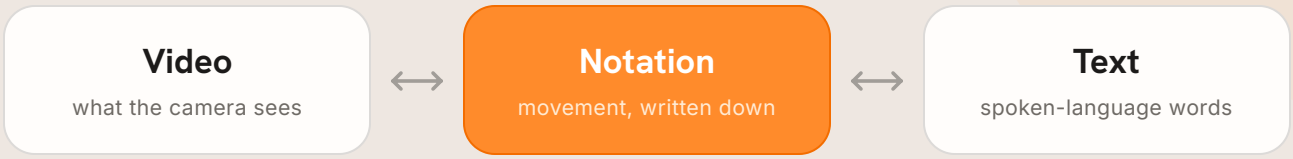
Rylo

RYLO RESEARCH · WHITE PAPER

The Future of Sign Language Translation is Transcription

Why teaching machines to *write down* movement — not guess at meaning — is the missing step that finally makes sign language translation work.



Dr. Amit Moryossef
Head of Research, Rylo

FOR A GENERAL AUDIENCE
EDITION 2026

START HERE

The whole idea, in one minute

Today's AI can read and it can listen. It cannot really *watch*. That is why, after decades of effort, there is still no reliable app that translates a signed conversation into spoken-language text — or the other way around.

The argument of this paper, in plain words:

We can't translate sign language directly from video. We first have to teach machines to **write movement down** — to transcribe it into a simple, universal notation. Translation then becomes a textual language problem. **Advancement is unattainable without a form-based transcription system.**

THE PROBLEM

There's no large-scale video data — and what exists is expensive and slow to process. There's also no everyday written form.

THE MOVE

Put a written notation in the middle. **SignWriting** captures the *form* of a sign — hands, face, motion — on the page.

THE PAYOFF

One half becomes universal computer vision. The other becomes ordinary translation. Both suddenly tractable.

“Give signed languages the written step that spoken languages already take for granted — and the rest of the problem falls into place.”

01

THE PROBLEM

AI learned to read and listen. It never learned to watch.

Up to 70 million deaf people sign every day, across 200–300 different signed languages — and modern language technology leaves nearly all of them out. Here's why the obvious approaches keep hitting a wall.

WHO GETS LEFT BEHIND

The web runs on text. Signing doesn't.

Search, chat assistants, captions, translation — almost everything useful that AI does with language starts from *written words*. Signed languages have no everyday writing system, so they never make it into the pipeline. The result is a community of millions that today's tools simply don't see.

70M

Deaf people worldwide whose first language is a signed one.

200–300

Distinct signed languages — each with its own grammar and vocabulary.

0

Signed languages with a writing system in common, everyday use.

Sign language is a real language — not gestures

Signed languages are full natural languages with their own grammar, word order and regional dialects. They are **not** a hand-spelled version of English or any spoken language. Meaning is carried by the hands *and* the face, the eyes, the posture, and the space around the signer — often all at once.

The simultaneity catch

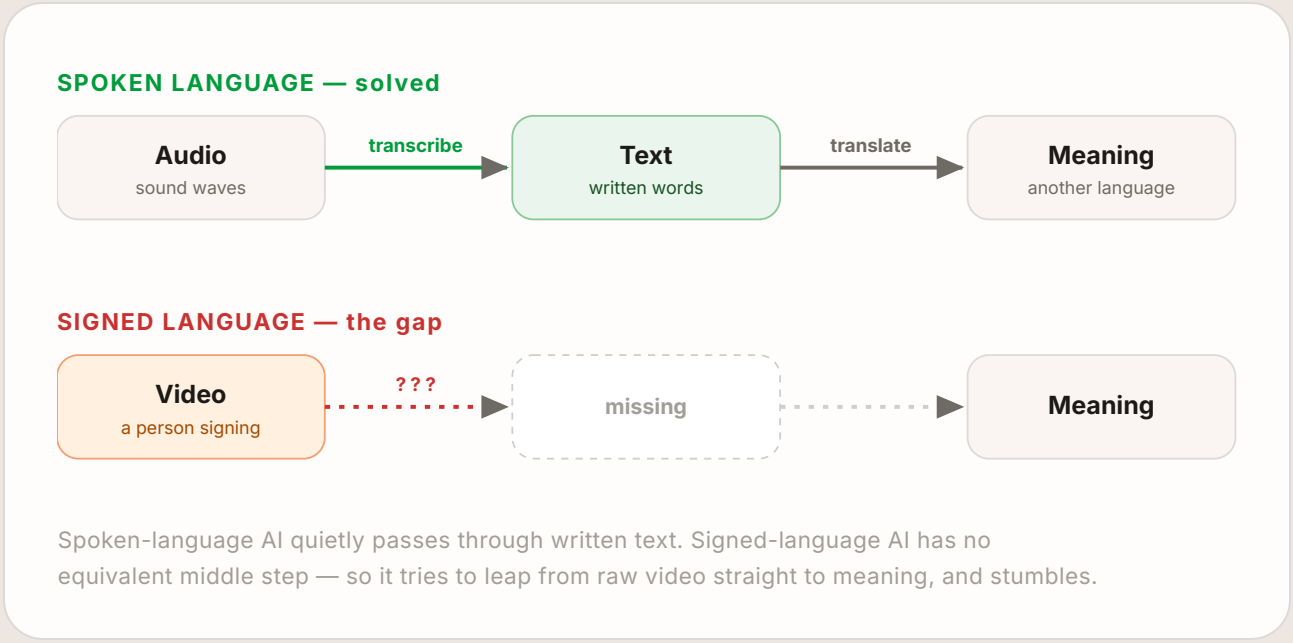
A signer can show *who*, *what*, and *how-they-feel-about-it* in a single moment. Written words are forced to put one thing after another. Any honest sign notation has to capture several things happening **at the same time**.

If your language has no written form, the entire modern AI stack skips you.

HOW SPEECH AI ACTUALLY WORKS

There's a hidden step you never think about

When you talk to a voice assistant, it doesn't jump from sound straight to meaning. It first **transcribes** your speech into written text. Everything clever happens *after* that. Signed languages are missing exactly this step.



Spoken-language AI quietly passes through written text. Signed-language AI has no equivalent middle step — so it tries to leap from raw video straight to meaning, and stumbles.

The fix writes itself. Don't invent a smarter leap. Build the missing middle box — a way to *write signing down* — and let the rest of the pipeline work the way it already does for speech.

WHY BRUTE FORCE WON'T SAVE US

The data desert

Modern AI is hungry — it learns from oceans of examples. For signed languages, that ocean is a puddle, and it's brutally expensive to fill. "Just train a bigger model" is not an option here.

50,000 h

Audio behind speech recognition — and that's just to **transcribe** speech into text, the easier task.

1,150 h

All the signing video we have — and we need it for **translation**, which is far harder. Only **50 hours** are public.

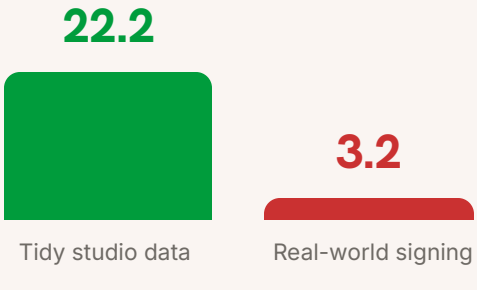
And it isn't even a fair fight. Those 50,000 hours only buy *transcription* — speech into text in the same language. Our 1,150 hours must buy *translation* into another language entirely: far harder, with far less data.

The annotation tax

Labeling signed video by hand can take up to **600 minutes of expert work for a single minute** of footage. That's roughly **ten hours of labor per minute** — which is exactly why the data stays scarce.

And the models don't survive the real world

On a tidy, narrow dataset — a single TV weather presenter — a standard model looks respectable. Point it at open, everyday signing and the quality collapses. Same model, same code: it simply hasn't seen enough of the world.



BLEU score (0–100, higher is better) · same model

When data is this scarce, the only way forward is to need less of it.

TWO TEMPTING SHORTCUTS

Why the obvious fixes don't work

SHORTCUT 1

Just write each sign as an English word ("glosses")

A *gloss* labels each sign with a spoken-language word — like subtitling a sign with HOUSE or WHAT. It's quick, and it's everywhere in research. It's also lossy.

- ✕ One line of words can't capture hands, face and motion happening together.
- ✕ The labels are different for every language — they don't transfer.
- ✕ The same sign gets different labels; the same label hides different meanings.

SHORTCUT 2

Build one big model: video straight to text

Skip the middle entirely and hope a single network learns the leap. With so little data, it doesn't — and the result is impossible to live with.

- ✕ Needs mountains of paired video↔text we simply don't have.
- ✕ A black box: when it's wrong, there's no middle to inspect or fix.
- ✕ Has to be rebuilt from scratch for every single language.

Beware "sign language translation" that isn't

Many flashy demos — including sensor gloves — actually just detect finger-spelling or a handful of isolated signs. They reduce a rich grammar to hand-waving, reinforce the myth that signing is "all in the hands," and are widely rejected by the Deaf community. Recognizing a few signs is not translation.

Both shortcuts skip the same thing: a faithful, shared way to write a sign down.

02

THE IDEA

Write the movement, not the meaning.

Put a single, universal notation in the middle of the pipeline. It captures *what the body did* — not what it meant. That one move splits an impossible problem into two solvable ones.

THE MISSING PIECE

Give signing a written form

The breakthrough isn't a bigger model. It's a humble one: a **notation** — a way to write down a sign on the page, the way letters write down sounds. Get the notation right and everything downstream gets easier.

What a good sign notation has to be

Universal

One system for every signed language — so data and tools are shared, not rebuilt 300 times.

About form, not meaning

It records the handshape, the motion, the face — the *how* — leaving translation for later.

Multi-dimensional

It can show several things happening at once, the way a real sign does.

Human-readable

People — and machines — can actually read and write it, check it, and trust it.

Why "form, not meaning" is the whole trick

Writing down *movement* doesn't require understanding any language. A camera in Tokyo and a camera in Nairobi capture the same kind of motion. So the hard, data-hungry part — vision — can be learned **once, for all signed languages at once.**

Letters didn't make speech smarter. They made everything after them possible.

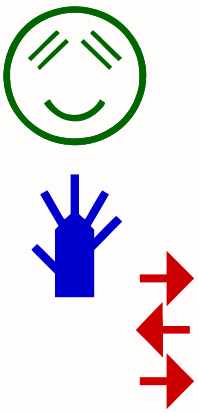
A NOTATION THAT ALREADY EXISTS

Meet SignWriting

Invented by Valerie Sutton in 1974, **Sutton SignWriting** writes a sign the way you'd see it — a little picture of the body in action. Handshapes, contact, movement arrows, and facial expressions are arranged in two dimensions, just as they happen.

Crucially, it's **visual, universal, and two-dimensional**. It can record several things at once, it works for any signed language, and a trained reader can look at it and know exactly what the body did. There's even a computer-friendly text encoding, so models can read and write it like any other string.

SIGNWRITING — REAL NOTATION



A single sign written in Sutton SignWriting — handshapes, facial expression and movement, colorized by symbol type, all on one 2D canvas.

Why this one, over the alternatives

- ✓ **Beats glosses:** it keeps the form that English-word labels throw away.
- ✓ **Beats other notations:** it stays human-readable and faithfully 2D, instead of a cryptic single line.
- ✓ **One alphabet, every language:** the same symbols serve ASL, DGS, LSF and the rest.

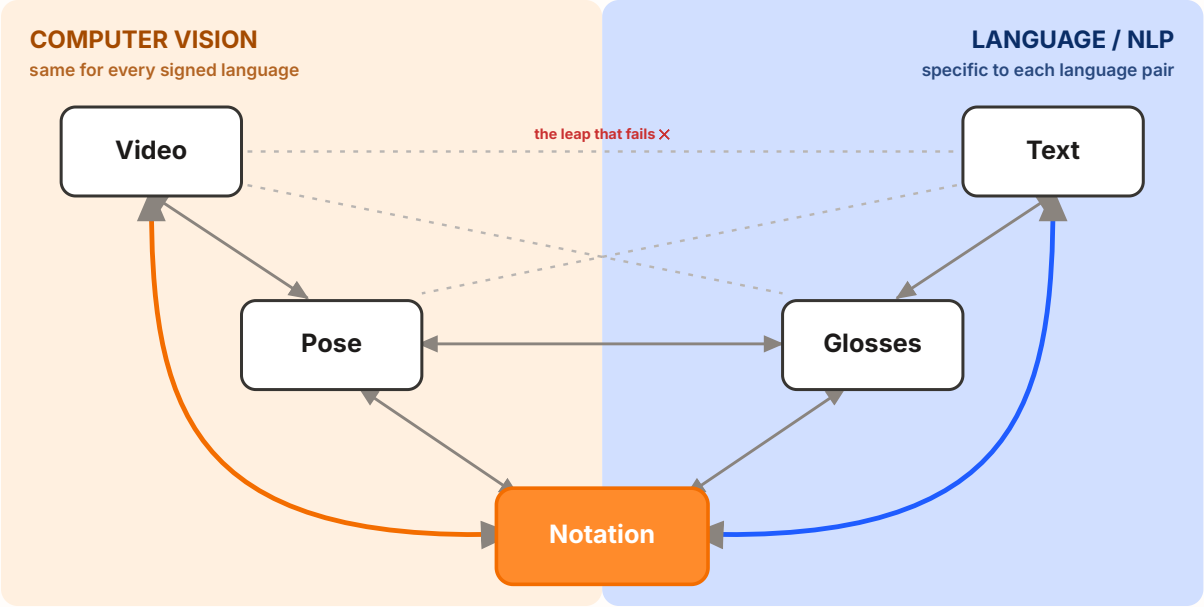
One phrase, five ways to write it

The same ASL question — *"What is your name?"* — can be stored as raw **video**, a tracked **skeleton**, **SignWriting**, a rival notation, or a string of **glosses**. Only SignWriting keeps the full, language-independent *form* in a way both people and machines can use.

THE MAP

One pivot splits the field in two

Picture every way of representing a sign as a dot, and every conversion as a line. Drop **Notation** in the center, and a clean border appears: a **vision** side that's the same for all languages, and a **language** side that's ordinary translation.



Every valid route runs **through Notation**. The orange curve (Video↔Notation) is universal vision; the blue curve (Notation↔Text) is ordinary translation. The dashed grey lines are the cross-the-gap shortcuts that don't hold up.

HALF ONE · COMPUTER VISION

Video ↔ Notation: writing & animating movement

This half is pure vision, and it's **language-agnostic**. Read a video, write the SignWriting. Read SignWriting, animate it back into a moving avatar. No translation required — so it can be learned once and reused for every signed language on Earth.

VIDEO → NOTATION

Transcription

Watch the hands, face and body, and write down the form as SignWriting — the same job your phone does turning speech into text, but for movement.

NOTATION → VIDEO

Animation

Turn SignWriting into a smooth, signing avatar — a 3D character or a photo-realistic one — so written signs become watchable again.

Why this is the unlock for the data desert

Because writing movement needs *no* language understanding, every scrap of annotated signing — from any country — trains the **same** model. A dictionary of German signs helps transcribe French ones. Scarce data stops being 300 tiny problems and becomes one shared, growing pool.

Honest status

A fully automatic video-to-SignWriting transcriber doesn't exist yet — it's the field's most important open task. The point of this paper is that it is the *right* task: the one piece that makes everything else fall into place.

- ✓ Learned **once** for all languages
- ✓ Every dataset helps every language
- ✓ A pure, well-defined vision problem

HALF TWO · LANGUAGE

Notation ↔ Text: ordinary translation, at last

Once a sign is written as SignWriting, translating it to English (or back) is just **text-to-text translation** — the most mature problem in all of AI. No exotic tricks needed: feed it through a normal translation model.

Quality beats quantity — dramatically

Cleaning up a half-million-example sign translation dataset (**no new footage**, just better data) sent translation scores from near-zero to the 20s and 30s on the BLEU scale. The whole cleanup cost about **$530** in AI usage — a reminder that better data, not just more of it, is what moves the needle.

~0
BLEU before cleanup (raw, noisy data)

22–30
BLEU after cleanup — same examples, tidied

1M+
Clean training pairs, built for ~$530

The universal-language dividend

Because SignWriting is shared across signed languages, this side inherits decades of progress in machine translation — and benefits from multilingual training, where languages help each other learn.

Keep it simple

Earlier work tried elaborate, custom machinery to handle SignWriting. Treating it as plain text — and cleaning the data — beat the fancy approach. Simpler *and* better.

Two clean problems beat one impossible one.

03

THE PAYOFF

Build the transcriber, and the future arrives.

One universal way to write movement down turns an unsolved moonshot into two problems the field already knows how to attack — and opens doors well beyond the Deaf community.

THE BOTTOM LINE

Transcription first. Everything else follows.

> **“Advancement is unattainable without a form-based transcription system.”**

It isn't a bigger model or more compute that's been missing. It's a written form. Adopt SignWriting as the pivot and the field cleaves neatly in two:

1. **Computer vision** learns to transcribe video into SignWriting, and to animate it back — once, for all signed languages.
2. **Language research** translates between SignWriting and spoken-language text, riding decades of machine-translation progress.

It reaches beyond signing, too

Hearing people "talk" with their bodies constantly — a shrug, an eye-roll, a thumbs-up that flips a sentence's meaning. A way to *write movement down* could let AI finally read the half of human communication that words leave out.

THE CALL TO ACTION

- → Treat **transcription** as the field's top priority.
- → Split the work cleanly: **vision** vs **language**.
- → Invest in **clean, shared, universal** data.

Writing systems revolutionized spoken language. It's signing's turn.

ABOUT THIS PAPER

Written movement is the key to the whole problem.

About the author

Dr. Amit Moryossef is Head of Research at Rylo. His doctoral work on real-time multilingual sign language processing introduced the transcription-first paradigm this paper summarises, and powers the open *sign.mt* translation demo.

About Rylo

Rylo builds accessible communication tools for people who are deaf and hard of hearing. We invest in the research that makes truly inclusive technology possible.

Go deeper

The full thesis
arxiv.org/abs/2412.01991

The peer-reviewed demo
aclanthology.org/2024.emnlp-demo.19

Try it live
rylo.com/sign/translate

Source: Moryossef, *Real-Time Multilingual Sign Language Processing* (2023/2024).

Rylo