo que mantém um registo, apresenta-o ao modelo de linguagem como uma restrição rígida, e depois verifica o resultado devolvido e corrige as violações diretamente.

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.

Consistência não é precisão

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."

Os sistemas ao nível da frase falham de forma ruidosa, e a falha ruidosa tem uma propriedade subestimada: o leitor repara nela. A inconsistência é legível como um defeito. Um sistema que está errado de forma consistente parece um sistema que quer dizer aquilo. Para um leitor Deaf que recebe o resultado em vez de o verificar contra a fonte, um referente errado mas estável não é sequer obviamente um erro; é simplesmente o que o documento diz. Isto inverte a suposição habitual de que uma melhor coerência interna é mais segura. A coerência aumenta o custo de um erro no exato momento em que reduz a visibilidade desse erro, e nenhuma média de pontuações por frase o mostrará.

O que as novas métricas medem, e o que as validou

O artigo é cuidadoso aqui, e o cuidado vale a pena ser lido de perto porque estabelece o teto do que pode ser afirmado de forma responsável mais adiante. As novas métricas medem a consistência do índice espacial, a estabilidade do mapeamento conceito-gloss, e a adequação do uso das Cláusulas de Pergunta-Resposta. Duas das três medem se o sistema fez a mesma coisa ao longo do texto, o que é uma propriedade do resultado considerada nos seus próprios termos e não uma medida de se um leitor Deaf compreendeu a passagem. Para validar as métricas, dois avaliadores fluentes em ASL

classificaram trinta e duas histórias, cento e cinquenta e duas frases, numa escala de cinco pontos. A consistência espacial correlacionou-se moderadamente com os seus juízos. A consistência conceito-gloss correlacionou-se. A métrica de adequação para as Cláusulas de Pergunta-Resposta não atingiu significância, e o artigo atribui isto ao facto de os avaliadores discordarem sobre o que sequer conta como tal cláusula, com a concordância ponderada entre eles descrita como moderada.

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 ainda não é a língua

A cobertura que apresenta isto como IA para línguas de sinais a ir além da tradução frase-a-frase está a ir mais longe do que aquilo que foi construído. O DiscoSign produz gloss, uma sequência de rótulos escritos que substitui os sinais manuais, e os autores afirmam diretamente que os marcadores não-manuais não estão incluídos. Em ASL, grande parte da estrutura do discurso que o artigo tenta preservar reside precisamente nesses canais. A marcação de tópico, a fronteira entre as metades de pergunta e resposta das próprias estruturas que o framework modela, as mudanças de papel que assinalam de quem é a perspetiva a ser expressa, o olhar que liga uma referência de volta a um local no espaço: estes são transportados pela face, cabeça e corpo. Uma cadeia de gloss pode registar que um referente é index-j; não pode registar o olhar que faz a referência funcionar.

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.

O conjunto de dados que não é nomeado

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.

O que muda na avaliação quando o sistema tem memória

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.

Os limites desta análise

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.

O aspeto útil deste artigo é que torna visível um problema oculto. Os sistemas ao nível da frase estavam sempre a falhar no discurso; simplesmente não havia um instrumento que o conseguisse ver. Agora existe o início de um, e chega com uma conclusão que deve ser lida tanto como um aviso quanto como um resultado. Tornar um sistema consistente não o torna correto. Torna-o correto ou incorreto do início ao fim, e decide, antes de alguém ler uma única palavra do resultado, qual dos dois será.

Fontes

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 Rese-

arch, 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.

Como citar

Grizzle, H. M. (2026, September 11). Consistentemente Errado: Apple DiscoSign e IA para Língua de Sinais. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003304