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/.

Her Zaman Yanlış: Apple DiscoSign ve İşaret Dili Yapay Zekası

Heather Grizzle · September 11, 2026

Apple ve Gallaudet araştırmacıları, hafızası olan bir işaret dili yapay zekası geliştirdi. Bir sistem durumu kaydettiğinde, erken bir hata rastgele olmaktan çıkar ve sistematik hale gelir.

İçindekiler

1. Makalenin iddia ettikleri ve kendisi hakkında söyledikleri
2. Tutarlılık doğruluk değildir
3. Yeni ölçütler neyi ölçüyor ve bunları ne doğruladı
4. Gloss hâlâ dil değildir
5. Adı verilmeyen veri kümesi
6. Sistem hatırladığında değerlendirme için değişen şey
7. Bu analizin sınırları
8. Kaynaklar

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 explicit 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.

Makalenin iddia ettikleri ve kendisi hakkında söyledikleri

DiscoSign üç olguyu ele alıyor. Birincisi uzamsal eşgönderim: ASL'de varlıklara işaret uzayında konumlar atanır ve daha sonra onlara geri işaret ederek gönderme yapılır, bu yüzden bir sistem Alice'in sola, Bob'un sağa yerleştirildiğini hatırlamak zorundadır. İkincisi, bir işaretleyicinin retorik bir soru sorup hemen cevapladığı konu-yorum yapıları olan Soru-Cevap Yan Tümceleri'dir; bunlar bazı söylem konumlarında uygun, bazılarında ise beceriksizdir. Üçüncüsü kavram-gloss tutarlılığıdır, yani bir belgede beş kez geçen bir kavram, yakın eşanlamlılar arasında değişmek yerine her seferinde aynı işaretle gösterilmelidir. Her biri bir kayıt tutan, bunu dil modeline katı bir kısıtlama olarak belirten ve ardından döndürülen çıktıyı kontrol edip ihlalleri yerinde düzelten bir modül tarafından ele alınır.

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.

Tutarlılık doğruluk değildir

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."

Cümle düzeyindeki sistemler gürültülü bir şekilde başarısız olur ve gürültülü başarısızlığın hafife alınan bir özelliği vardır: okuyucu bunu fark eder. Tutarsızlık bir kusur olarak okunabilir. Tutarlı bir şekilde yanlış olan bir sistem, kastedildiği gibi görünür. Çıktıyı kaynakla karşılaştırmak yerine alan Sağır bir okuyucu için, sabit bir yanlış gönderge açıkça bir hata bile değildir; basitçe belgenin söylediği şeydir. Bu, daha iyi iç tutarlılığın daha güvenli olduğu yönündeki olağan varsayımı tersine çevirir. Tutarlılık, bir hatanın görünürlüğünü azalttığı anda maliyetini artırır ve cümle başına puanların hiçbir ortalaması bunu göstermez.

Yeni ölçütler neyi ölçüyor ve bunları ne doğruladı

Makale burada dikkatli ve bu dikkat, aşağı akışta sorumlu bir şekilde ne iddia edilebileceğinin tavanını belirlediği için yakından okunmaya değer. Yeni ölçütler uzamsal indeks tutarlılığını, kavram-gloss eşleme kararlılığını ve Soru-Cevap Yan Tümcesi kullanımının uygunluğunu ölçüyor. Üçünden ikisi, sistemin baştan sona aynı şeyi yapıp yapmadığını ölçer; bu, çıktının kendi başına değerlendirilen bir özelliğidir ve bir Sağır okuyucunun metni anlayıp anlamadığının bir ölçüsü değildir. Ölçütleri doğrulamak için, ASL'ye akıcı iki değerlendirici, otuz iki hikayeyi, yüz elli iki cümleyi beş puanlık bir ölçekte derecelendirdi. Uzamsal tutarlılık, yargılarıyla orta düzeyde ilişkiliydi. Kavram-gloss tutarlılığı ilişkiliydi. Soru-Cevap Yan Tümceleri için uygunluk

ölçütü anlamlılığa ulaşmadı ve makale bunu, değerlendiricilerin böyle bir yan tümcenin ne sayıldığı konusunda anlaşamamasına bağlıyor; aralarındaki ağırlıklı uyum orta düzeyde olarak tanımlanıyor.

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 hâlâ dil değildir

Bunu, işaret dili yapay zekasının cümle cümle çevirinin ötesine geçtiği şeklinde çerçeveleyen haberler, inşa edilenin ötesine uzanıyor. DiscoSign, el işaretlerinin yerini tutan yazılı bir etiket dizisi olan gloss üretir ve yazarlar el dışı işaretleyicilerin dahil edilmediğini açıkça belirtir. ASL'de, makalenin korumaya çalıştığı söylem yapısının büyük bir kısmı tam olarak bu kanallarda yaşar. Konu işaretleme, çerçevenin modellediği aynı yapıların soru ve cevap yarıları arasındaki sınır, kimin bakış açısının ifade edildiğini belirten rol değişimleri, bir göndergeyi uzaydaki bir konuma bağlayan göz bakışı: bunlar yüz, baş ve vücut tarafından taşınır. Bir gloss dizesi bir göndergenin indeks-j olduğunu kaydedebilir; göndermeyi yerine oturtan bakışı kaydedemez.

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.

Adı verilmeyen veri kümesi

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.

Sistem hatırladığında değerlendirme için değişen şey

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.

Bu analizin sınırları

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.

Bu makalenin yararlı yanı, gizli bir sorunu görünür kılmasıdır. Cümle düzeyindeki sistemler zaten söylem konusunda başarısız oluyordu; bunu görebilecek bir araç yoktu, hepsi bu. Şimdi bunun bir başlangıcı var ve bir sonuç kadar bir uyarı olarak okunması gereken bir bulguyla geliyor. Bir sistemi tutarlı kılmak onu doğru kılmaz. Onu baştan sona doğru ya da yanlış kılar ve çıktının bir kelimesi bile okunmadan önce, hangisinin olacağına karar verir.

Kaynaklar

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

TİPS

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.

Nasıl alıntılanır

Grizzle, H. M. (2026, September 11). Her Zaman Yanlış: Apple DiscoSign ve İşaret Dili Yapay Zekası. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003304