MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

Mini Sign Language Workshop at Max Planck Institute for Psycholinguistics
June 3, 2024

From pose estimation to pretrained models: Doing sign language research with computer vision

Onur Keleş
Department of Linguistics
Boğaziçi University, Istanbul, Turkey

Flow

This presentation will summarize recent findings on TİD (Turkish Sign Language) from the Computational Sign Language Lab at the Department of Linguistics, Boğaziçi University.

Pose Estimation in Sign Language:

1.Morphophonology and Articulatory Energy in Expressing Complex Motions Events in TİD and Age of Acquisition Effects
2.Complexity of Telicity Marking in TİD
3.Getting to the Point: Deciphering the Linguistic Multifunctionality of Pointing in TİD

Use of Transformers in Sign Language
4.A road-map

Morphophonology and Articulatory Energy in Expressing Complex Motions Events in TİD and Age of Acquisition Effects

Onur Keleş
Emre Bilgili
Kadir Gökgöz

Introduction

In a complex motion-event, an agent moves along a path with a manner[1].

Languages encode different ways of exponentiating manner and path in these event.

[1] Talmy 1985; Benedicto et al. 2008; Özyürek et al. 2015; Supalla 1990

Introduction

Sign languages in particular can encode path and manner by using:

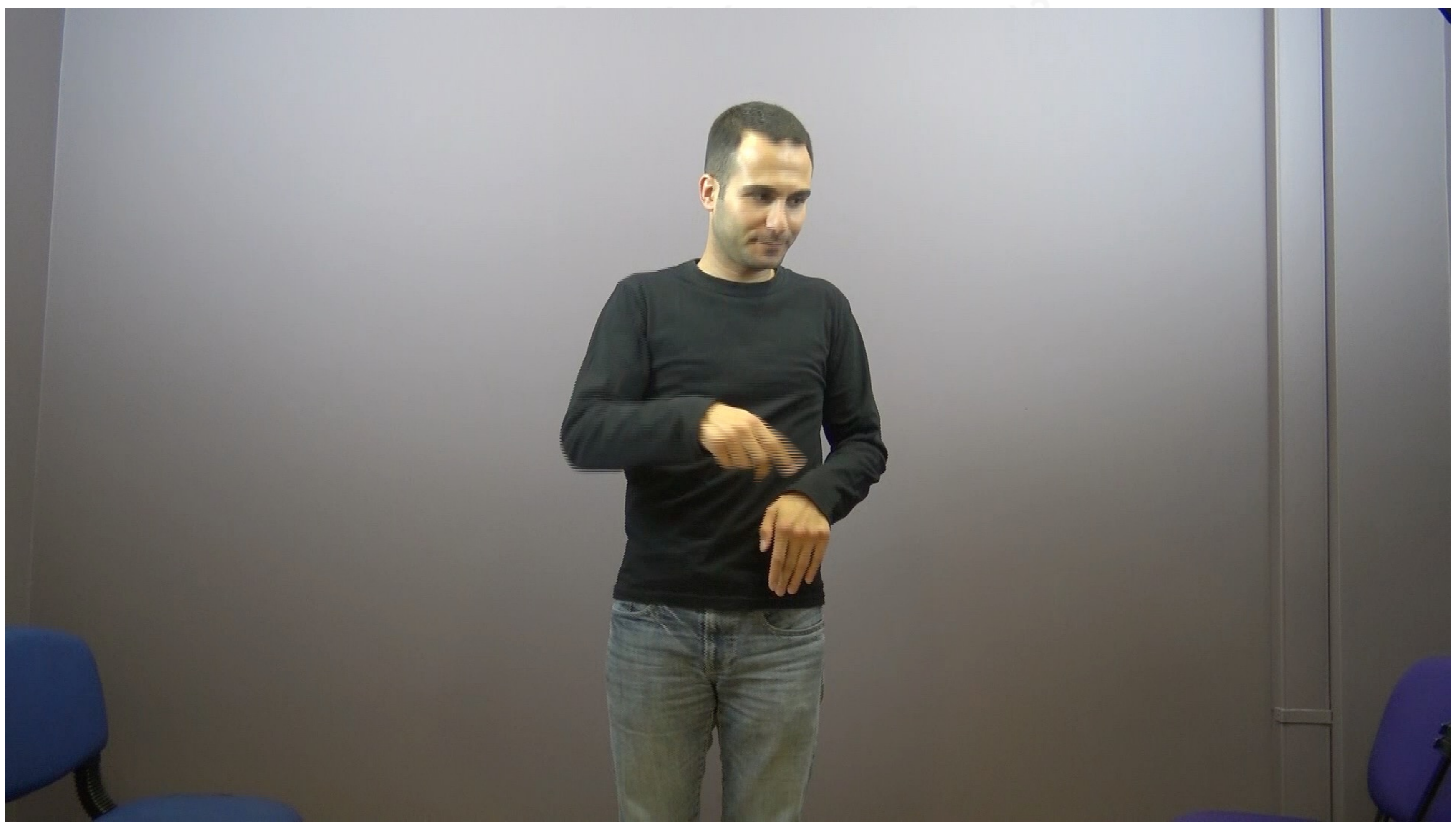
- separate signs of manner and path (sequential form)
- as single sign (conflated form)

[2] This is what Slobin and Hoiting 1994 calls "path-focused (p. 493)"

Introduction

Sign languages in particular can encode path and manner by using:

- separate signs of manner and path (sequential form)
- as single sign (conflated form)



[2] This is what Slobin and Hoiting 1994 calls "path-focused (p. 493)"

Introduction

Sign languages in particular can encode path and manner by using:

- separate signs of manner and path (sequential form)
- as single sign (conflated form)
- or a combination of both (mixed form)[2]

[2] This is what Slobin and Hoiting 1994 calls "path-focused (p. 493)"

Introduction

(1) Sample expressions for Walking on Toes on a Curved Path.

Linguistic Deprivation

- 90 percent of all Deaf are late signers (i.e., born into hearing & non-signing parents[3])
- Late signers receive frequent & regular language exposure after infancy starting from age 4 and onward[4]

[3] Mitchell and Karchmer 2004
[4] Mayberry 2007; Mayberry et al. 2011

Background

Late signers may exhibit linguistic deprivation effects on:

- linguistic abilities in morphosyntax and sentential processing[5]
- lexicon[6]
- pragmatic abilities[7]
- and possibly on spatial language development

[5] Sevgi and Gökgöz 2023; Kayabaşi and Gökgöz 2013; Cheng and Mayberry 2019; Newport 1990
[6] Keleş et al. 2022; Sehyr et al. 2018
[7] Keleş et al. 2023; Cormier et al. 2013

Research Questions

- How are complex motion events expressed in Turkish Sign Language (TİD)?
- Can we estimate the energy spent during the articulation of these complex events with computer vision?
- Do native and late signers differ the energy spent on the expression of complex motion events in TİD?

Stimuli

- 54 items: 9 Manners (Running, Walking-on-Toes, etc.) with 6 Paths each (Curved, Circle, Zigzag, etc.) adapted from[8].

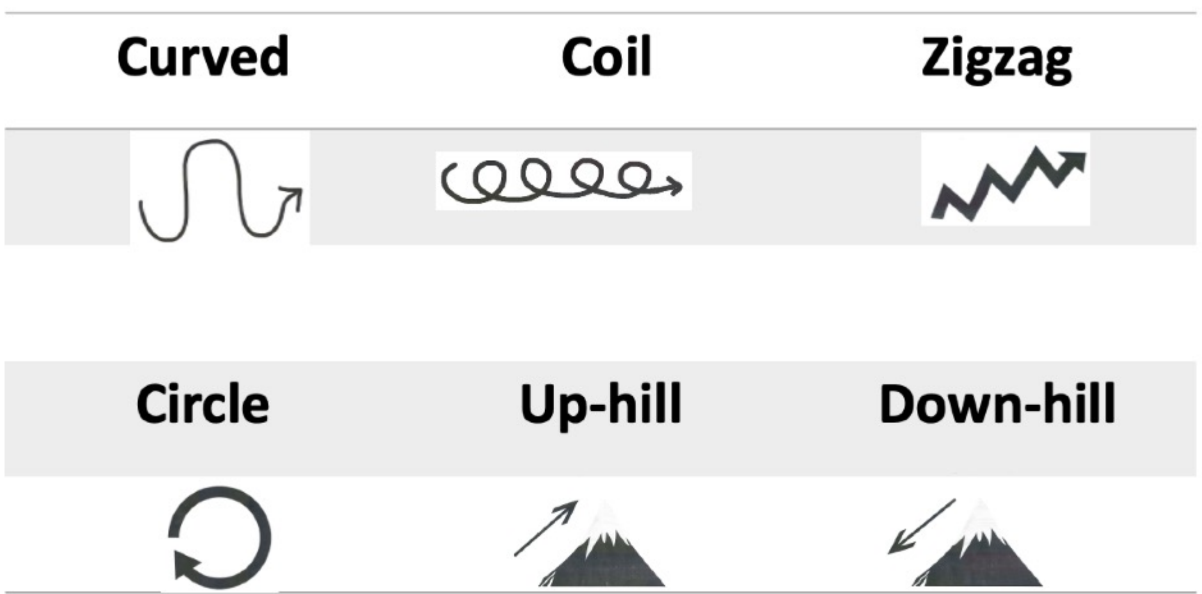


Figure 3: Manner Stimuli

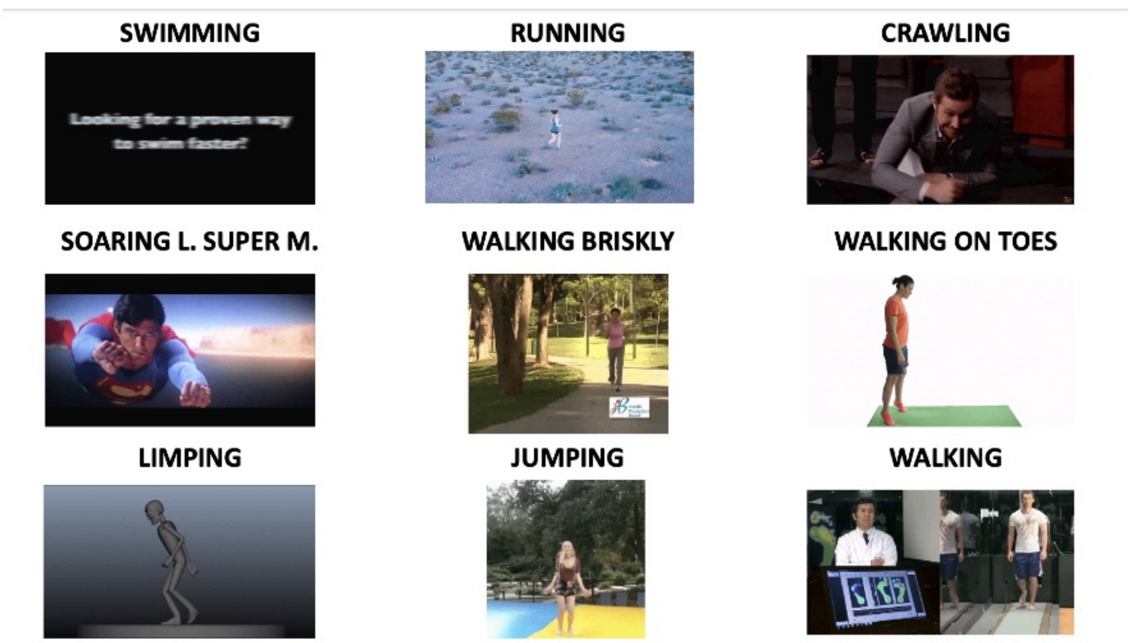


Figure 4: Path Stimuli

[8] Supalla 1990

Participants

10 adult native signers

- 6 females; 4 males, all right-handed.
- All have deaf parents.
- Exposed to TİD from birth.
- Age range at testing: 18-35 (mean age: 27.6).

10 adult late signers

- 4 females; 6 males, all right-handed.
- All have hearing parents.
- Delayed exposure to TİD.
- TİD learning starts with enrollment in a deaf school.
- Age range at testing: 25-51 (mean age: 37.6).
- Mean number of years TİD used: 30.7.

Participants

10 adult native signers

- 6 females; 4 males, all right-handed.
- All have deaf parents.
- Exposed to TİD from birth.
- Age range at testing: 18-35 (mean age: 27.6).

10 adult late signers

- 4 females; 6 males, all right-handed.
- All have hearing parents.
- Delayed exposure to TİD.
- TİD learning starts with enrollment in a deaf school.
- Age range at testing: 25-51 (mean age: 37.6).
- Mean number of years TİD used: 30.7.

Coding

We coded the string types as

- Sequenced (separate manner and path)
- Conflated (simultaneous manner and path)
- Mixed (at least one separate manner or path, followed/preceded by a conflated form)

(2) Sample expressions for Walking on Toes on a Curved Path.

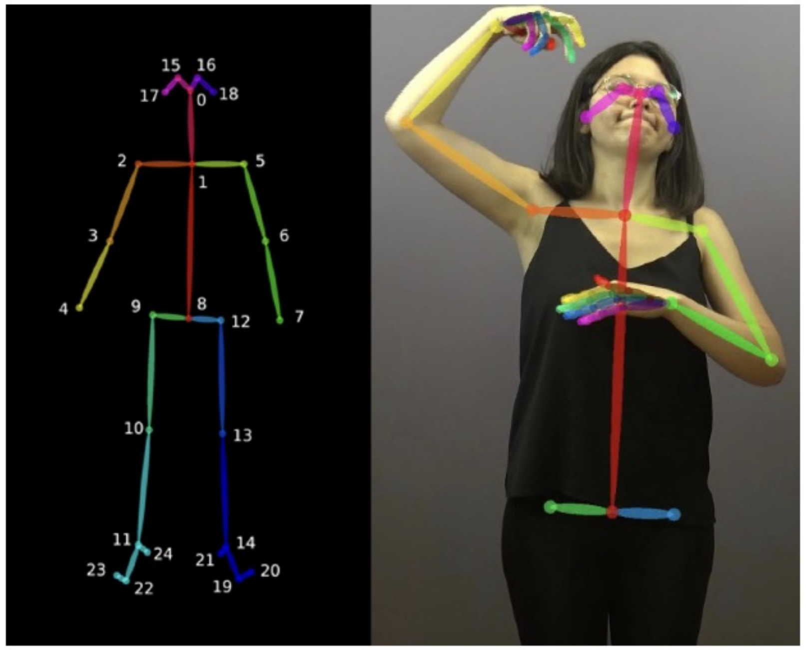
Pose Estimation and Energy Calculation

- For pose estimation, we took a set of screenshots from each movement and processed the images in OpenPose library[9] in Python by marking the torso, shoulder, elbow, wrist and fingers (3).
- To calculate a value for the estimated energy spent on a movement within an expression we assigned relative values to joints according to the body-mass moved by each factoring in the duration of active joints (4).
- We calculated total and average values for each expression. We measured the Right and Left side of the body separately.

[9] Cao et al. 2021

Pose Estimation and Energy Calculation

(3) Joint reference numbers and an example output



- Step 1: Extract 8 frames from the videos per second.
- Step 2: Run OpenPose on each frame and get the coordinates of the joints.
- Step 3: Then, calculate the Euclid distance of each joint between the consecutive frames to calculate energy.

(4) Energy calculation

a. **Relative values assigned to the joints:** Body-midline = 5, Shoulder = 4, Elbow = 3, Wrist = 2, Fingers = 1

b. **Formula**

$$\text{Energy} = \sum_{i=1}^{n} \frac{\text{Duration of active joint } i}{\text{Duration of entire sign}} \times \text{Relative value of joint } i$$

Pose Estimation and Energy Calculation

Results

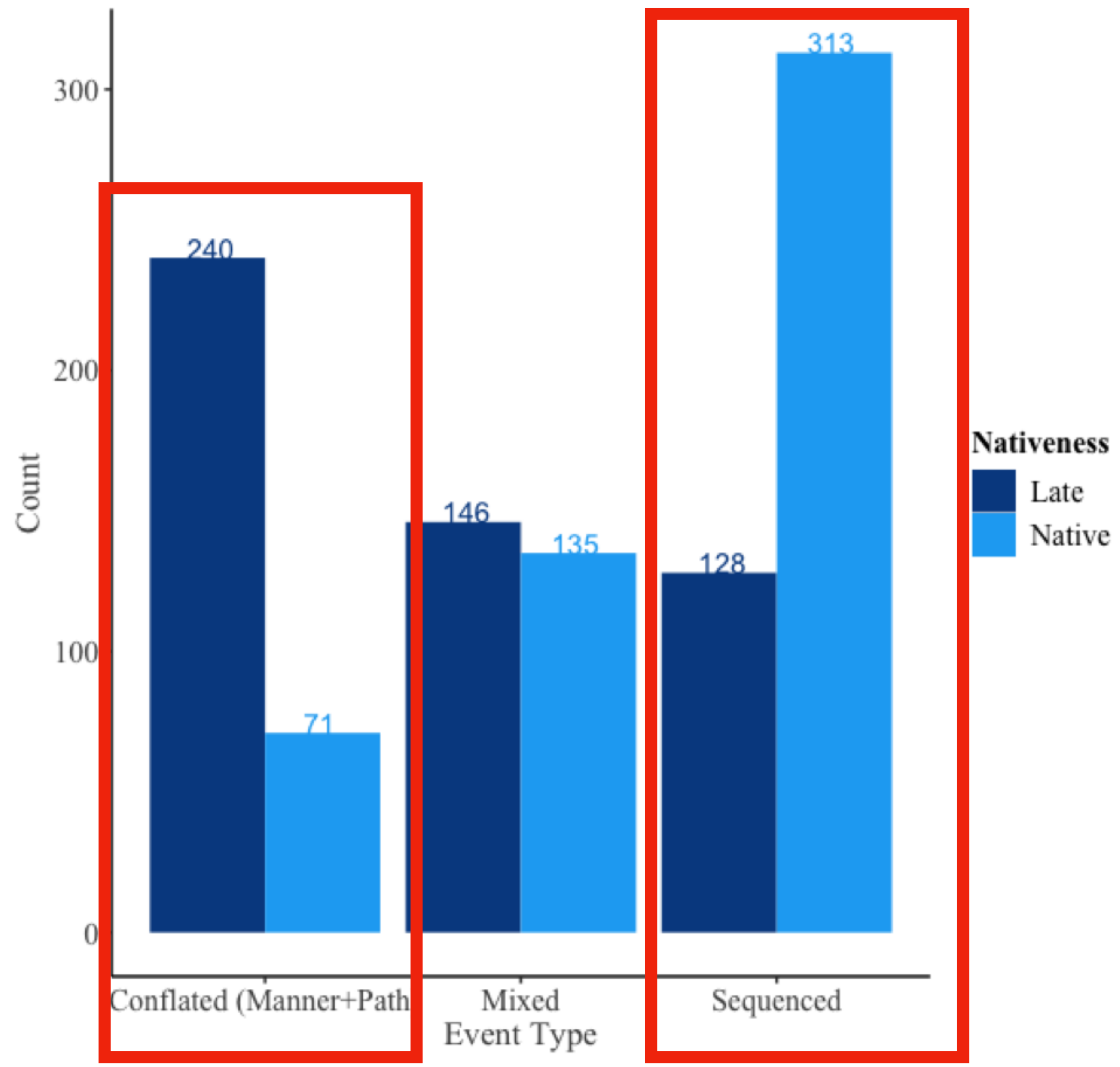


Figure 5: Production of Complex Motion Events by Event Type and Nativeness

- Late signers use conflated forms more than native signers.
- Native signers use sequenced forms more than native signers.

Results

	Number		
Predictors	*Incidence Rate Ratios*	*CI*	*p*
Intercept	15.13	14.11 – 16.19	**<0.001**
Conflated	0.74	0.60 – 0.91	**0.005**
Sequenced	1.75	1.46 – 2.10	**<0.001**
Late	1.10	0.96 – 1.26	0.174
Conflated*Late	9.44	6.30 – 14.35	**<0.001**
Sequenced*Late	0.14	0.10 – 0.20	**<0.001**
Observations	60		
R^2 Nagelkerke	0.980		

Figure 6: Poisson GLM Results

- Two Poisson GLMMs (one with Nativeness as predictor and other without this predictor) were fitted to the data.
- The model with Nativeness as predictor was found to be a better fit for the data.

Results

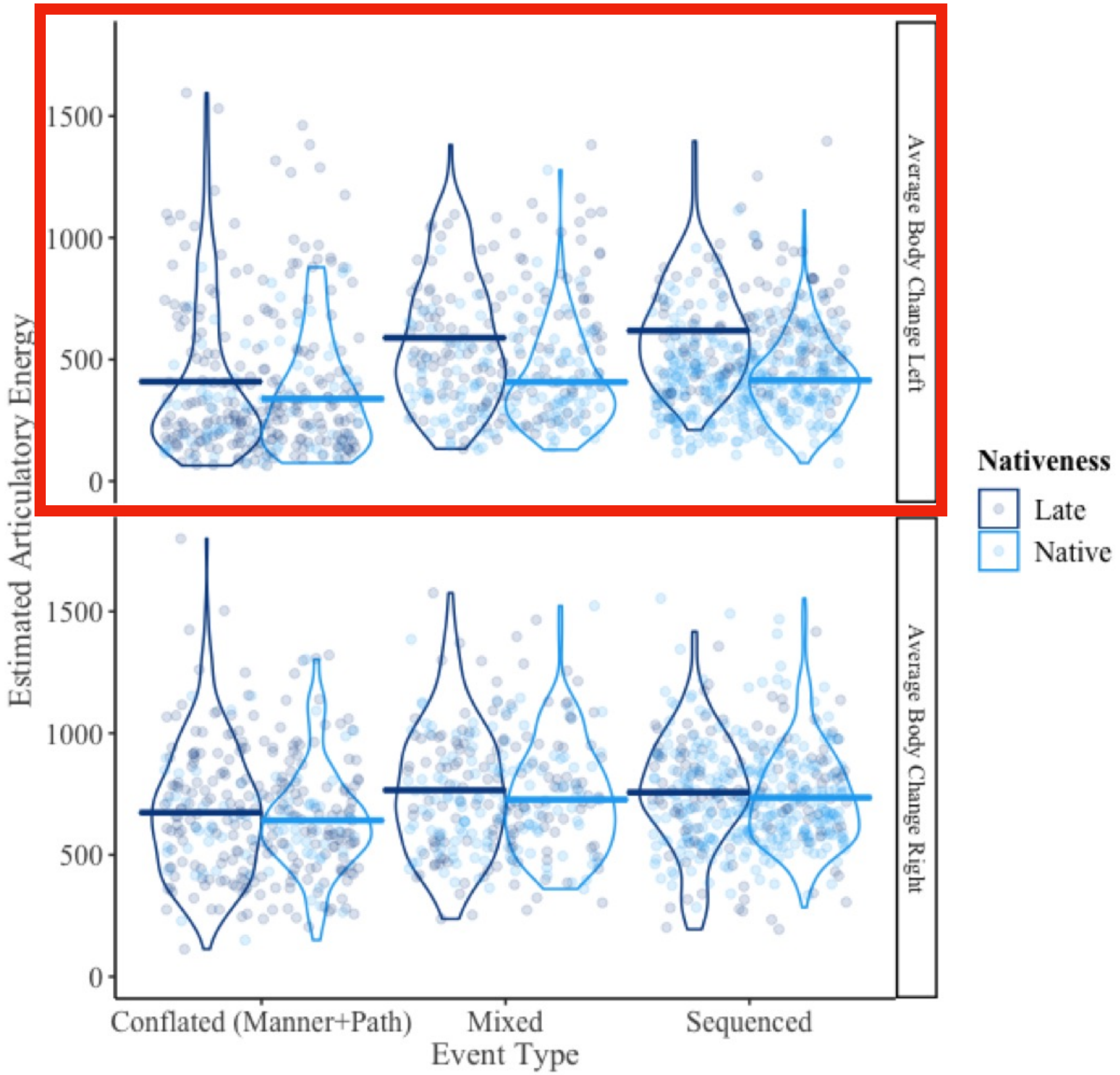


Figure 7: Articulatory Energy by Event Type and Nativeness

Energy spent on a complex motion event increases with

- Late acquisition
- Using a sequenced form
- Using the dominant side

Age of Acquisition Effects

If a signer is a late acquirer, more energy use on the non-dominant side (i.e., "left side of the body").

Results

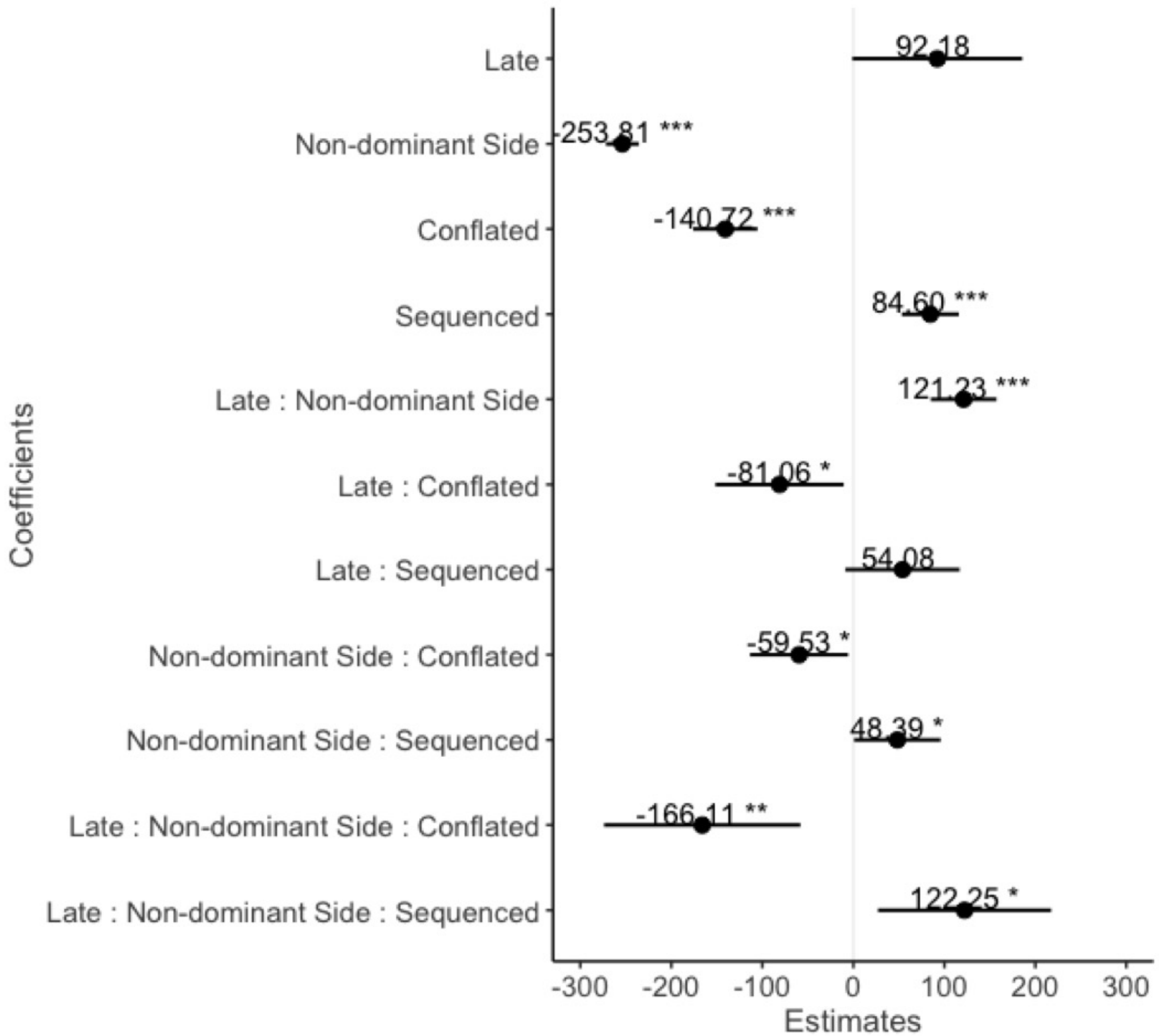


Figure 8: Mixed effects model results

- LME model fitted to Energy values with Nativeness, Body Dominance and Event Type as Fixed Effects.
- Participants and Frames as Random Effects.
- Two-way interaction between Nativeness and Body Dominance.

Results

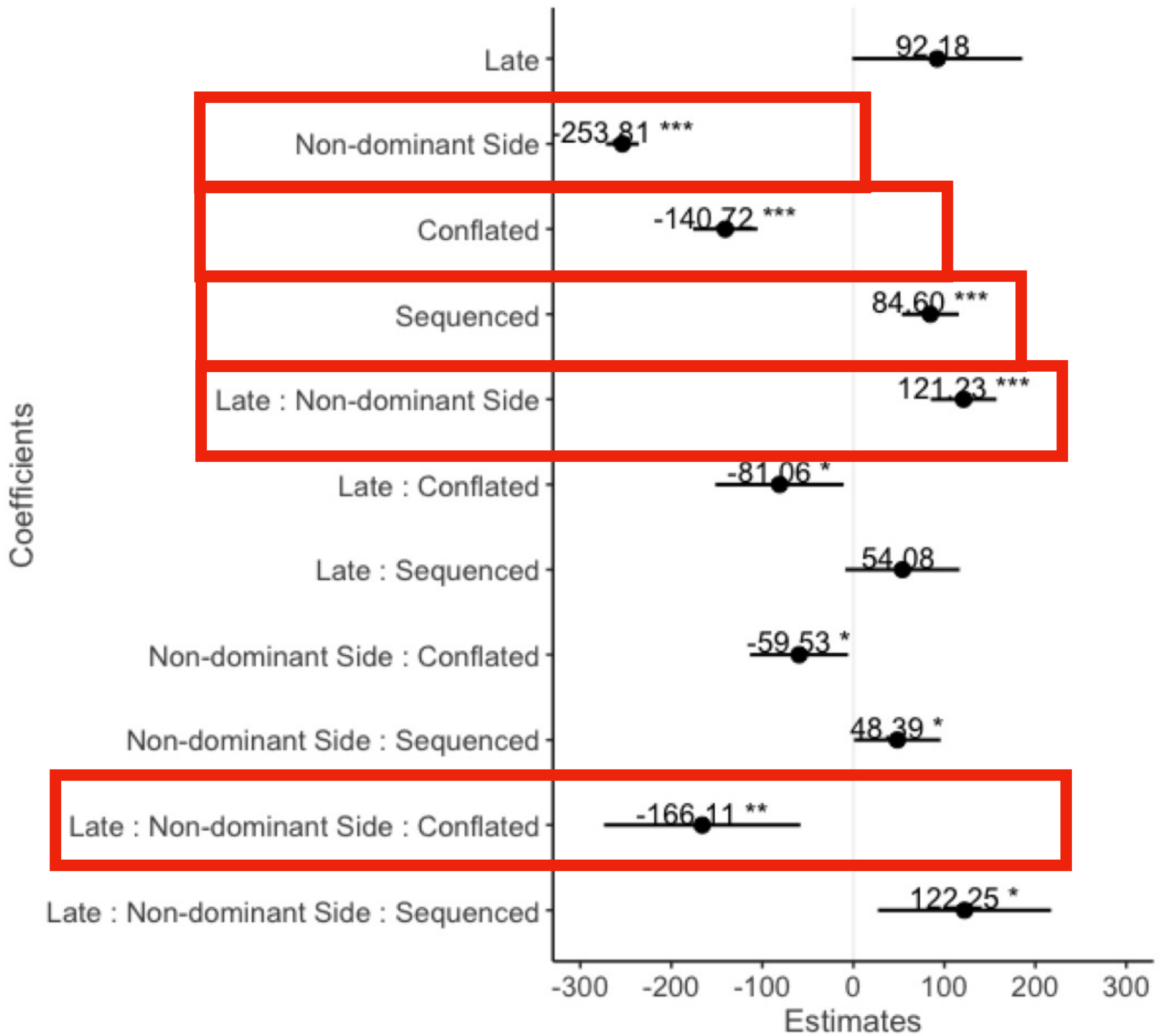


Figure 8: Mixed effects model results

- LME model fitted to Energy values with Nativeness, Body Dominance and Event Type as Fixed Effects.
- Participants and Frames as Random Effects.
- Two-way interaction between Nativeness and Body Dominance.

Discussion

Summary of Results

- Late signers use conflated forms more, similar to home signers[10].
- Native signers used more sequenced forms than late signers.
- Similar frequency of mixed forms in both groups.
- Although conflated forms decreased the use of energy, late signers used more energy overall than native signers.
- Dominance x Nativeness interaction: Late signers used more energy on the non-dominant side.

[10] Özyürek et al. 2015

Discussion

Implications

- Sequential exponence is the default strategy for natives.[a]
- Age-sensitivity in complex motion event production.
- Less inhibitation of the non-dominant side in late signers.
- The idea of Reactive Force[b]

[a] Supalla 1990
[b] Sanders and Napoli

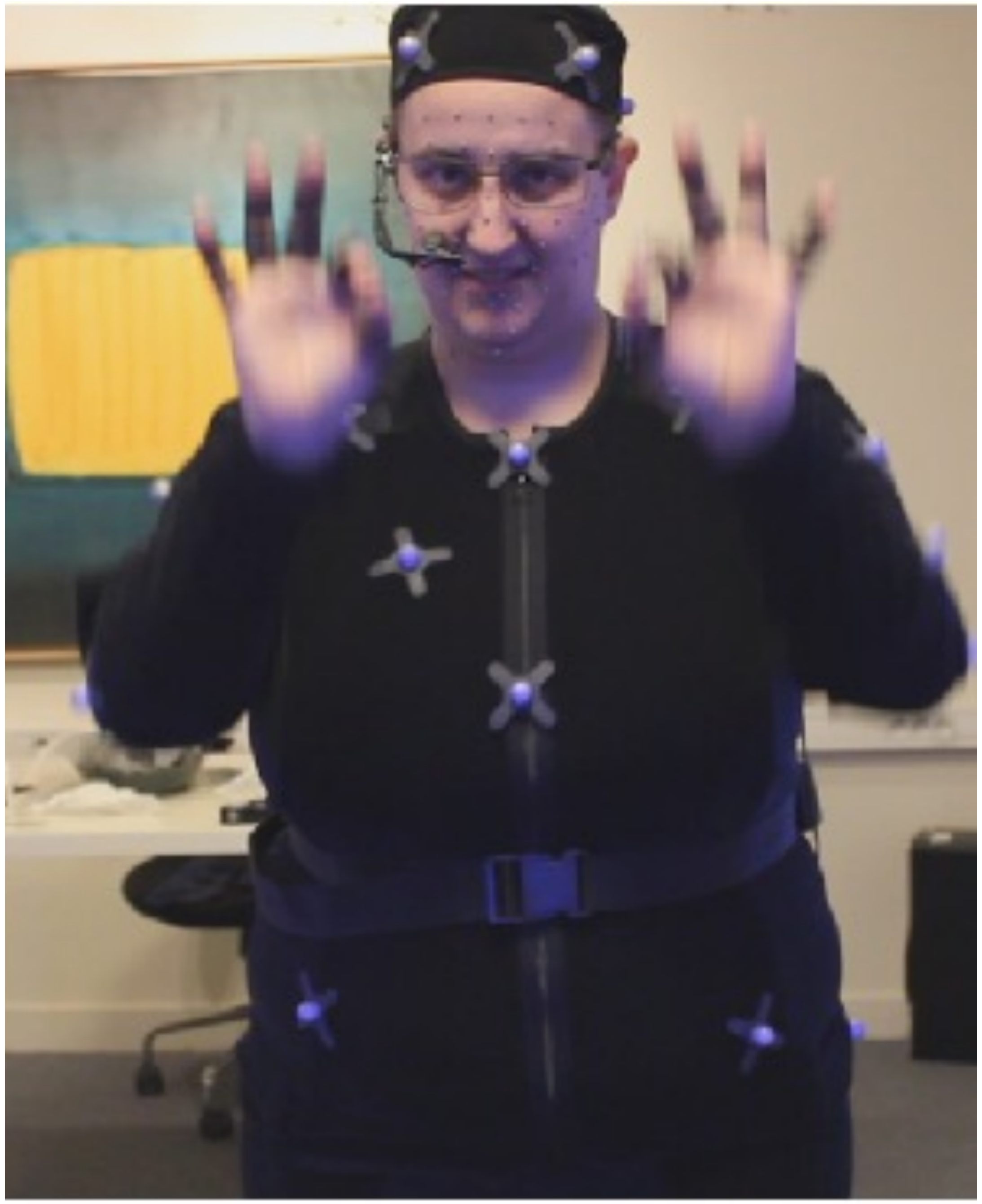
Pose Estimation Research 2

Complexity of Telicity Marking in TİD

Aysemin Yaşar
Bahadır Kisbet
Kadir Gökgöz

Introduction

Telic	Atelic
Semantically having an inherent end-point / goal	Semantically lacking an inherent end-point, perceived as ongoing
Bounded in nature	Unbounded in nature
Have the potential to reach a final state / completion	Lack the potential for completion, have the potential to continue indefinitely
Heterogeneous internal structure	Homogeneous internal structure
Phonologically have an overt end-marker	Phonologically not end-marked



Gibet (2018)

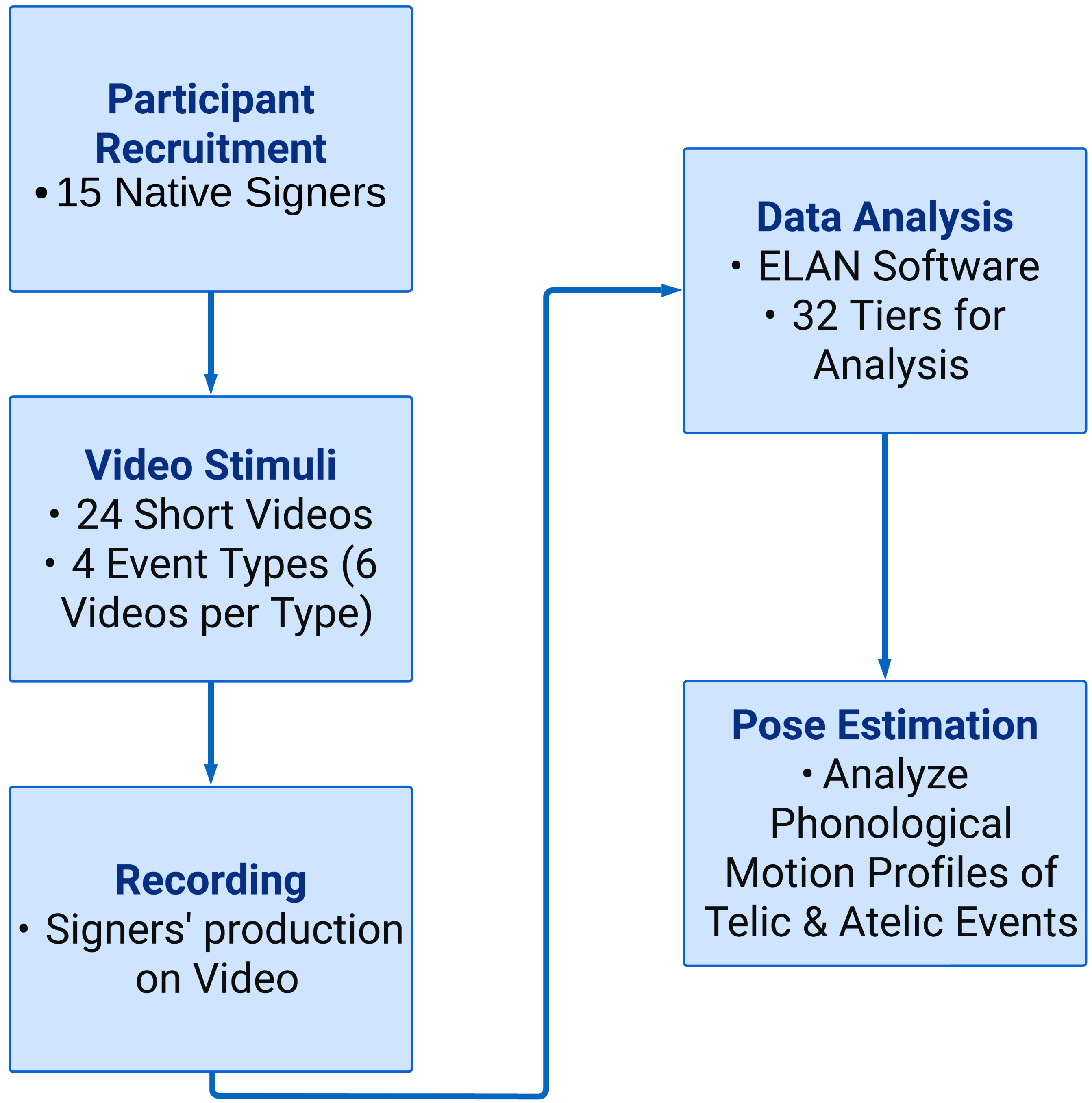
Kinematic Parameters: A Quantative Analysis

- Wilbur and her colleagues (2012) employed Motion Capture to calculate velocity values of event predicates
- Steeper slope of deceleration in telic events in ASL

However:

- Not cost-effective
- Signers do not really like using motion capture tools
- You cannot work with many people at the same time

Methodology



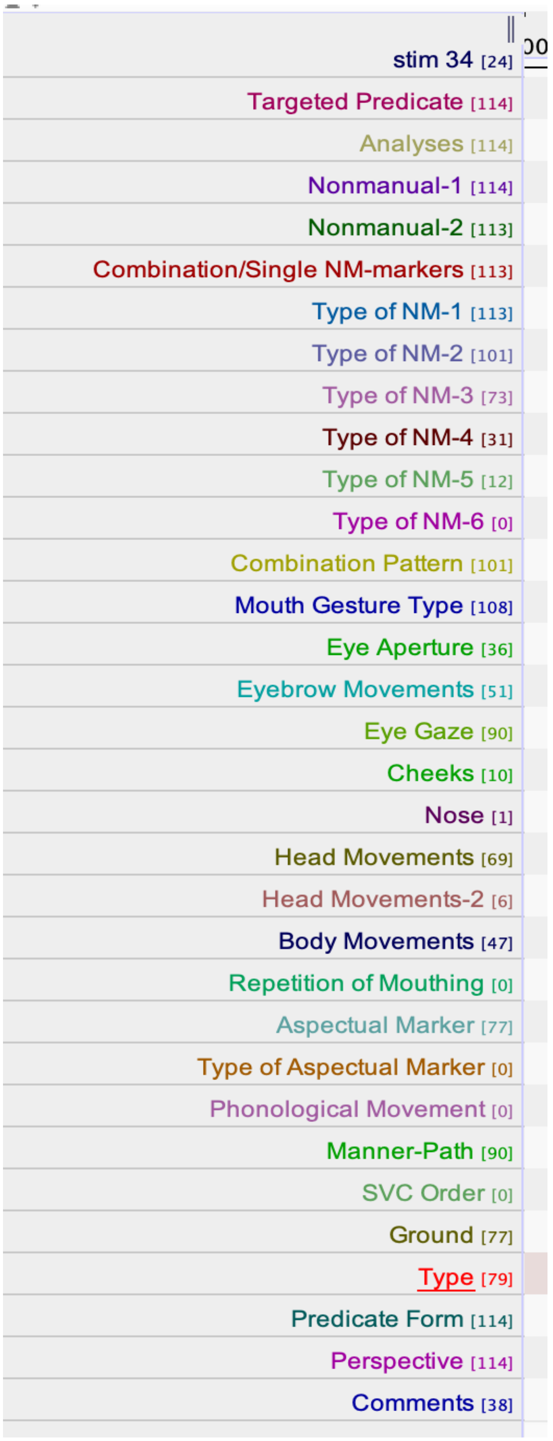
Methodology

- Our research draws on data from experiments conducted by our Sign Language Lab, where 15 native signers and 15 late signers viewed 24 short videos of various event types and provided productions, recorded on video

- Event predicates are detected and annotated along with their phonological specifications for movement, non-manuals, predicate forms and as well as accompanying aspectual markers if present

- A total of **984 event predicates** (Native signers only)

- Substantial distributional evidence which shows the multidimensionality of telicity marking.

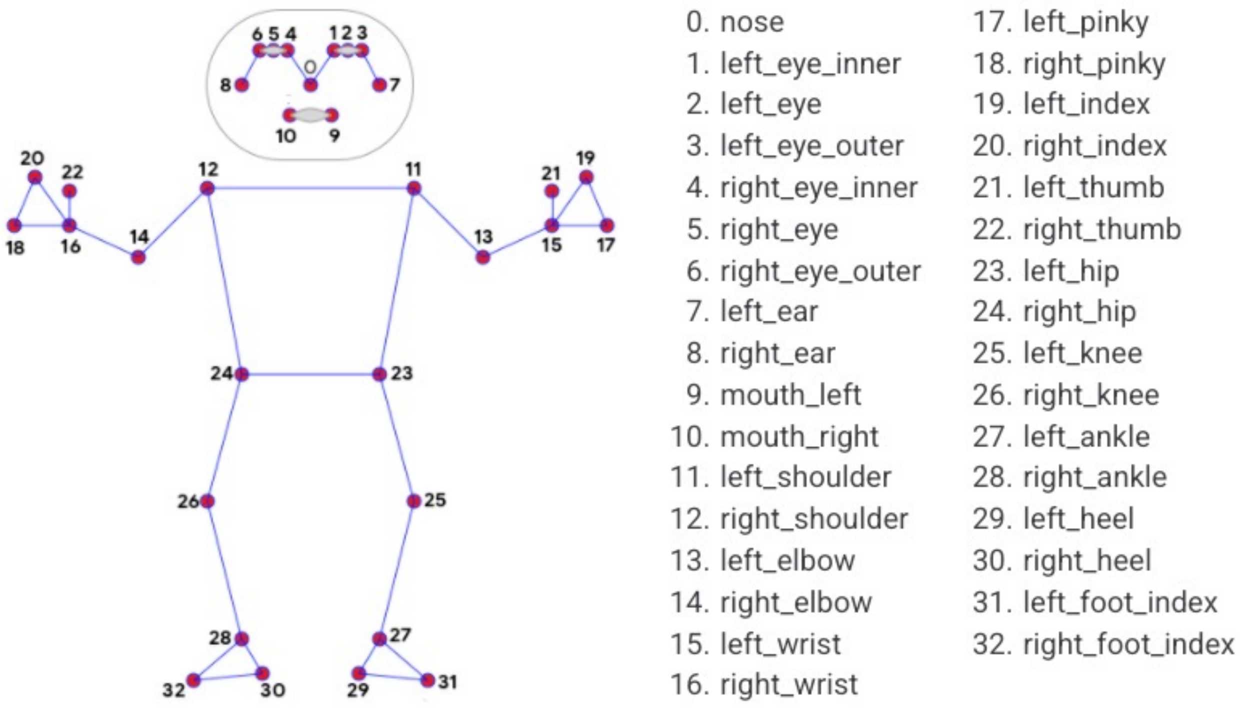


Analysis

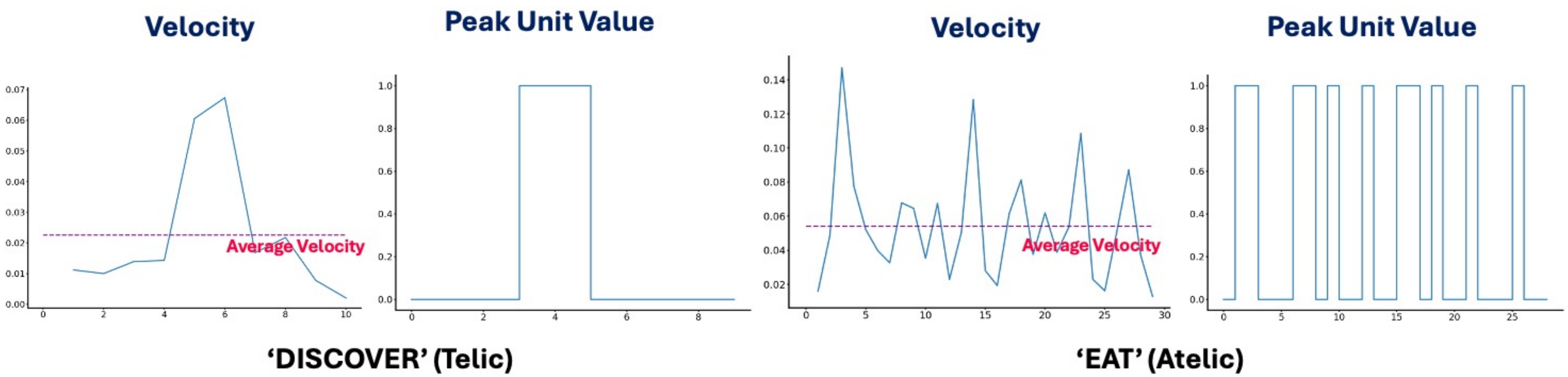
- A novel, cost-effective approach as an alternative for calculating similar values

- Initially computed the spatial positions of hand joints by tracking their movement along the x and y axes

- Relevant derivative values, including average velocity, peak velocity, and duration



Analysis



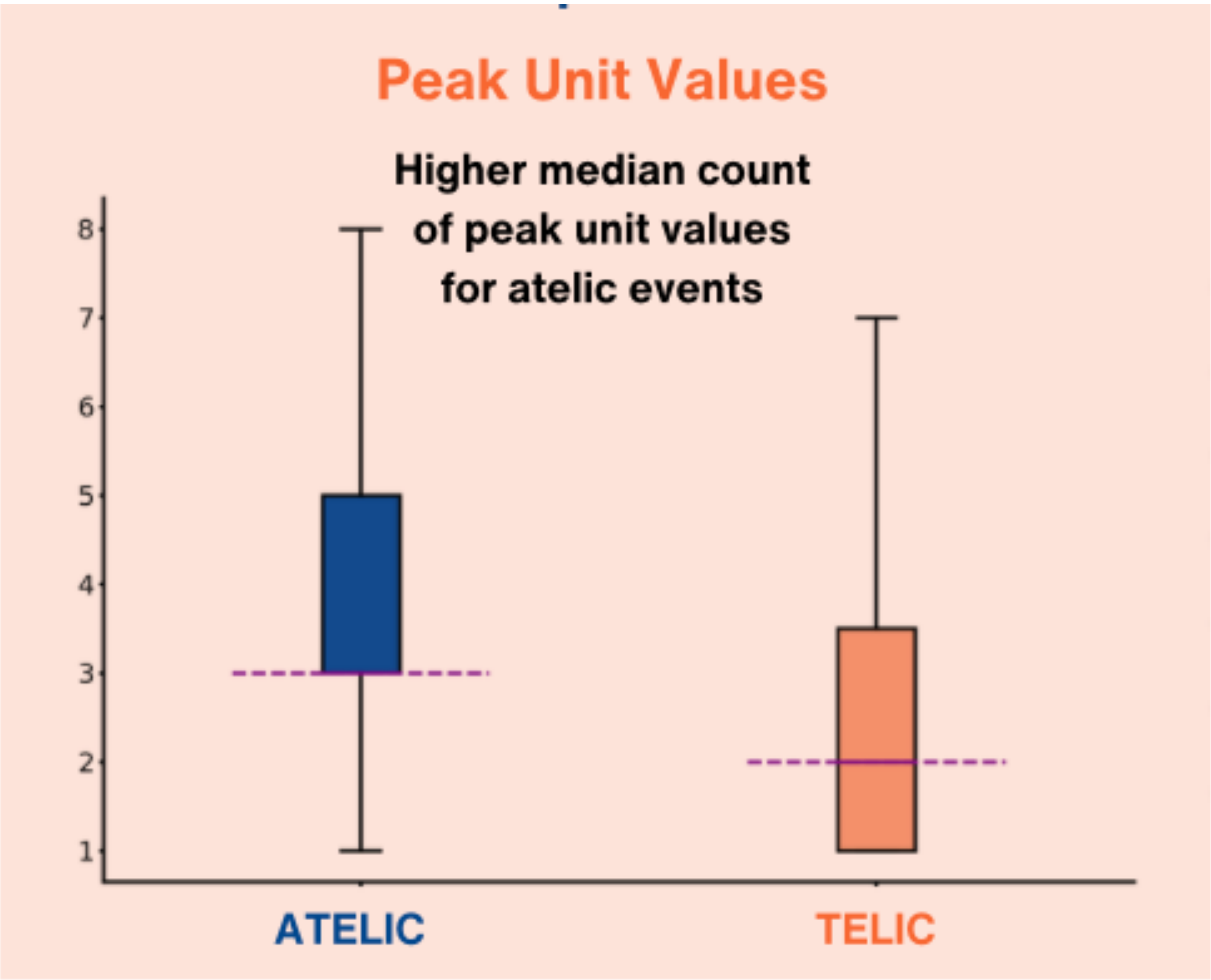
- The velocity graphs of events illustrated a **distinction in motion profiles**

- Calculating average velocity for each sign and subsequently determining the frequency of peak value units forming above the average velocity line for each group

- Higher median count of peak unit values for atelic events around the average velocity line, indicating **a harmonic motion pattern that fluctuates around the average velocity line**

- In contrast, **telic events displayed non- harmonic motion** with typically one or two peaks above the average velocity line

Distinct phonological motion profiles



The results seem to validate our initial impressionistic observations that these two event types do have distinct phonological motion profiles.

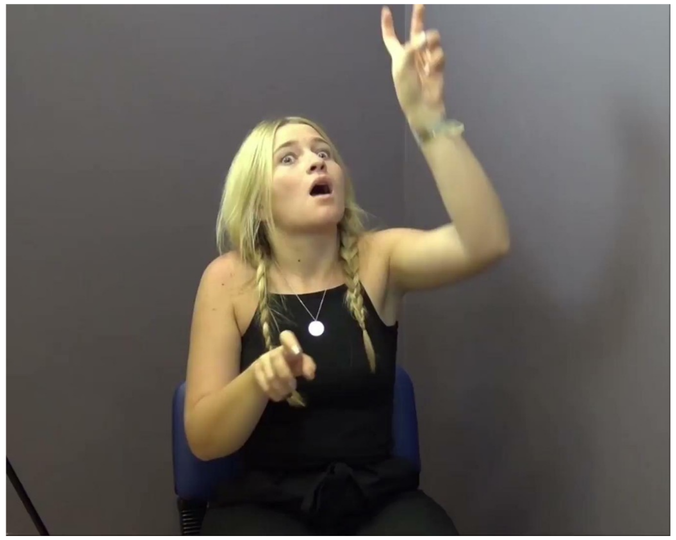
Getting to the Point: Deciphering the Linguistic Multifunctionality of Pointing in TİD

Ece Eroğlu
Kadir Gökgöz

Categorization

Locative

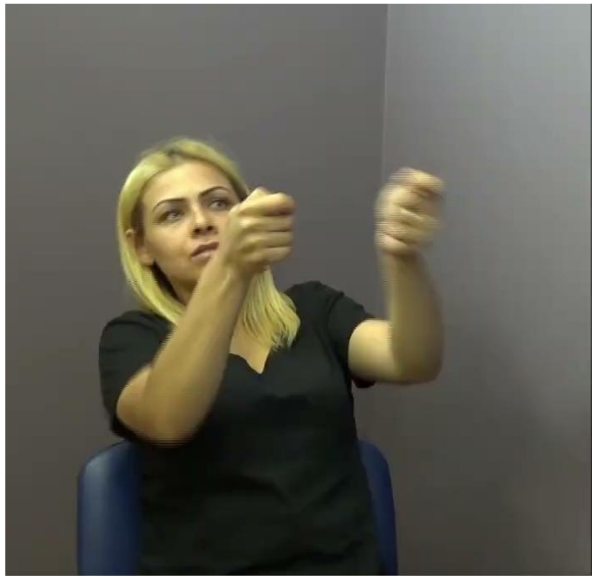
Pronoun



Demonstrative

Weak
Demonstrative
Clitic

Extra Examples



Locative

Pronoun

Demonstrative

Weak Demonstrative Clitic

Categorization

Possible Realizations:

	_	_V V_	_N N_
Corresponding to:	LOCATIVE	PRONOUN	DEMONSTRATIVE&CLITIC

- The most frequent distribution is _N. Pointing signs are interpreted to be a demonstrative when they occur preceding or following a noun. When they occur before, after or in between two Verbs

- In some cases, we see _(Modifier) N (Modifier)_ ; that is the pointing sign is observed at the beginning and the end of the NP.

  We see this pattern more frequently when the Noun is modified by an Adjective. (working hypothesis)

  This might be marking the domain of the NP. ‘sandwich’LOOK

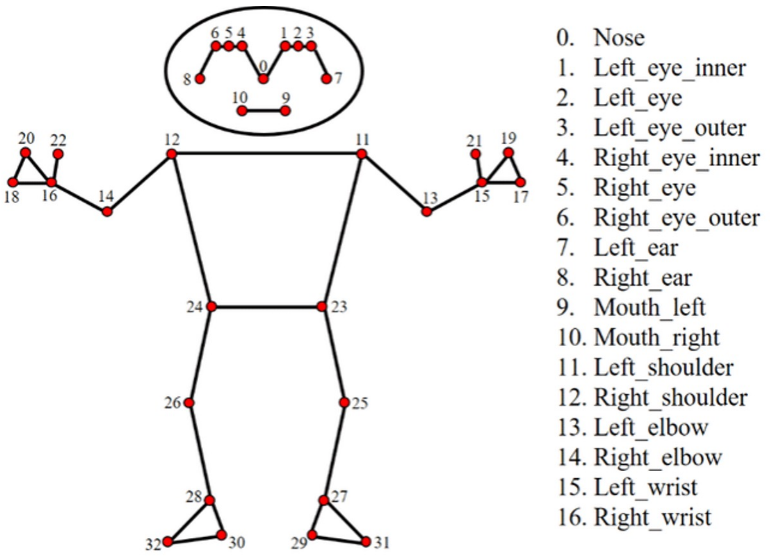
Analysis

Initial categorization based on observation is followed by validating the categories with Pose Estimation.

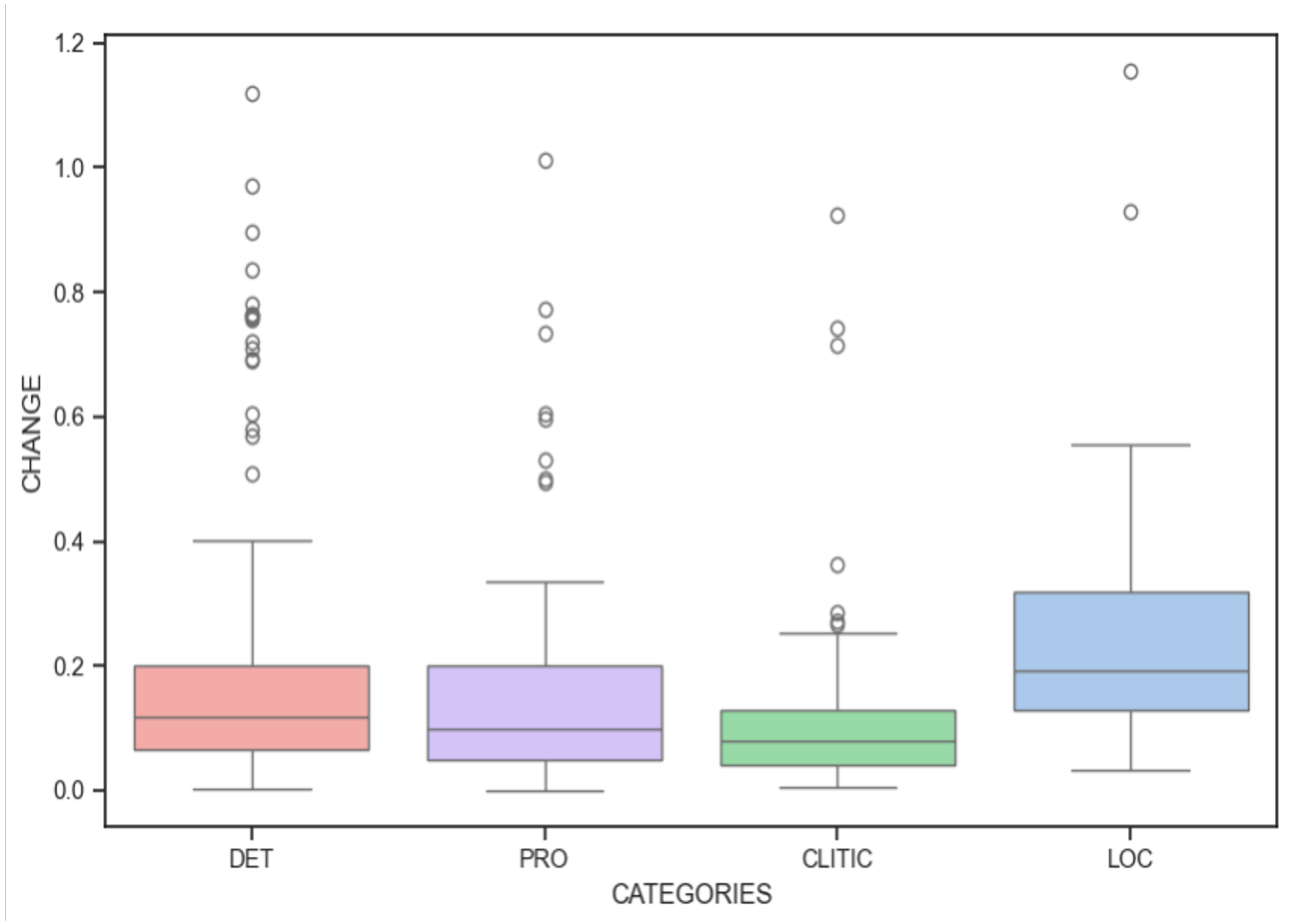
A blend of **qualitative and quantitative method**.

Q: Whether what we have seen as ‘distinctions’ could be mapped onto numbers.

With the help of Pose Estimation, we have transformed these pointing signs into solid numeric ranges upon which hypotheses can be built.



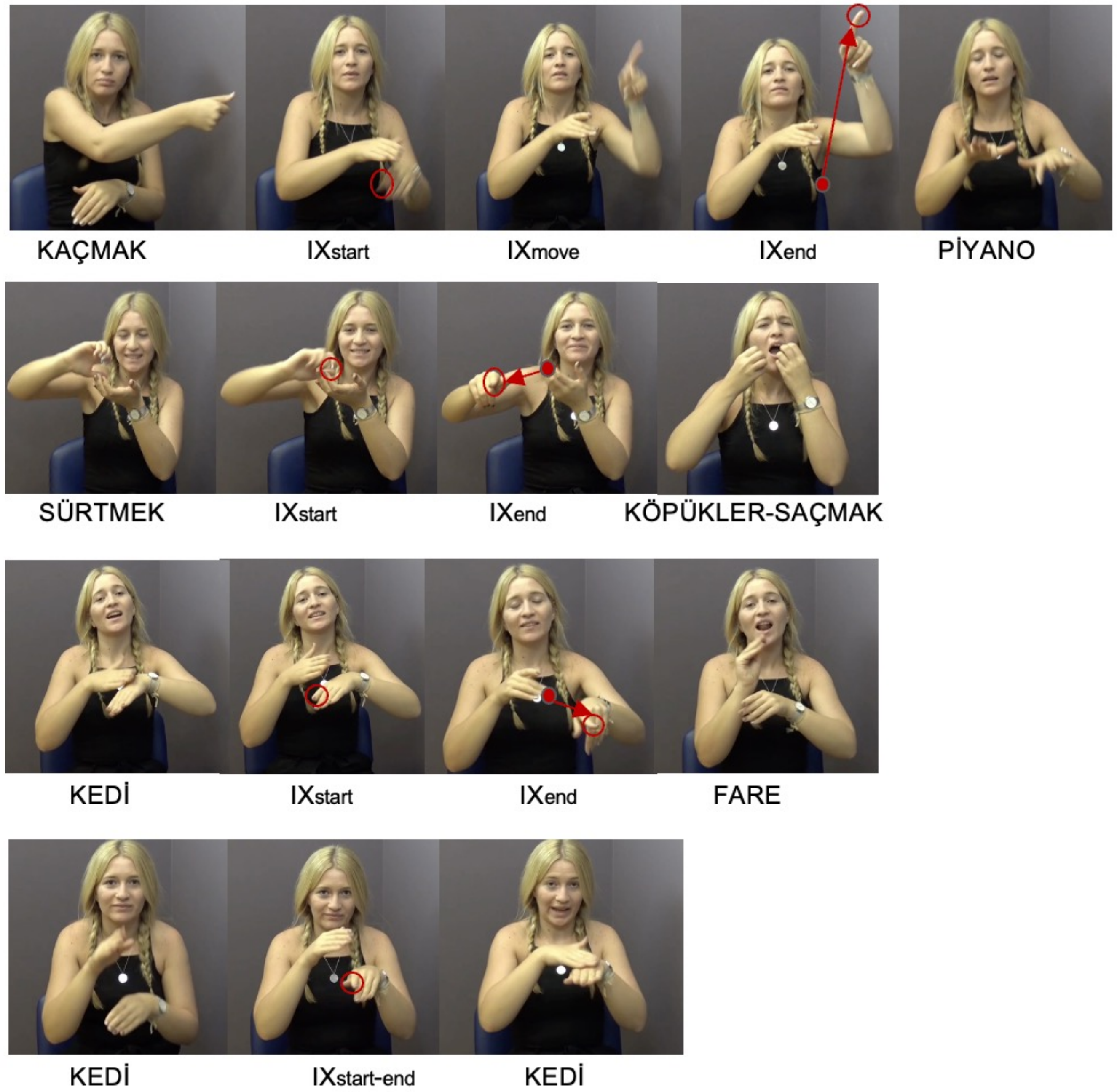
Quantifying Phonology: The Metrics of Pointing Signs

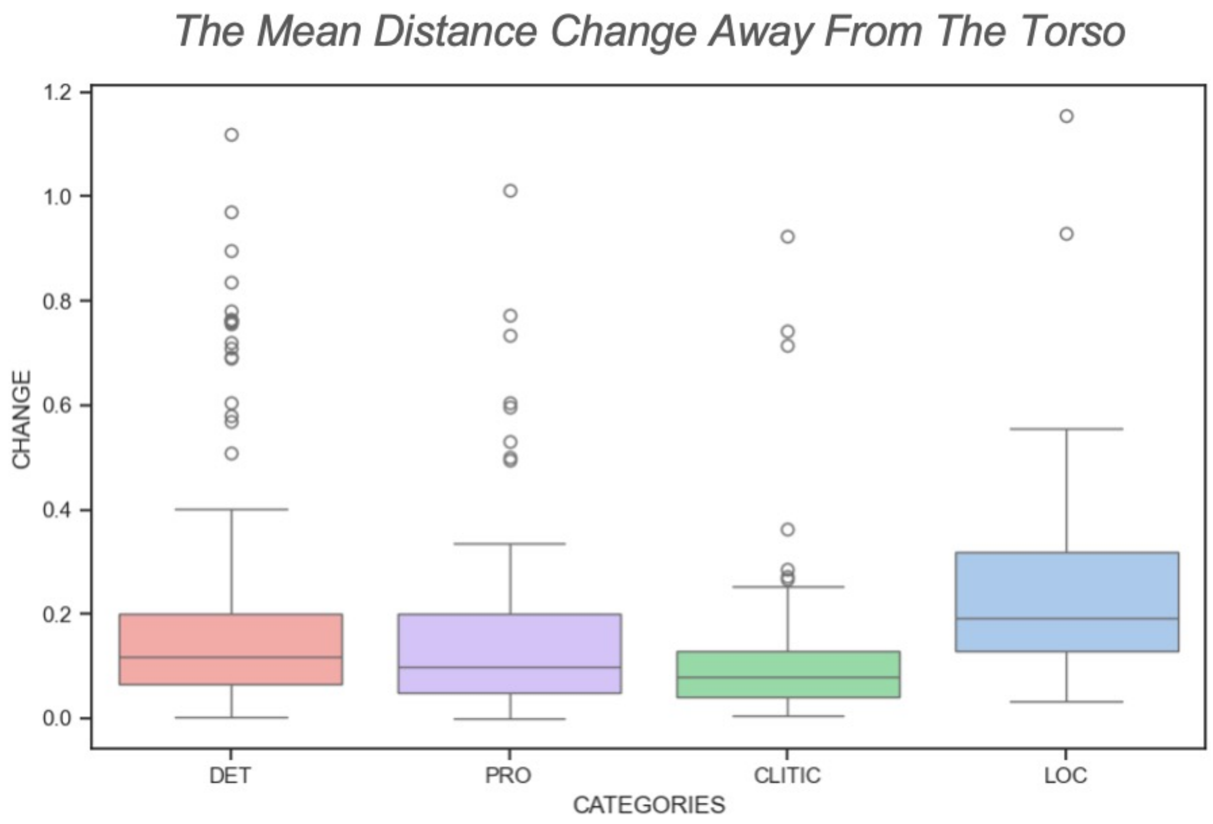


381 pointing signs in total (Free-Demonstratives, 228; Clitic-Demonstratives, 73; Pronouns, 54; Locatives, 26)

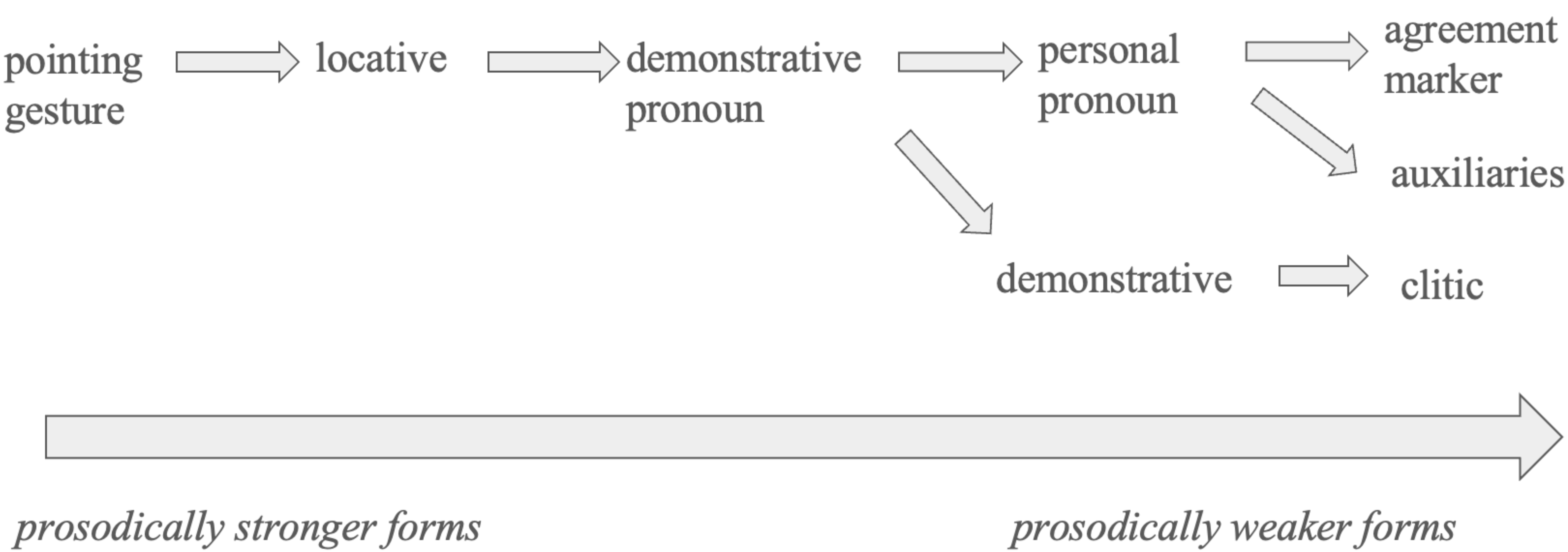
- The concern regarding terminology is irrelevant.
- Validates our initial observations showing that these signs are distinctly characterized by considerable variance.

Quantifying Phonology: The Metrics of Pointing Signs





Discussion


pointing gesture
locative
demonstrative pronoun
personal pronoun
agreement marker
auxiliaries
demonstrative
clitic
prosodically stronger forms
prosodically weaker forms

About Late Signers

Prediction:

As TİD is very actively making use of locatives, pronouns, and demonstratives.

In my opinion, these classes' frequencies might have more to say for the late signers' adoption of the system.

Late signers might be making use of all classes, but maybe with a lower rate of clitics.

They might not be using clitics at all.

Results (ongoing):

It seems that they are more actively using clitics than any other class!

No important statistical difference between the distributions.

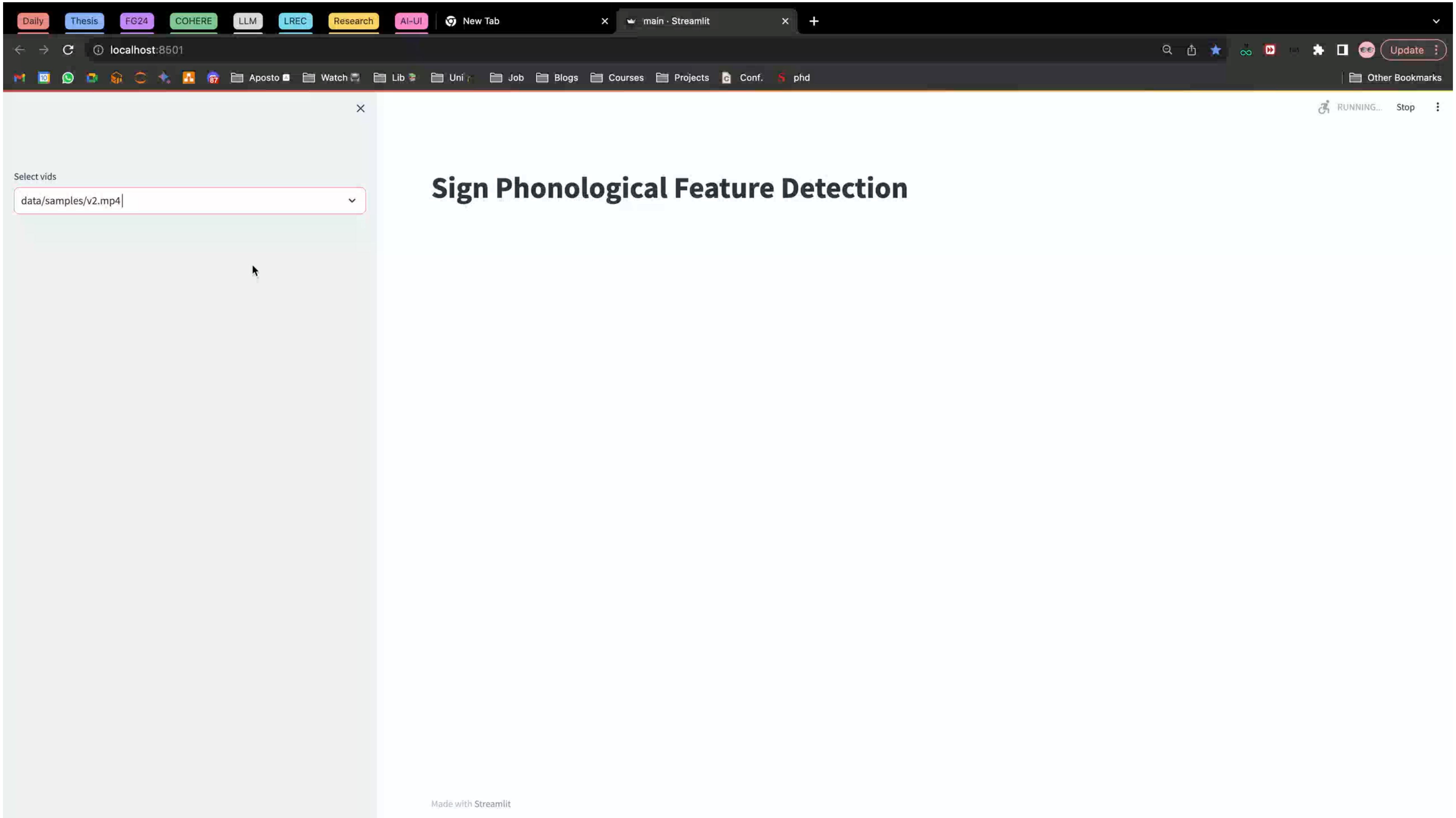
However, they seem to be making use of a spatial strategy, namely LOCUS less.

Prediction: They might be utilizing space less efficiently!

Use of Transformers in Sign Language

Many thanks to Karahan Şahin

Check out Karahan's Computational SL projects



Sign-LLM: All experiments regarding continuous sign language translation

sign-llm-base

All experiments regarding continuous sign language translation

List of experiments for SLR-SLT system

1. Dataset(s):

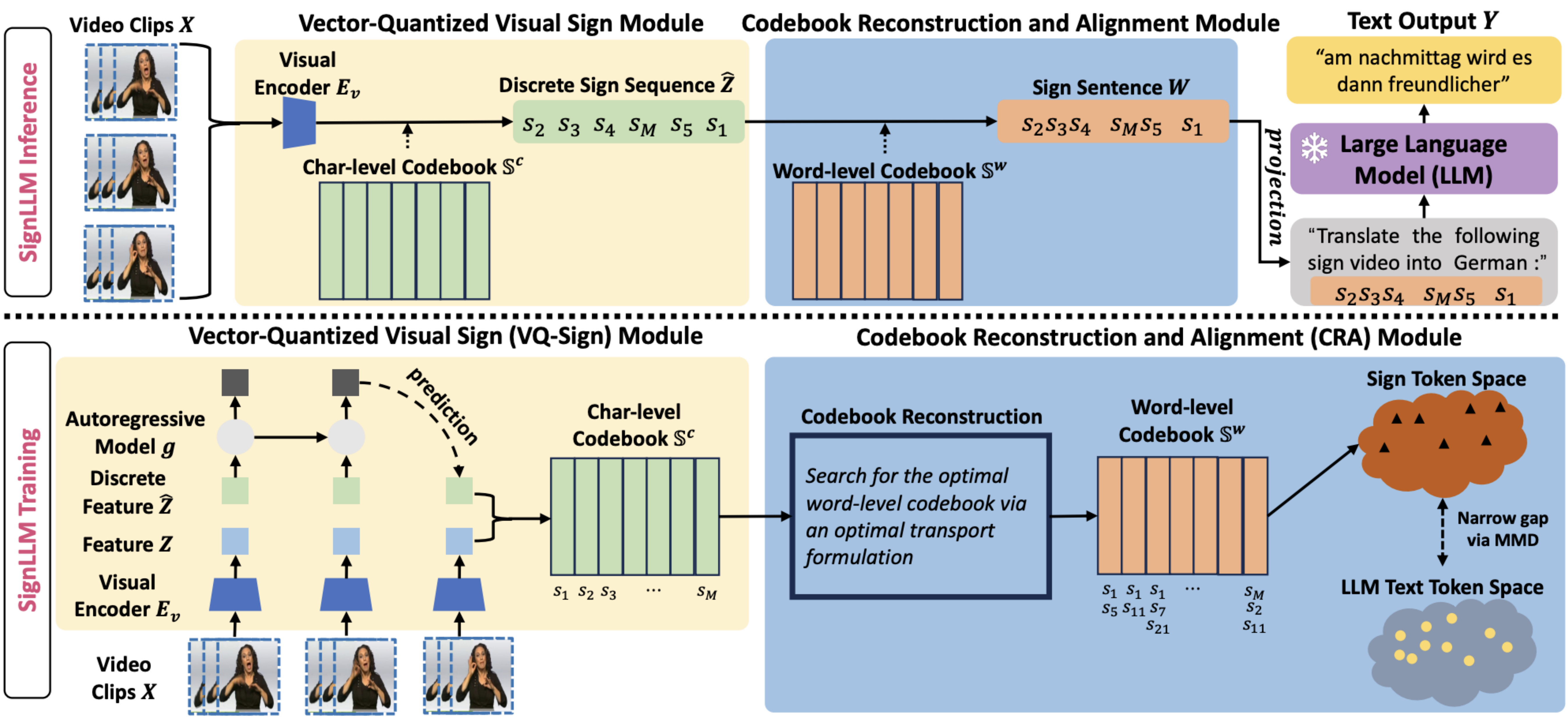
These are the datasets for experimentation. The continupus

Dataset Name	Description	Types	Status	Publish Status
TIDSözlük	Toy Dataset	Isolated/Cont.	Vid + Pose	-Publish (-SLR) (-SLT)
BSign22k	TID Benchmark 0	Isolated	Vid + Pose	+Publish (+SLR) (+SLT)
AUTSL	TID Benchmark 1	Isolated	Vid + Pose	+Publish (+SLR) (+SLT)
PhoenixWeather	DGS	Continuous	Vid + Pose	+Publish (+SLR) (+SLT)
Content4All	DGS / Swiss	Continuous	Vid + Pose	+Publish (+SLR) (+SLT)
DGS Corpus	DGS	Continuous	Vid + Pose	+Publish (-SLR) (-SLT)
SEBEDER	TID	Continuous	Vid	-Publish (+SLR) (+SLT)

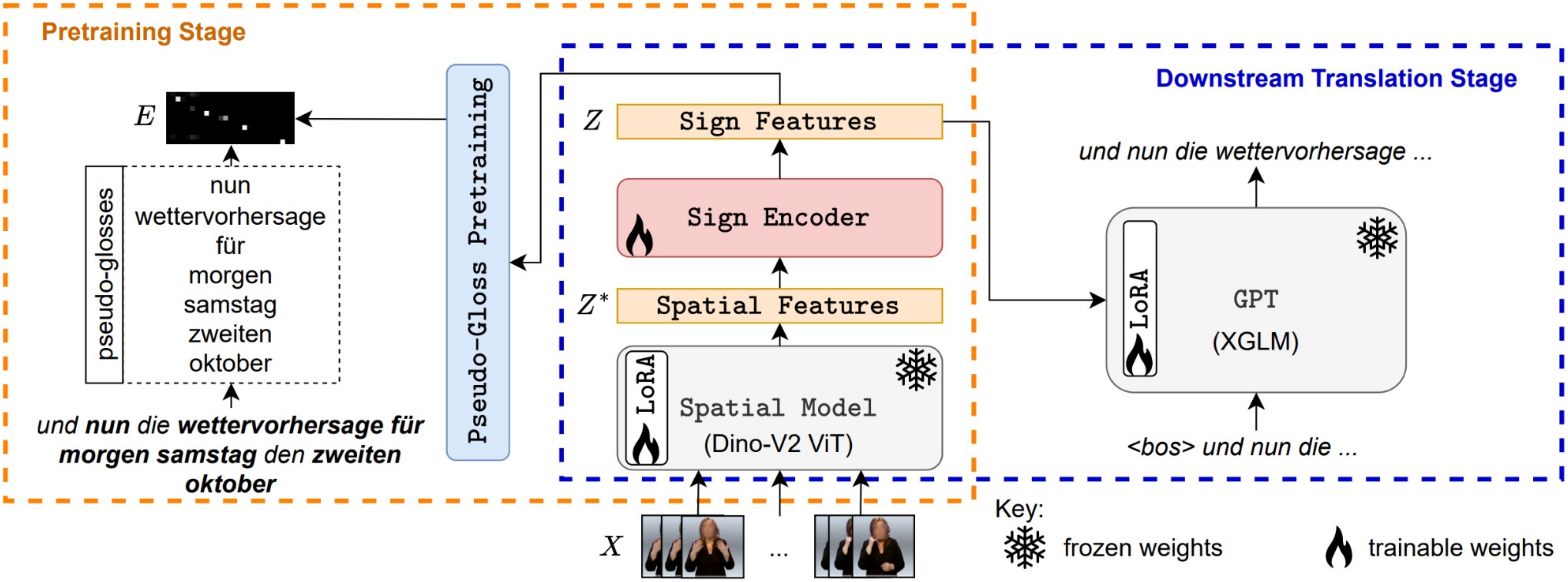
Follow Karahan Şahin on Github

Karahan’s Github Page:

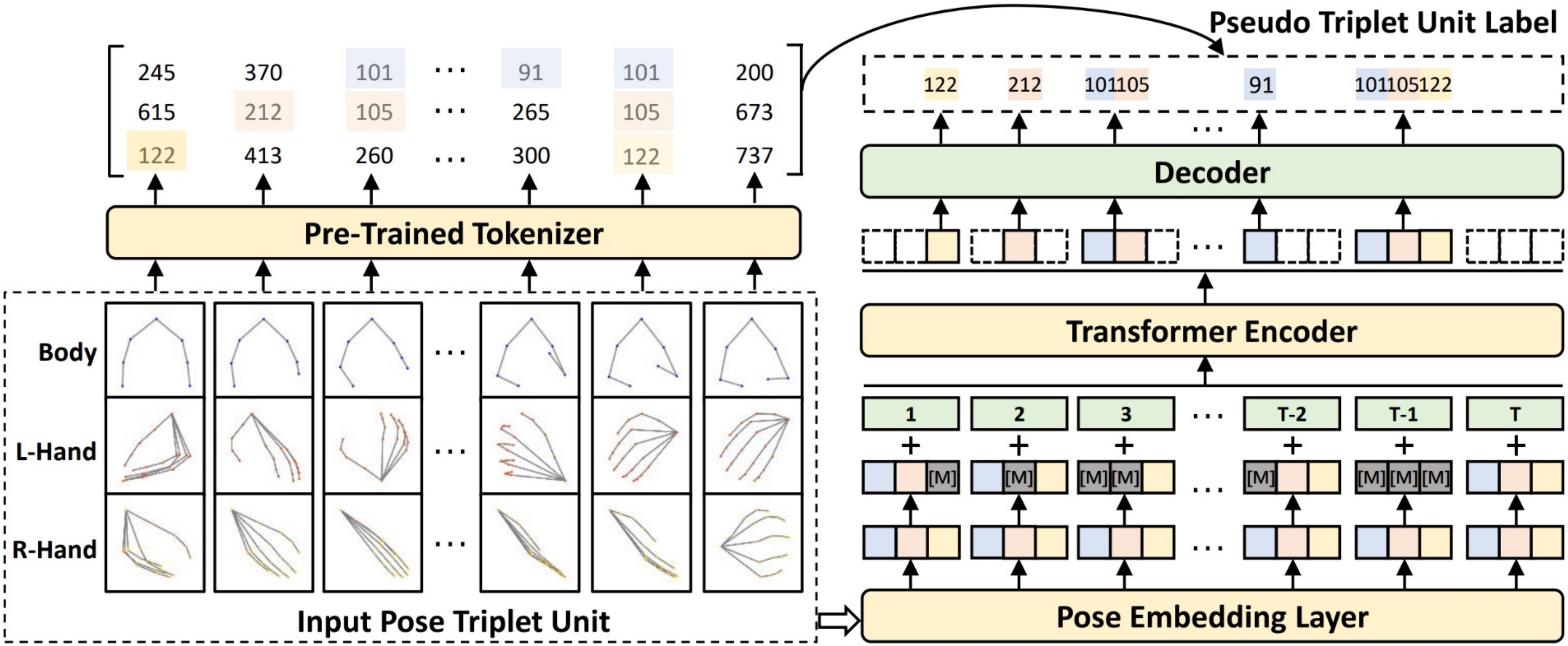
LLMs are Good Sign Language Translators



SIGN2GPT: Leveraging Large Language Models for Gloss-Free Sign Language Translation



BEST: BERT Pre-Training for Sign Language Recognition with Coupling Tokenization



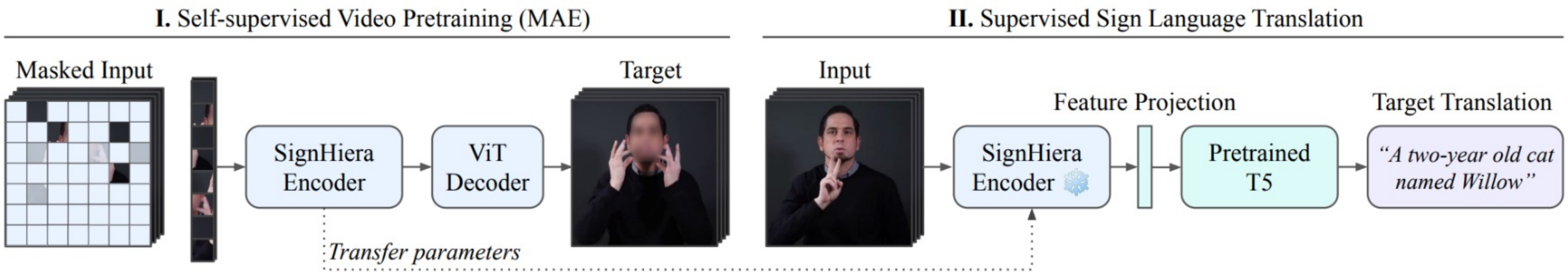


Figure 1: Overview of our two-stage SSVP-SLT method. The first stage consists of training a SignHiera encoder via masked autoencoding (MAE) on *blurred* video frames. In the second stage, a pretrained T5 model is finetuned for SLT while the pretrained SignHiera is kept frozen (❄️). The input video in the second stage *can be unblurred*.

What is next?

<u>ROADMAP</u>

1. Transform Sign Language Videos as **Structurally-Aware pseudo-words**

2. Apply Large Language Modeling Objectives

3. Map to SignLLMs with SpokenLLMs

THANK YOU!

Any questions or comments?

CONNECT WITH ME!