A glimpse into sign language corpus linguistics

the WHAT, the HOW & the WHY

Carl (Calle) Börstell

University of Bergen

LINNORD

Nordic Innovations in Linguistics: Multi-method approaches for understanding language

Jyväskylä, March 20–21, 2025

Me

My research

Probably **99%** of my research is mainly on **signed languages**

Early: **typologically**-oriented **descriptive** linguistics

Then: **iconic patterns** across languages and modalities

Now: distributional patterns in **corpora + computer vision**

I think ~80% is based on corpus data in some form

See more: borstell.github.io

Today’s presentation

Today's presentation topics

What is corpus linguistics in sign language research?

What research **questions can we answer** with current data?

How do we **extract and process** the data with these goals in mind?

Why do we want to **use corpus data** – or not?

Introduction

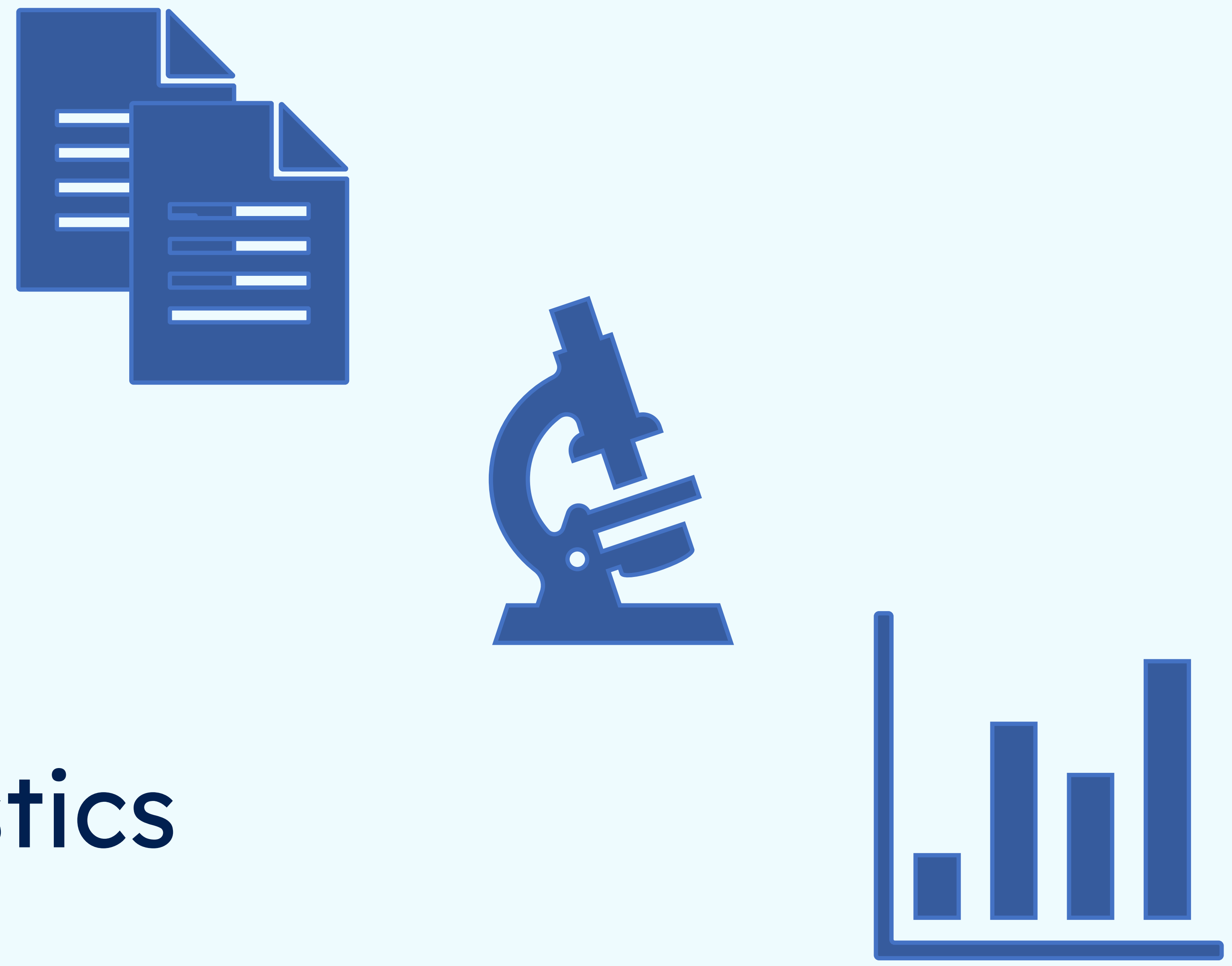
The Quantitative Turn

In the past 25 years, **linguistics** has become more **quantitative**

More **data**: corpora

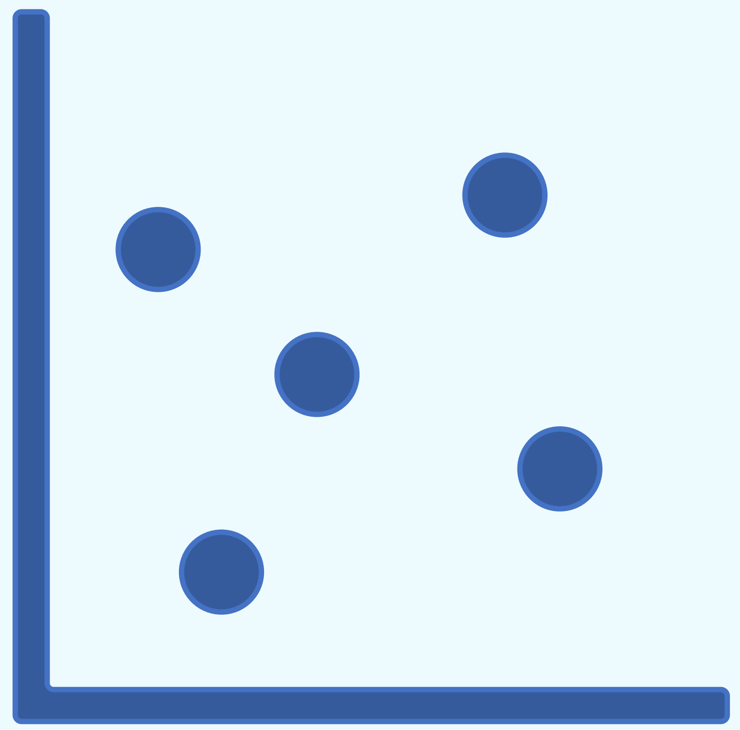
More **control**: experiments

More **complex methods**: statistics



Joseph (2008); Kortmann (2021)

Quantitative sign linguistics



The field **has followed this general trend**

However…

… still **small datasets** (few **participants**) also in experiments (small pools)

… **SL corpora** have been **built from scratch** (technically challenging)

Corpus-based or corpus-driven

Corpus-based research

– corpus is used to test and **verify theoretical claims**

Corpus-driven research

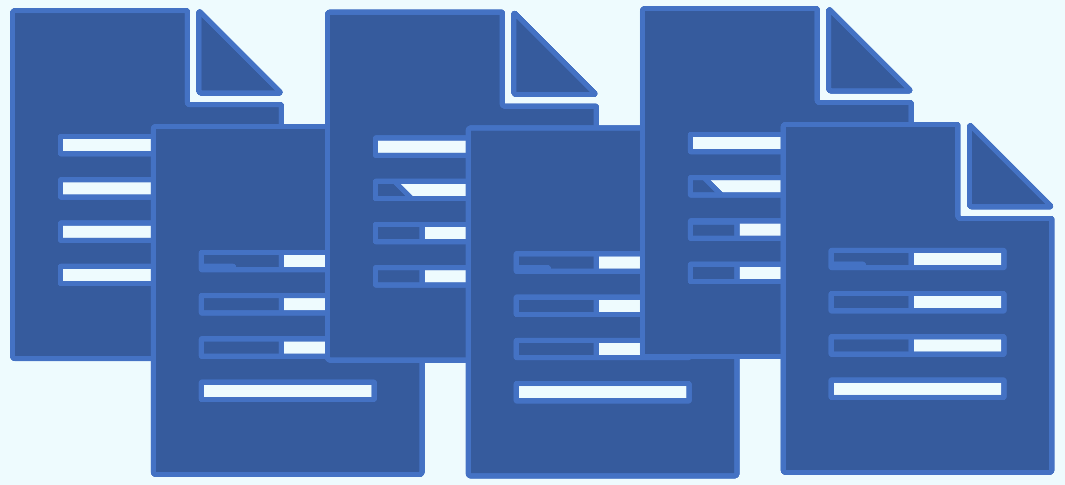
– corpus itself is the **source** from which **patterns are identified**

SL work has been mostly corpus-based, but some corpus-driven?

– corpus size, balance and representativeness are issues here

Tognini-Bonelli (2001: 17); see also O'Keefe (2018) and Aijmer (2018)

What is a corpus?



Data(base) ≠ corpus

A **collection** of natural(istic) **texts**

Machine-readable transcriptions: written, spoken or signed

Should be **large enough**; can target a specific purpose
… anything <1 million words is often seen as small

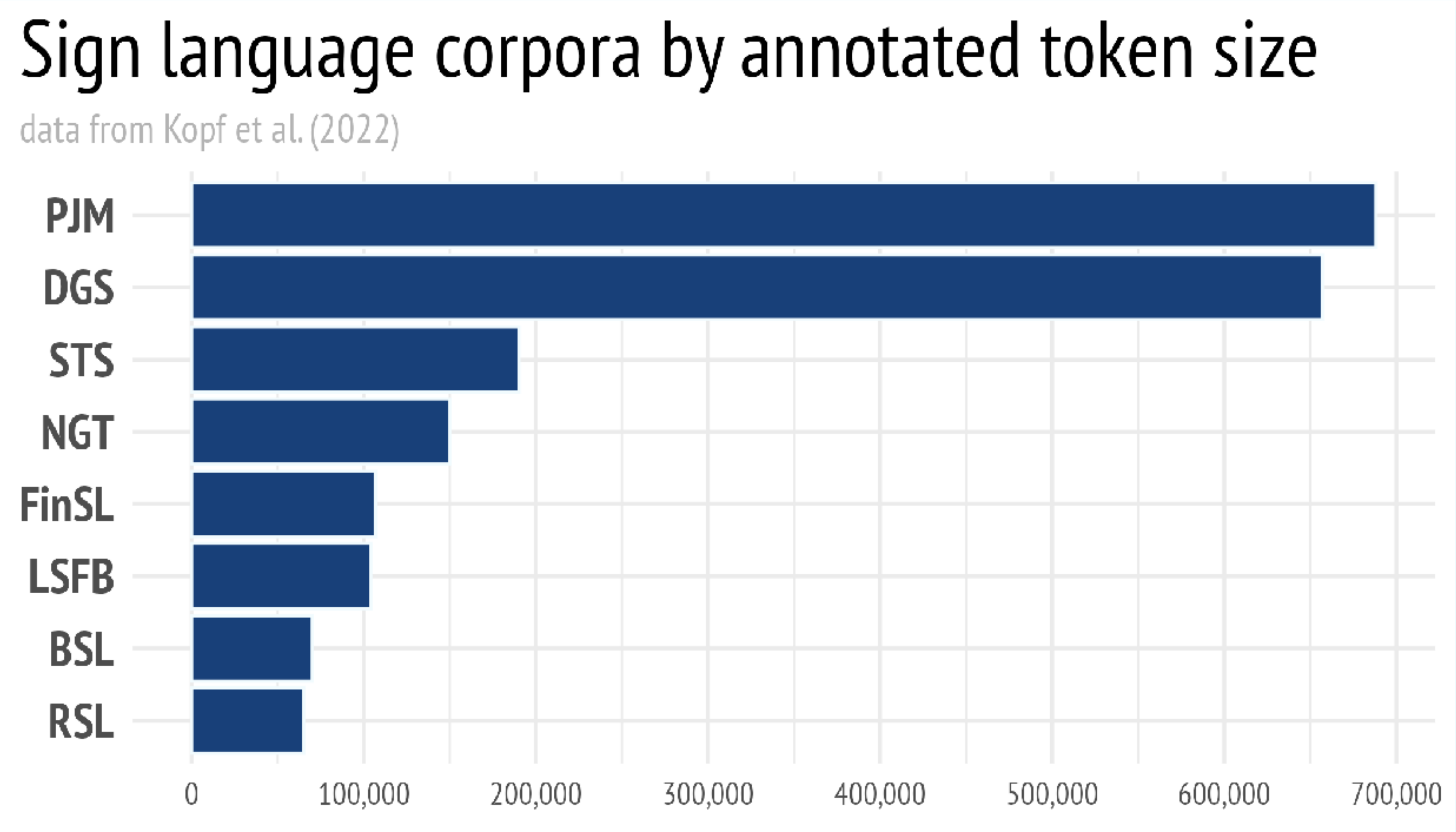
Biber et al. (1998); McEnery & Hardie (2011); Kennedy (2014)

Sign language corpora today

Only two sign language corpora that are equivalent in size to (smaller) spoken language corpora:
Polish SL (PJM) and German SL (DGS)

What does this mean for research?

Kopf et al. (2022)

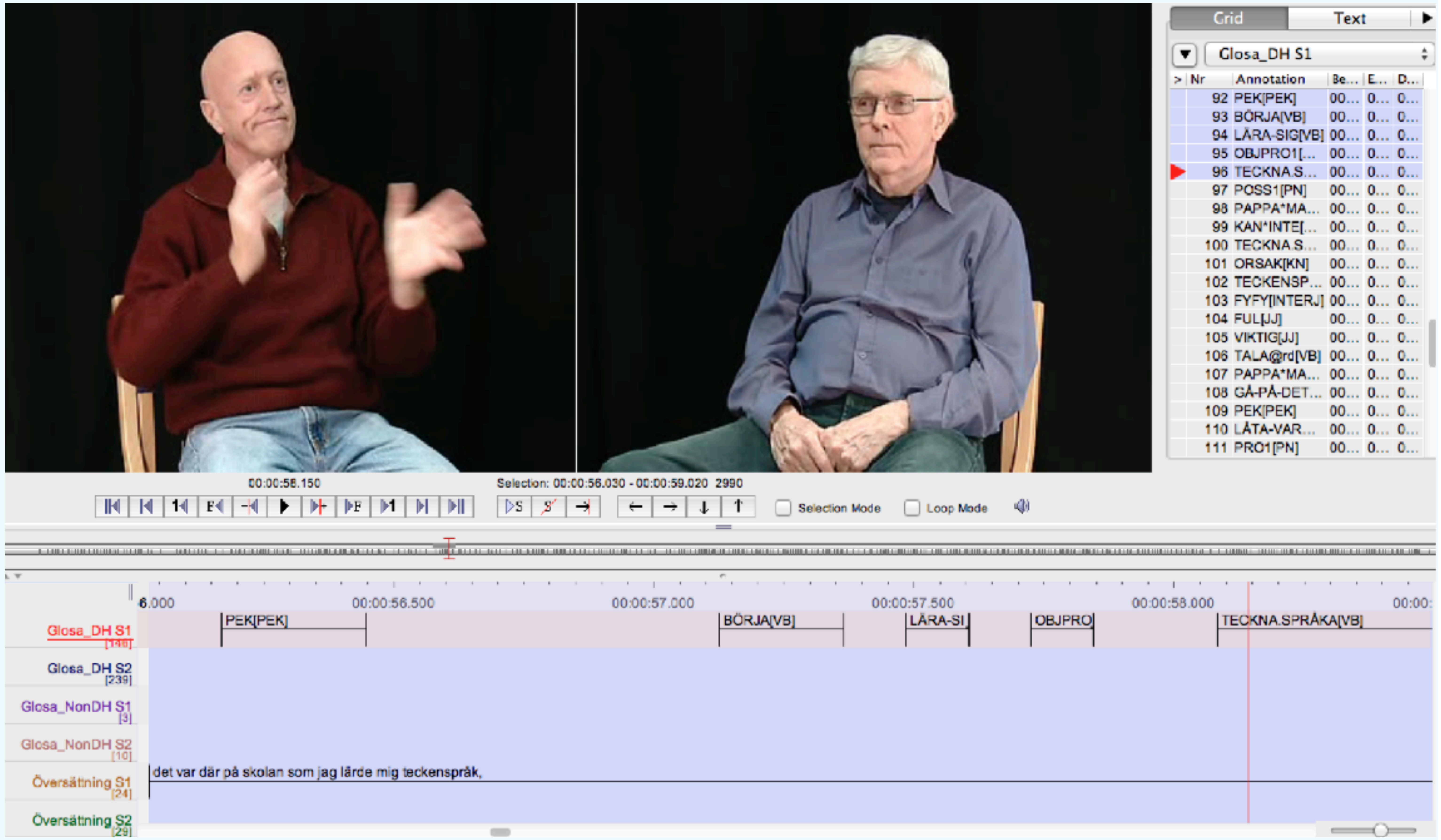


Sign language corpora

Most sign language corpora use **ELAN for annotation**

ELAN is also the main tool **for viewing and searching** the corpus

Basic interface and annotations: **signs + translations** (here STS)



Mesch et al. (2012); Öqvist et al. (2018)

the WHAT &
the HOW

What data do SL corpora have?

Lexical: Segmentation + gloss annotation of **individual signs**

Translations: written (or spoken) translations into a **spoken language**

Metadata: sociolinguistic profile of signers

Other: morphology (limited), **form**-descriptors (hs, #hands), **pose**

Johnston (2010; 2014); Fenlon & Hochgesang (2022) ; Kopf et al. (2022)

What data do SL corpora lack?

Morphology: Reference, modification, lemmas (cf. DGS)

Syntax: segmentation and dependencies (cf. Auslan & STS)

Discourse: interactional segmentation + functions

Tools: user-friendly search tools for frequency, concordances and context

Östling et al. (2017); Börstell (2022; 2024a,b); Fenlon & Hochgesang (2022); Kopf et al. (2022)

Frequency

Lexical frequency in sign languages

One of the **first and easiest** things to look at in a corpus

Some of the **earliest papers** used a “corpus”
- **ASL** data was 4,000 sign tokens
- **NZSL** used text transcriptions (but: 100,000 tokens)

With first “real” SL corpora, **lexical frequencies** were investigated
- **text type** influences sign types: **depiction** more in **narratives**

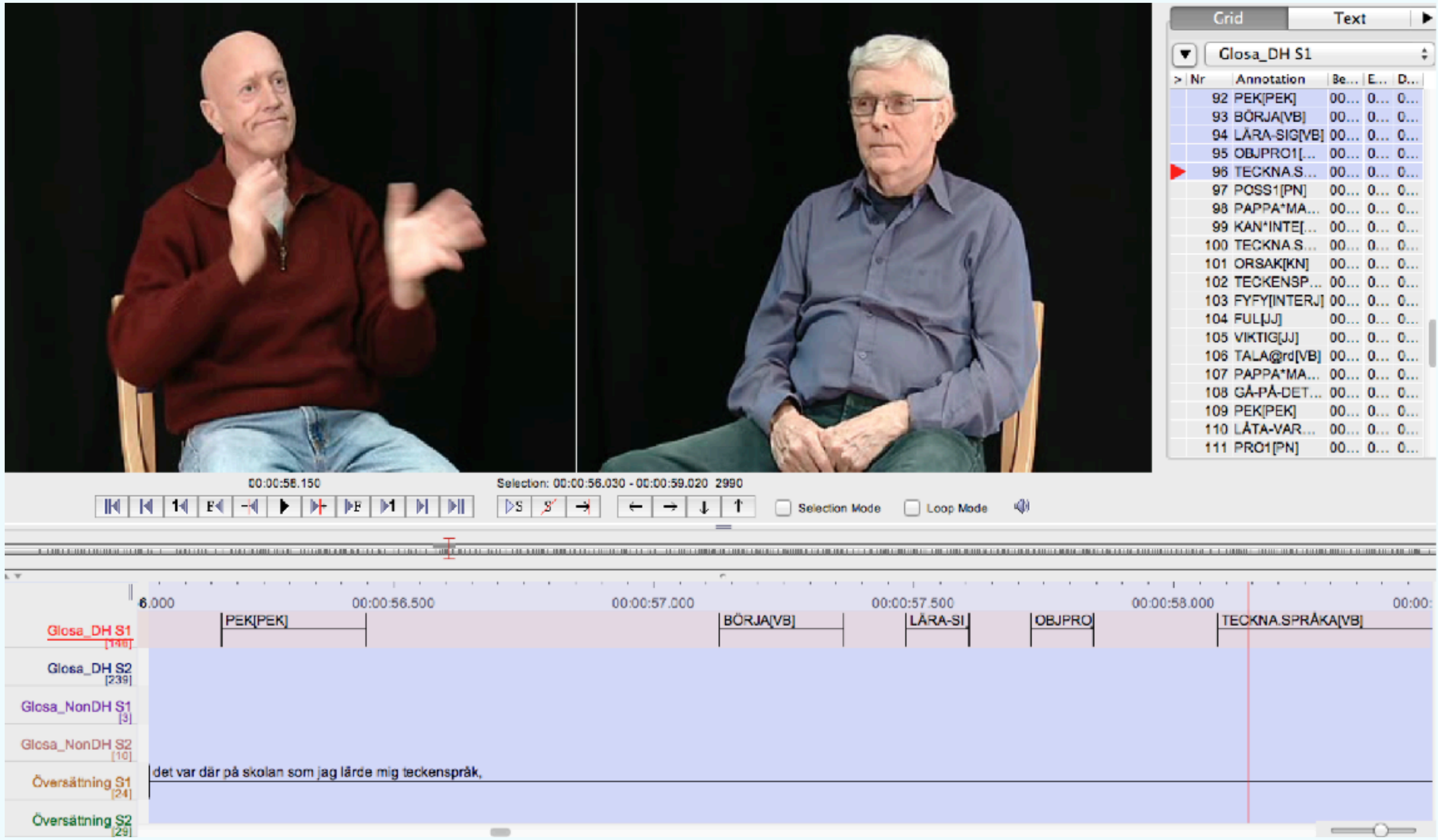
Morford & MacFarlane (2003); McKee & Kennedy (2006); Johnston (2012); Fenlon et al. (2014); Börstell et al. (2016)

Lexical frequency in STS

We (5th?) could compare **STS sign frequencies** with previous research

Similar frequency **patterns**:

- **functional** (points, etc.)
- **culturally relevant** concepts (e.g., DEAF)
- **sign** type ↔ **text** type



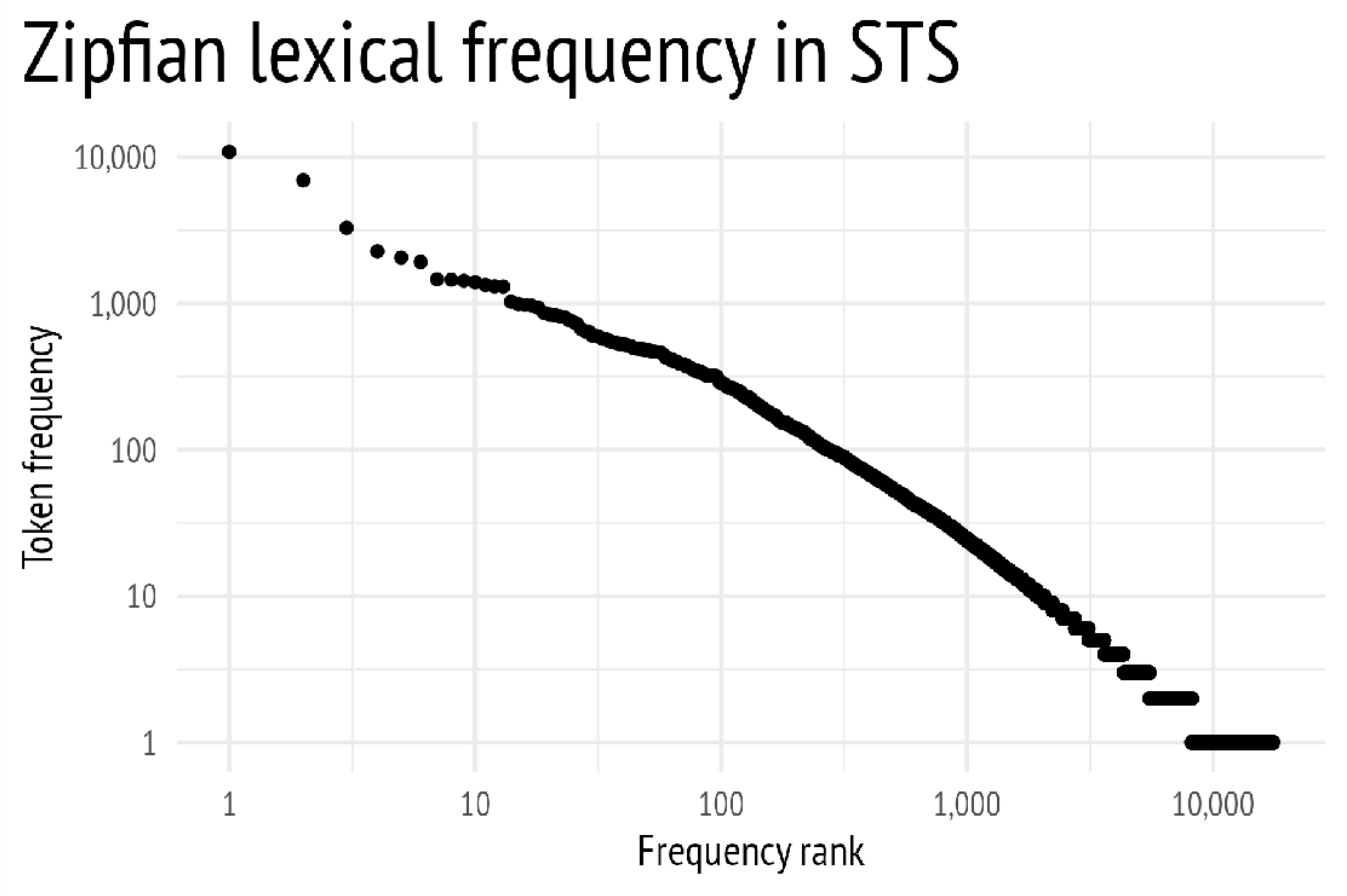
Börstell et al. (2016)

Lexical frequency in STS

We could also (1st?) show that STS lexical frequency showed a **Zipfian distribution** (expectedly)

The **frequency rank** (1st, 2nd, 3rd, …) is inversely **correlated** with the **token frequency** of signs

Börstell et al. (2016); Zipf (1935); see also Kimchi et al. (2023)



Word class in STS

In 2015, we **word class** (or part-of-speech) **annotated** the STS corpus

Often misunderstood, but our semi-automatic approach was **manually corrected** (~3,000 types)

– done on the type-level

Östling et al. (2015)

Word class in STS

With these word class annotations, we could also look at **frequencies on a "higher level"**

Expected patterns:

- **Content** classes have more types to tokens
- **Function** classes have more tokens to types

Börstell et al. (2016)

Category	Word class	Types	Tokens
content	noun	4,878	28,579
	verb	3,752	42,794
	adjective	692	8,211
	adverb	522	18,337
function	pronoun/point	446	31,190
	preposition	77	3,382
	conjunction	60	4,356
	article	5	7
other	depict/CA	2,263	9,383
	numeral	393	4,246
	gesture	130	5,154
	interjection	111	7,627
	buoy	70	2,602

Lexical variation

With **signer metadata**, we can also look at lexical variation

In practice: this is difficult!
- “small” corpora = **few relevant items**
- without targeted topics/themes, **content varies**

We have tried to **identify candidates** from the data, but it is hard!
- a combined approach may be best

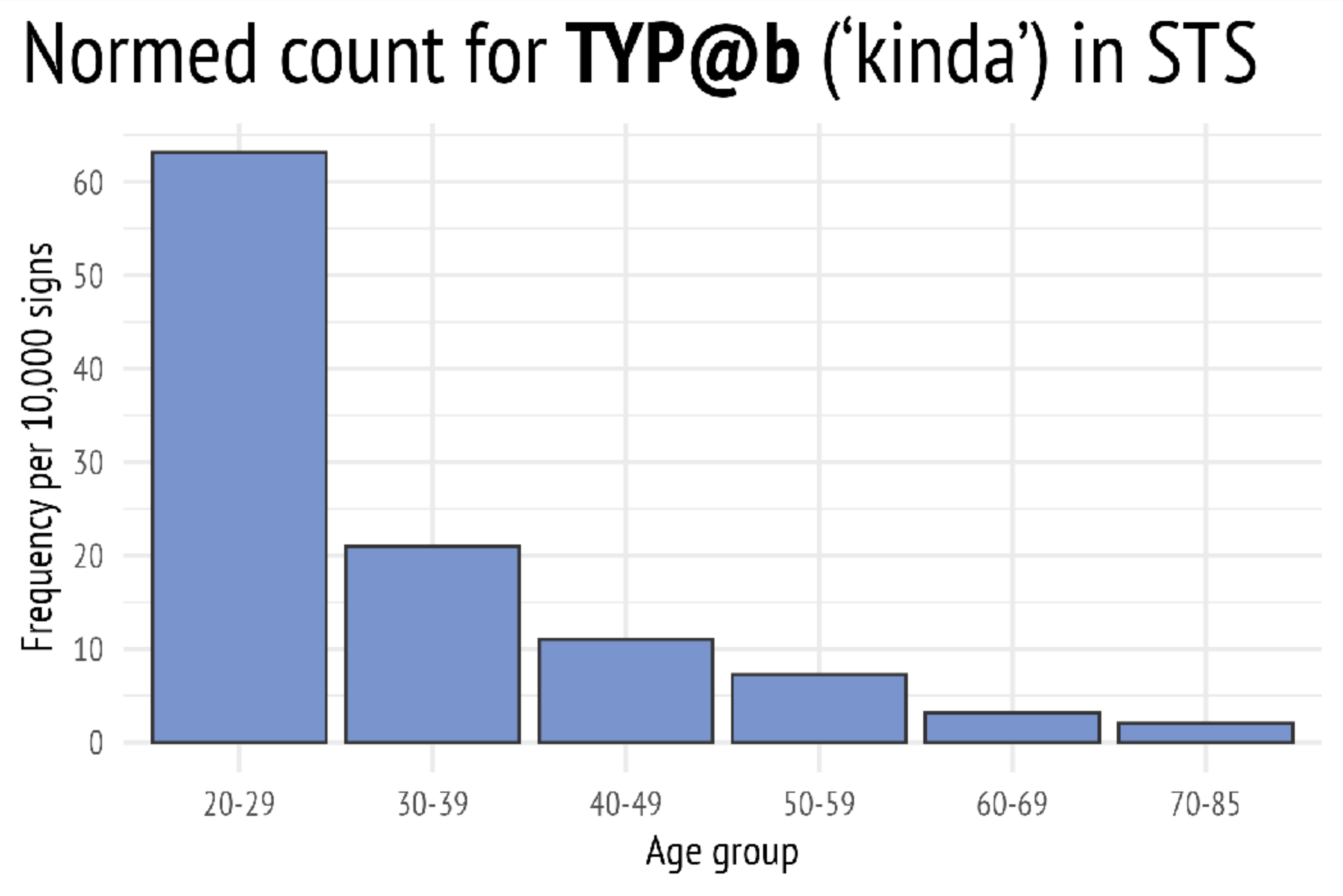
Börstell & Östling (2017); Börstell (2024); Börstell et al. (under review)

Lexical variation

Easier when

- you **know** specific candidates
- they are **frequent**
- they are themselves or with variants **dispersed** across signers

Example: the sign TYP@b →



Börstell & Östling (2017); Börstell (2024)

Lexical variation

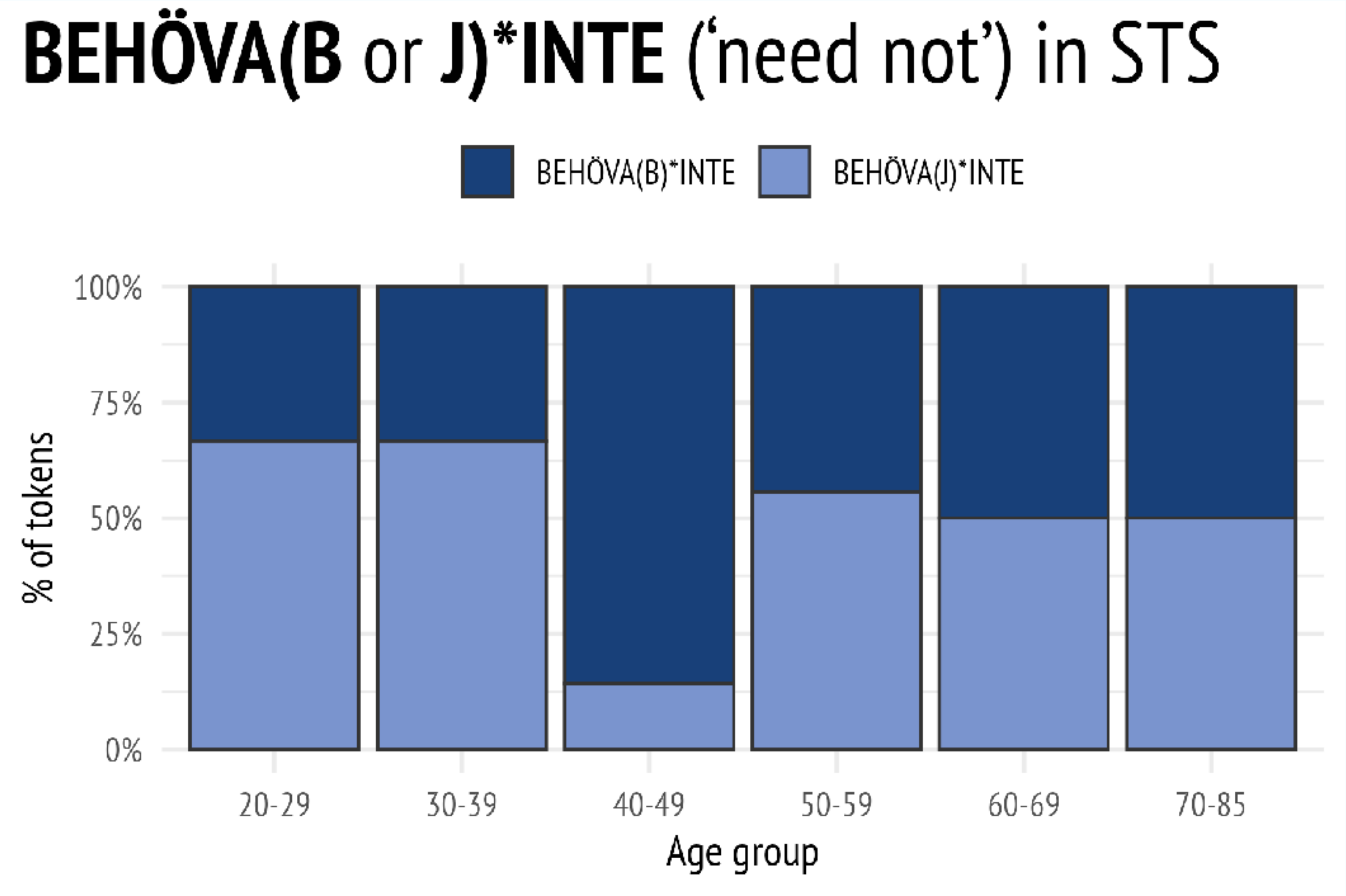
Easier when

– you **know** specific candidates

– they are **frequent**

– they are themselves or with variants **dispersed** across signers

Example: BEHÖVA(**B** or **J**) →

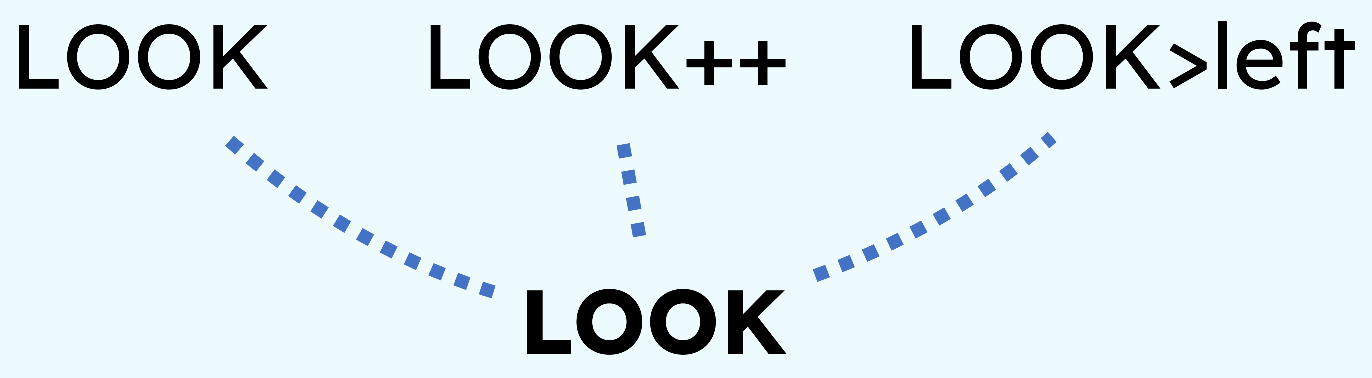


Börstell & Östling (2017); Börstell (2024)

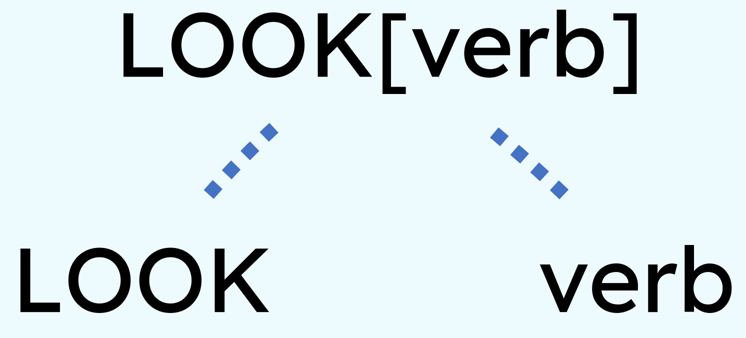
Frequency: how

Exported (and imported) sign **annotations from ELAN**

Count number of **occurrences per sign gloss**
– may involve some **trimming**/lemmatizing!



Extract **word class tags** from sign glosses
Count number of types per word class
Count total tokens per word class



Börstell (2022); Börstell et al. (2016)

Lexical variation: how

Count number of **occurrences per sign gloss**

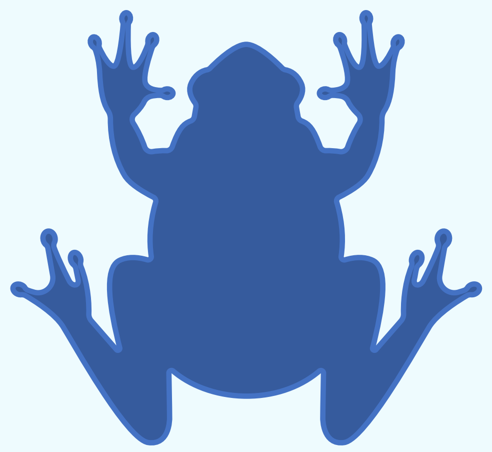
Combine data with **metadata** (age, gender, region, etc.)

Trim and match with **meanings or lemmas**

Normalize as **rate** (frequency relative to signer/group total)

Targeted **topics/texts** for specific items

Targeted interviews or **tasks** for specific items

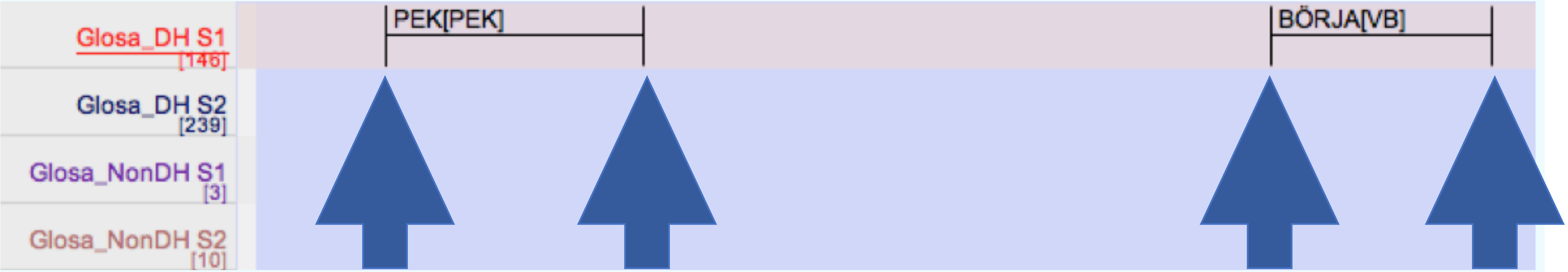


Börstell & Östling (2017); Börstell (2022, 2024a); Stamp et al. (2014)

Duration & articulation rate

Frequency and duration in STS

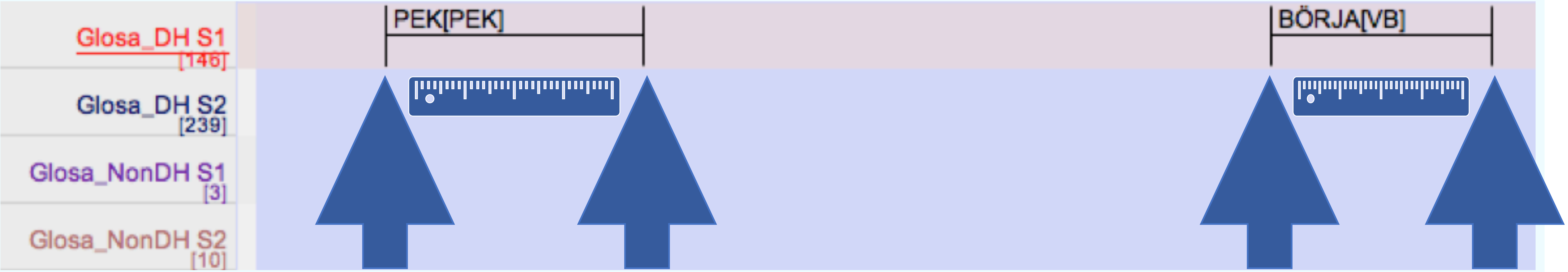
But wait, those annotations have an extension in time!



Börstell et al. (2016)

Frequency and duration in STS

But wait, those annotations have an extension in time!
– this means we get **durations** for free!



Börstell et al. (2016)

Frequency and duration in STS

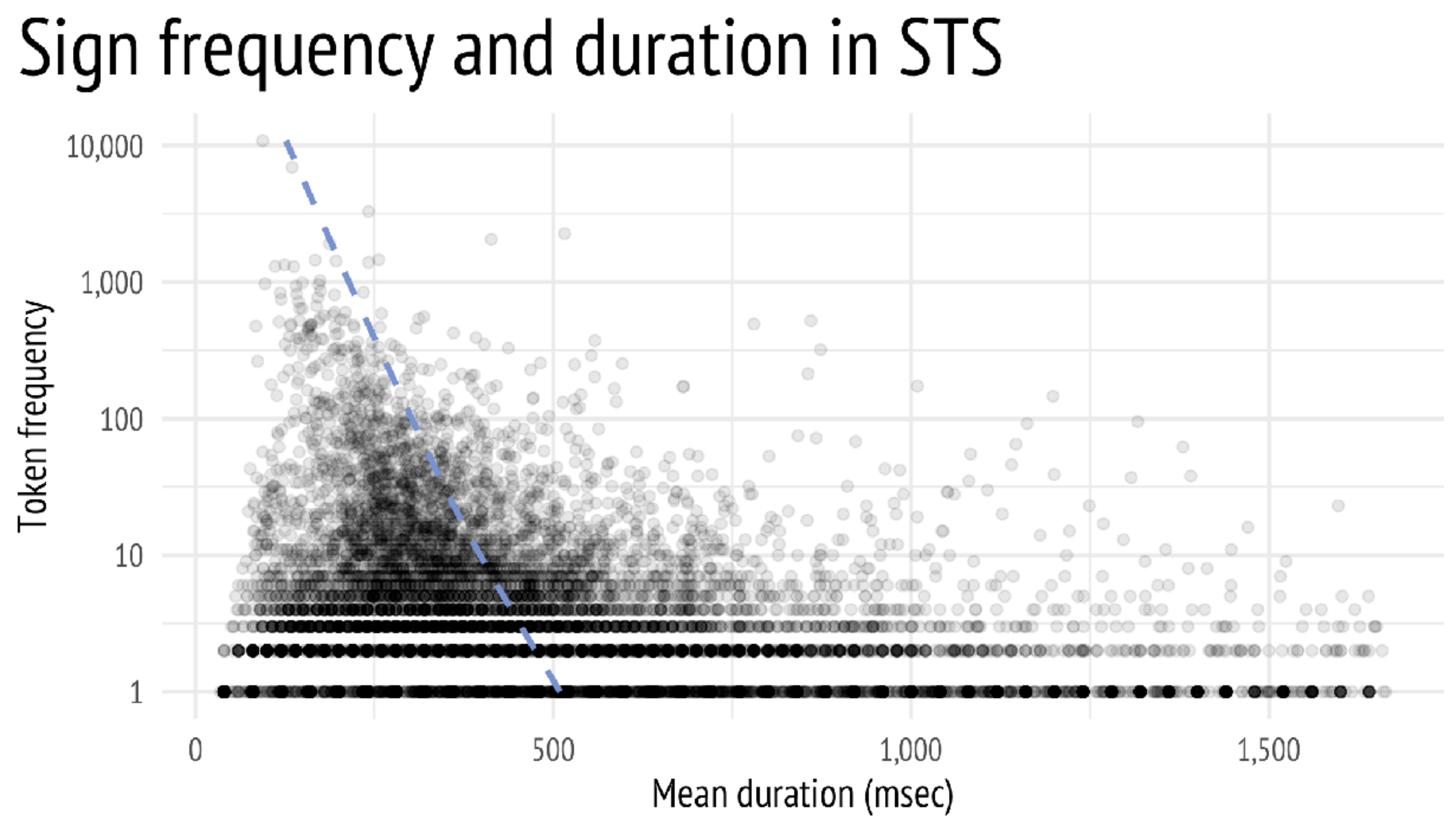
We took the durations in
ELAN and found:

Börstell et al. (2016)

Frequency and duration in STS

We took the durations in ELAN and found:

1) More frequent = shorter duration



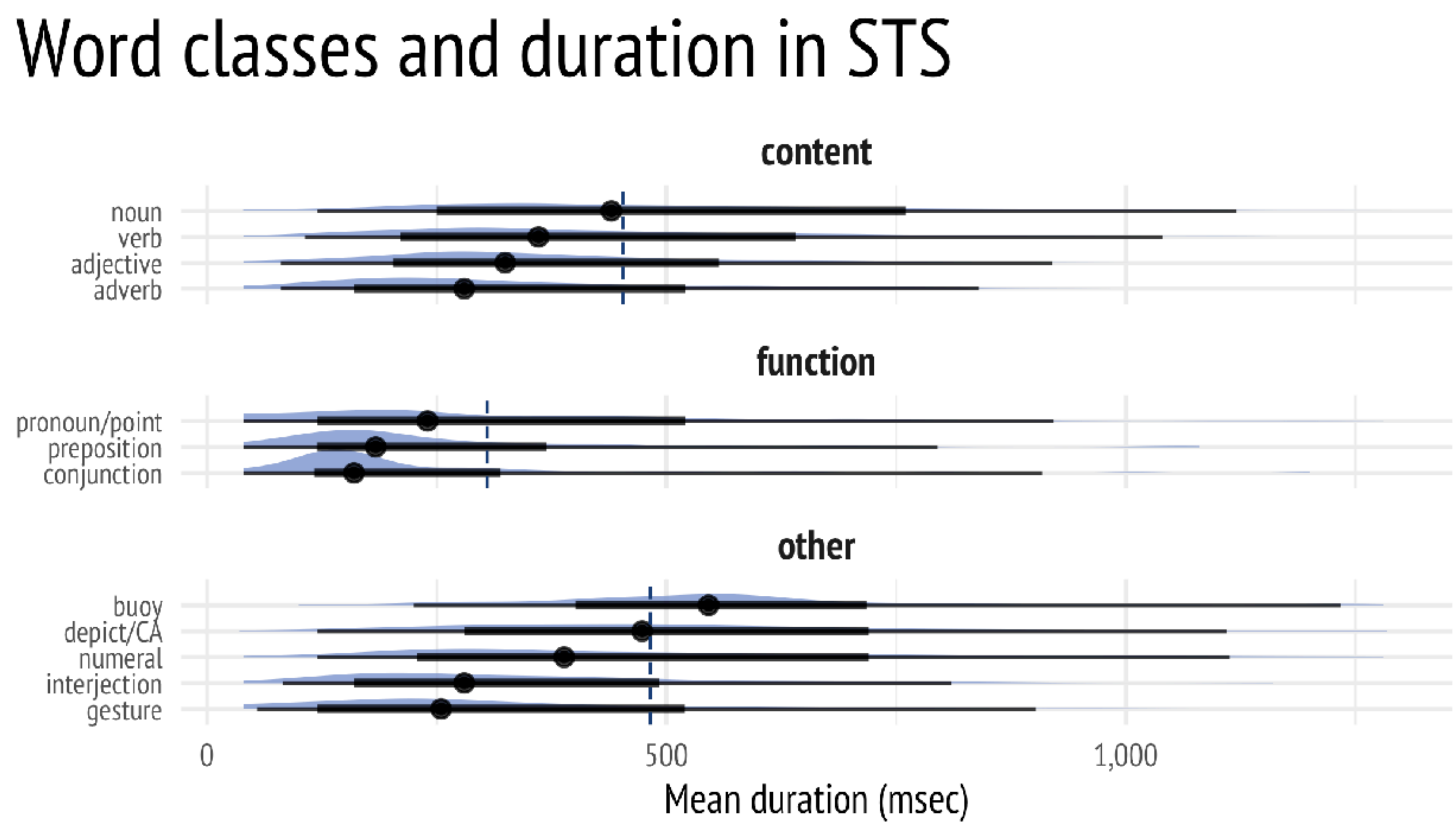
Börstell et al. (2016)

Frequency and duration in STS

We took the durations in ELAN and found:

1) More frequent = shorter duration

2) Content word classes > functional



Börstell et al. (2016)

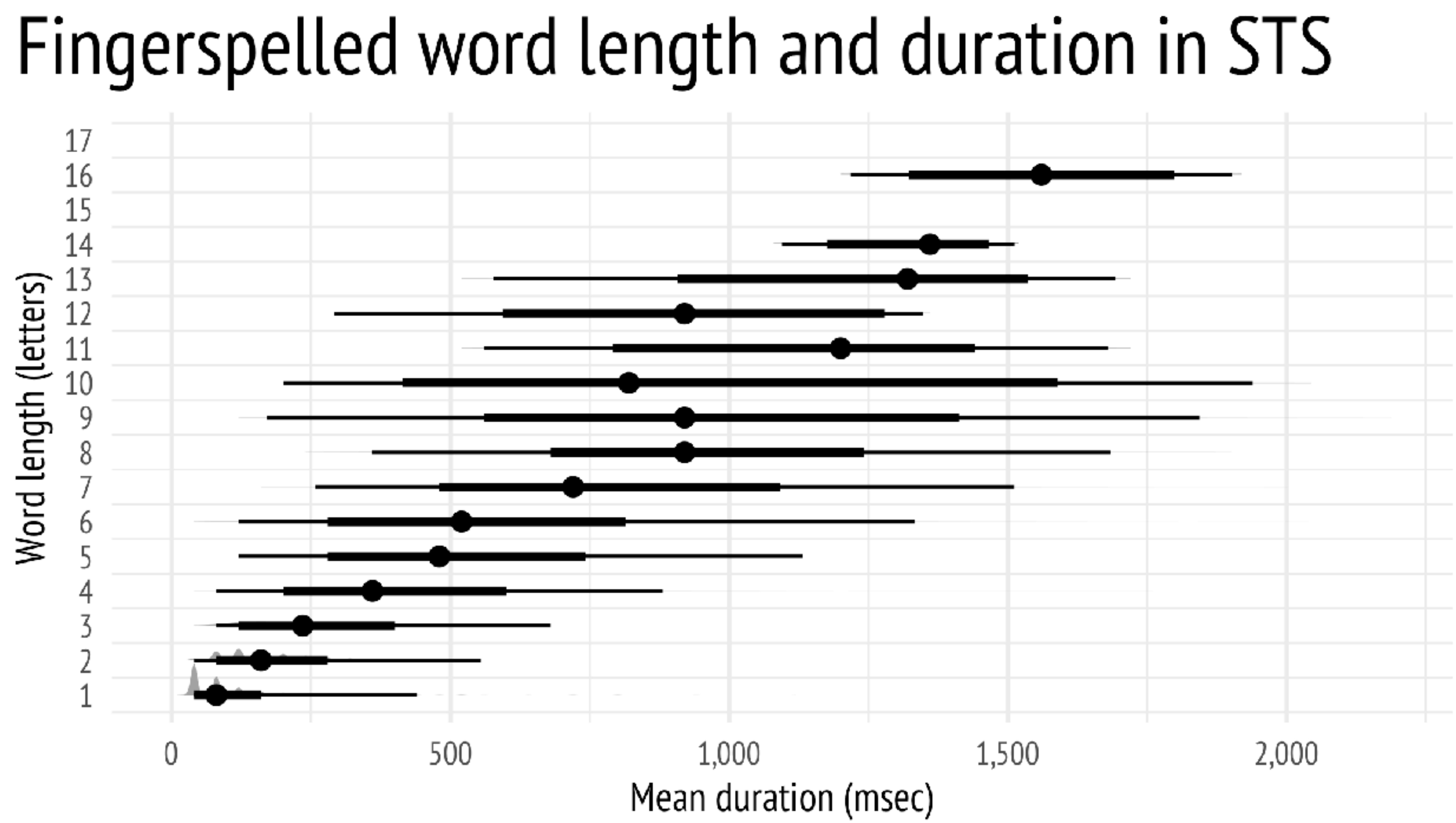
Frequency and duration in STS

We took the durations in ELAN and found:

1) More frequent = shorter duration

2) Content word classes > functional

3) Fingerspelling duration ↔ length



Börstell et al. (2016)

Frequency and duration in STS

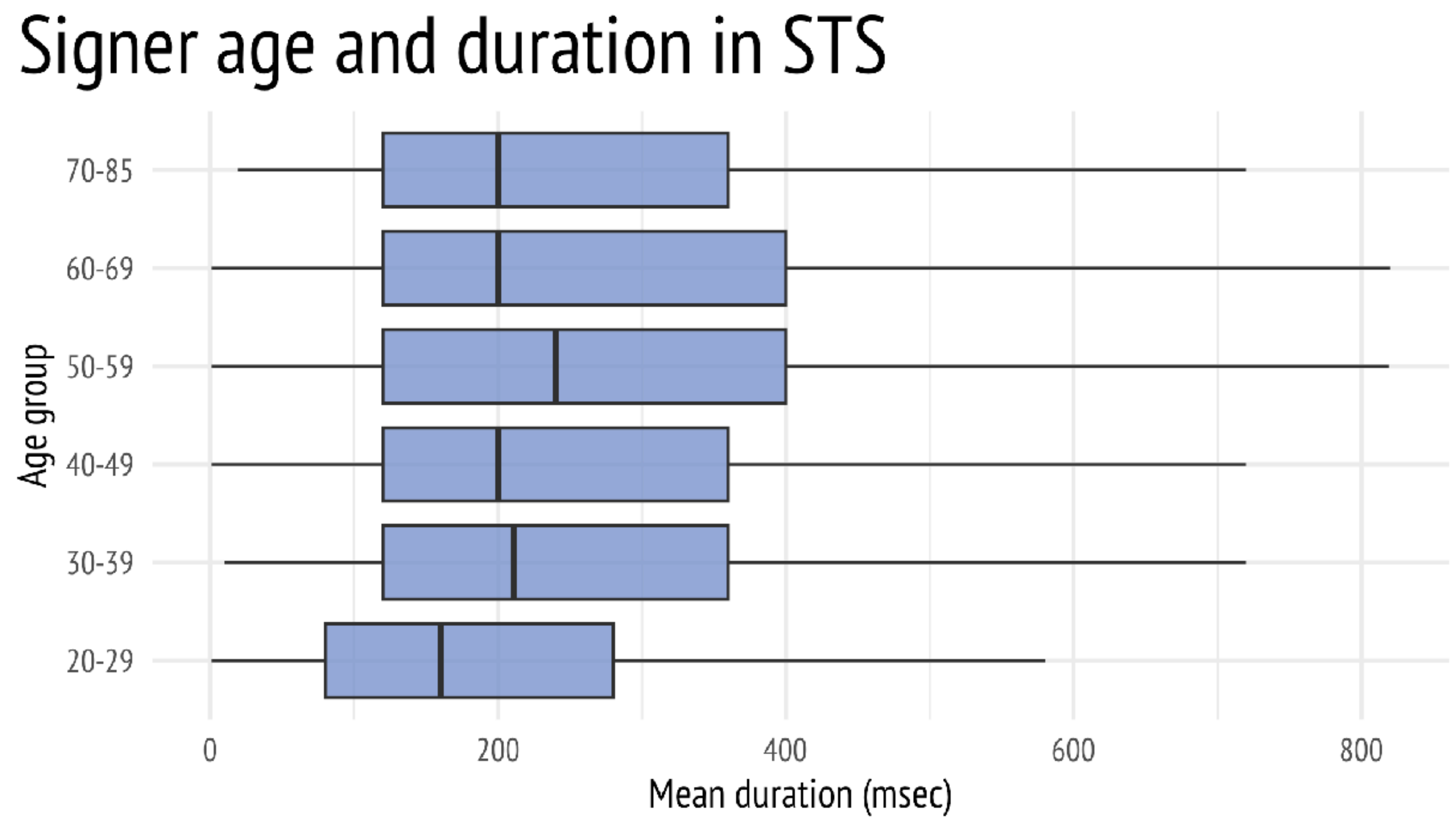
We took the durations in ELAN and found:

1) More frequent = shorter duration

2) Content word classes > functional

3) Fingerspelling duration ↔ length

4) age → duration

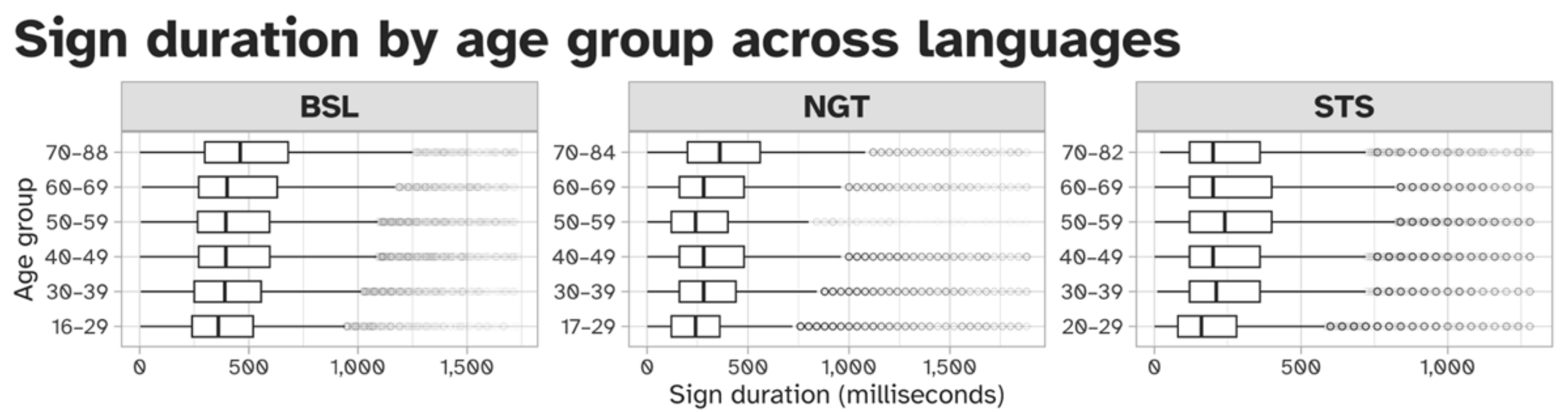


Börstell et al. (2016)

Signing rate in BSL, NGT and STS

We replicated and added to this with three SL corpora:

1) age → duration



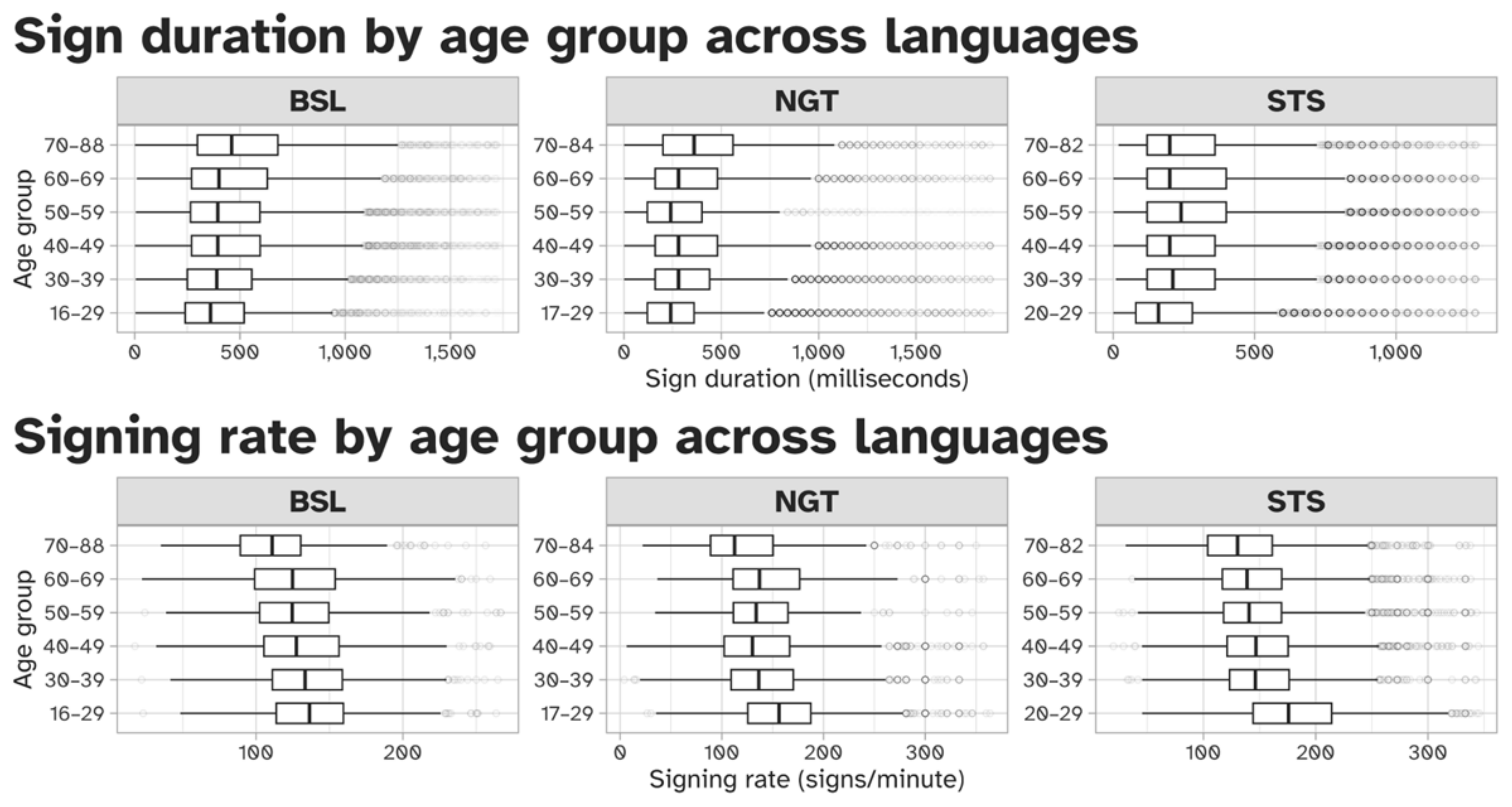
Börstell et al. (2024)

Signing rate in BSL, NGT and STS

We replicated and added to this with three SL corpora:

1) age → duration

2) age → rate



Börstell et al. (2024)

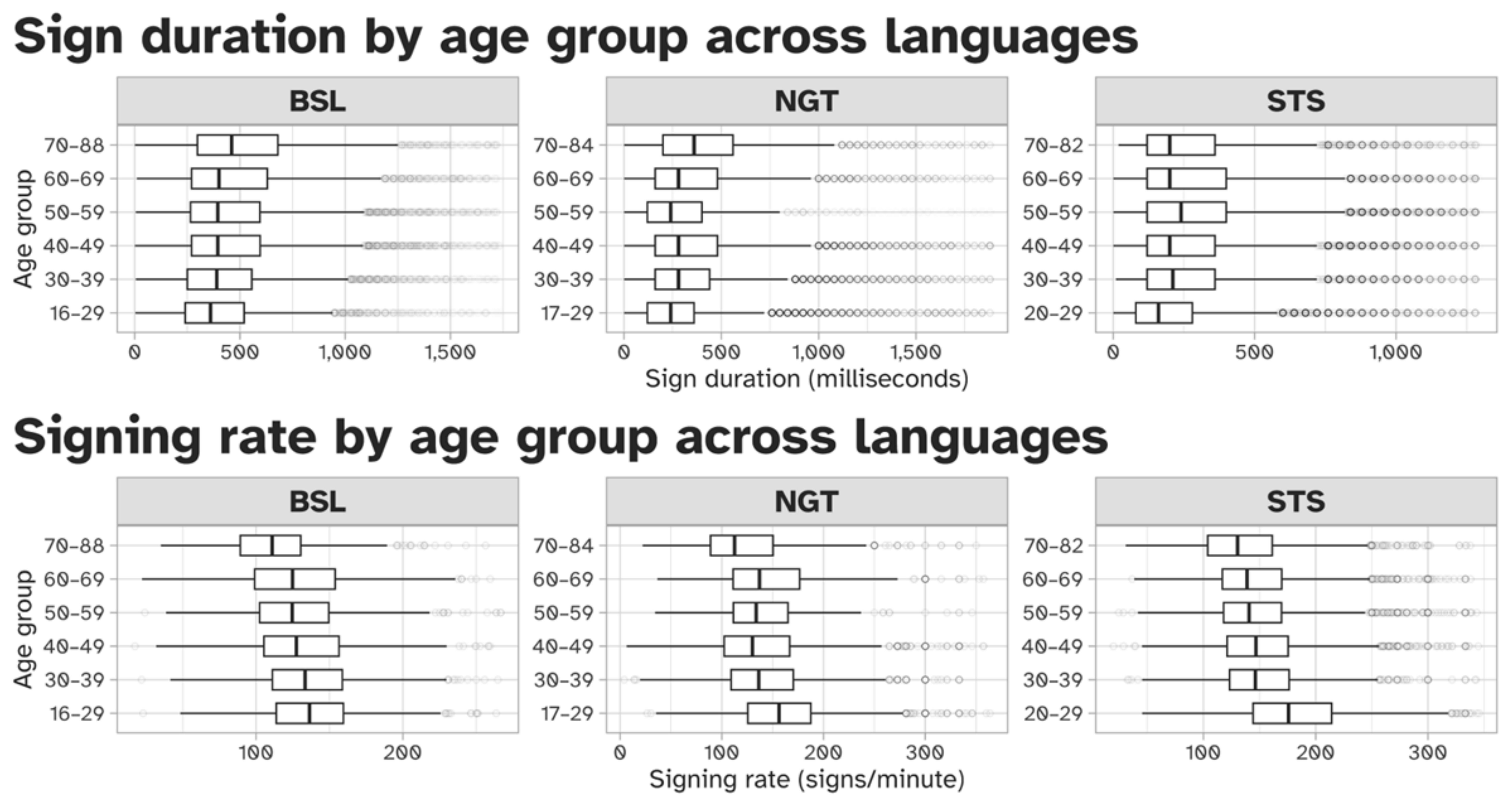
Signing rate in BSL, NGT and STS

We replicated and added to this with three SL corpora:

1) age → duration

2) age → rate

3) no effect of gender or family



Börstell et al. (2024)

Signing rate in BSL, NGT and STS

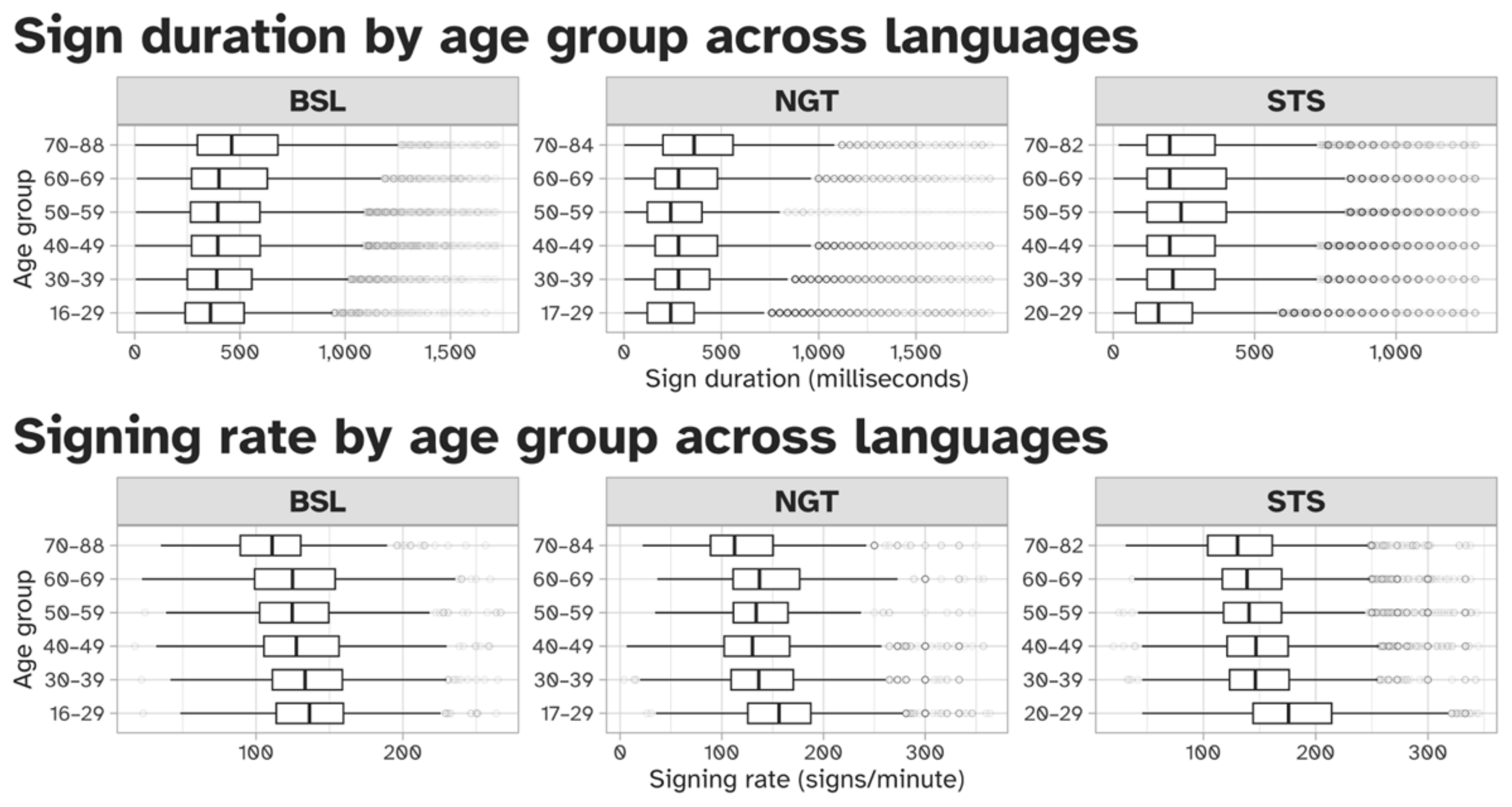
We replicated and added to this with three SL corpora:

1) age → duration

2) age → rate

3) no effect of gender or family

4) possible effect of region for BSL



Börstell et al. (2024)

Duration and rate: how

Exported (and imported) sign **annotations from ELAN**

Calculate duration based on timestamps (end – start)
– optional: adjust timestamps to match frame rate (fps 25 → 40 msec)

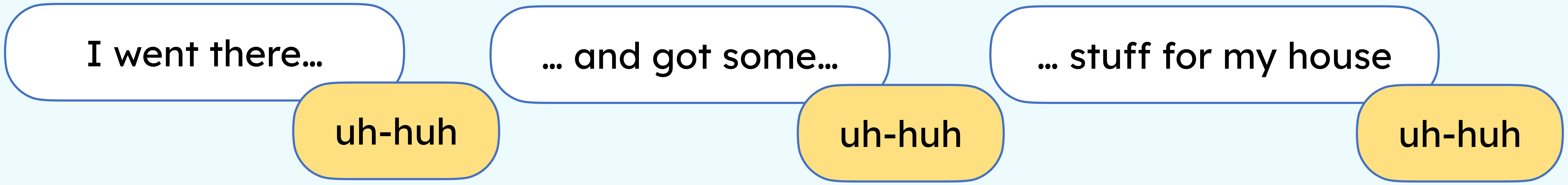
Count **number of signs** divided by **time** (e.g., utterance)
– optional: **infer utterances** from longer pauses (500 msec?)
– **note:** overlapping signs: OK; two-handed lexical signs: not OK
– **note:** alignment of annotations → cross-linguistic comparison?

Börstell (2022); Börstell et al. (2016; 2024)

Distributional patterns

Finding continuers (backchannels) in STS

With spoken language corpora, Dingemanse et al. (2022) developed a **language-agnostic** sequential search method for identifying **continuers**



With a few different approaches for **inferring utterances and turns**, I applied this sequential method to the STS Corpus data…

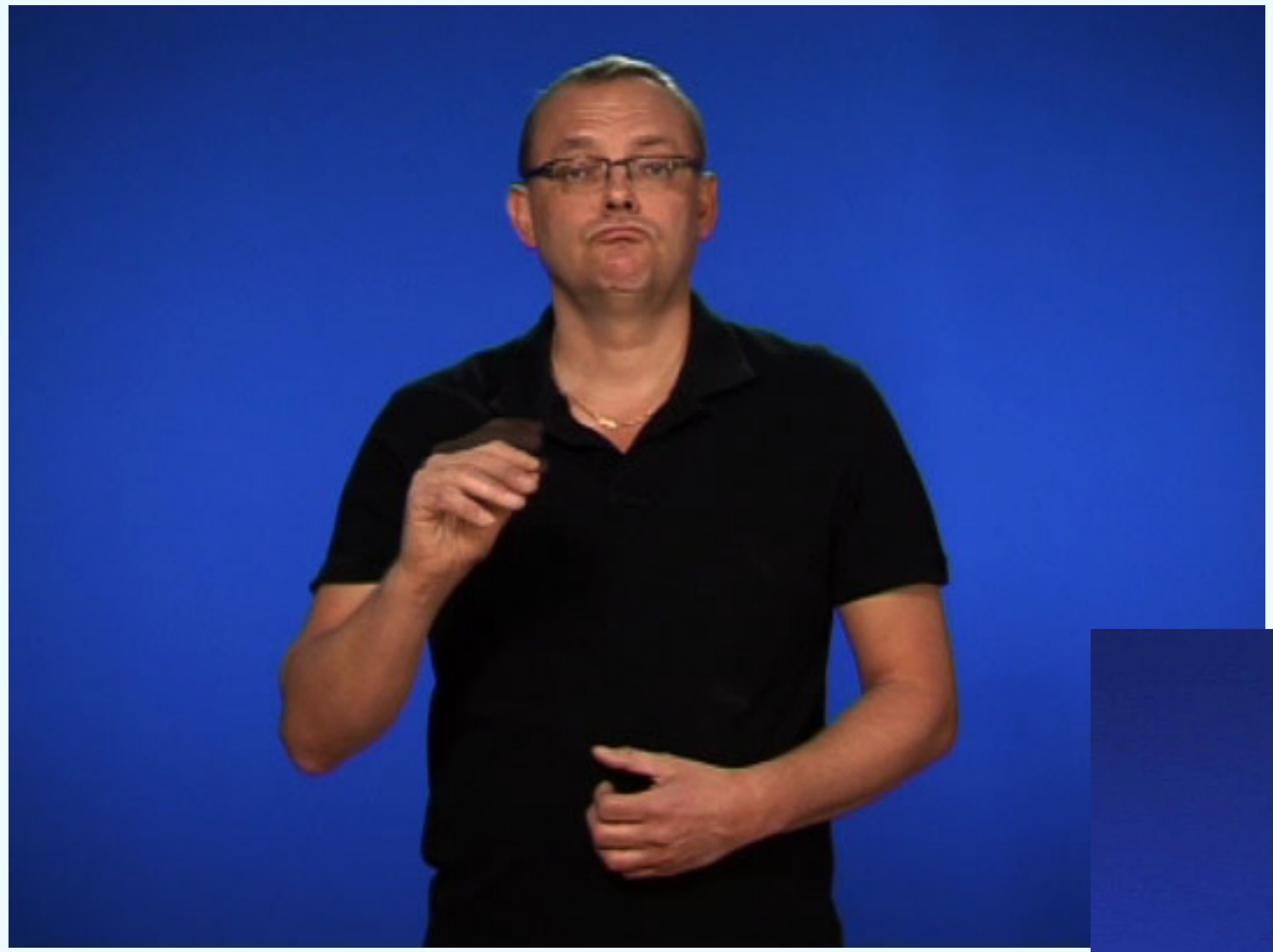
Dingemanse et al. (2022); Börstell (2024b)

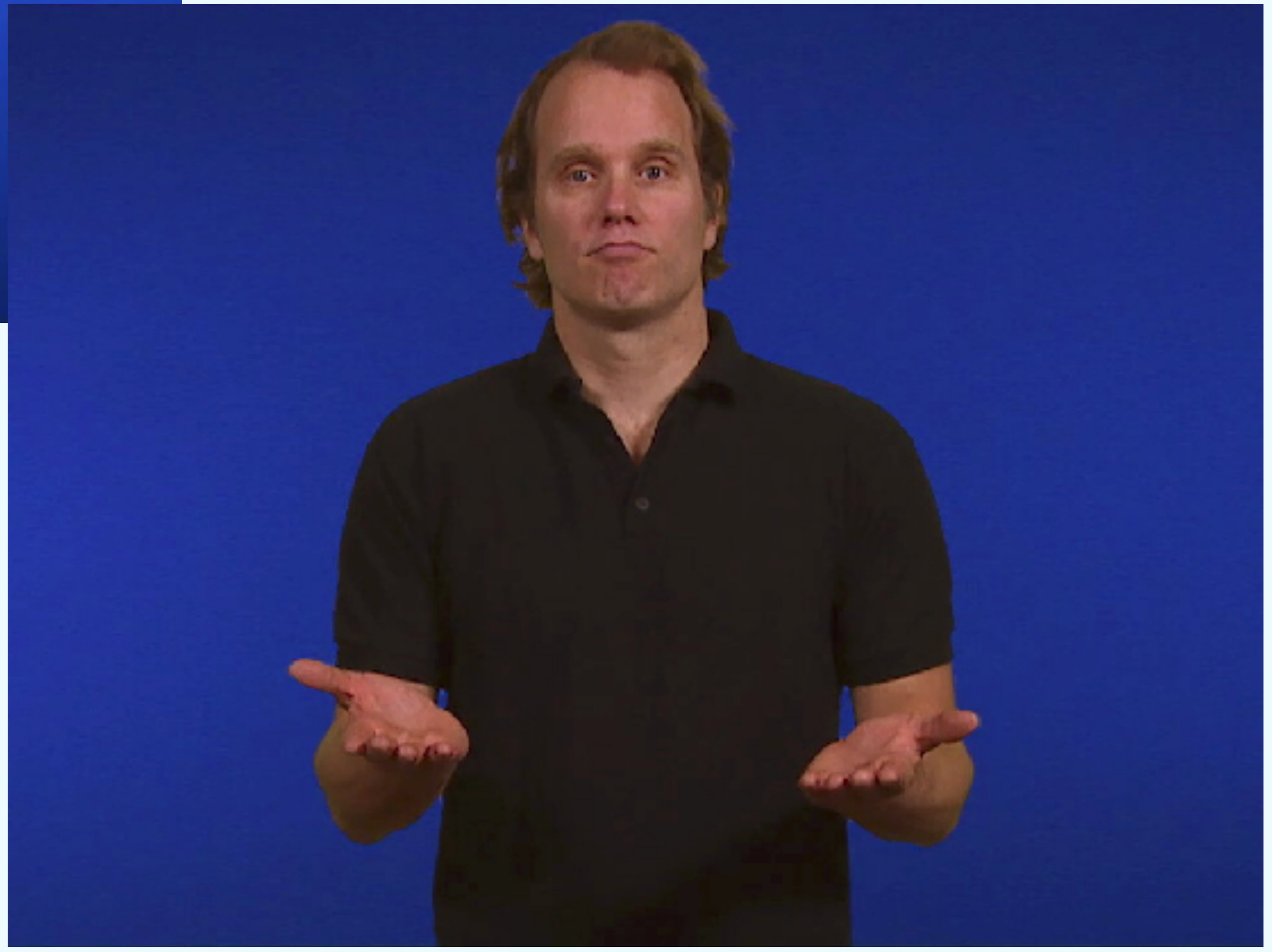
Finding continuers (backchannels) in STS

Two signs were **identified as** likely **continuers** in STS

JA@ub ‘yes’ (reduced)

PU@g (palms-up)





Further analysis found they are **longer than expected**, but also **visually less obtrusive** (**lower** and **less movement** in space)

Börstell (2024b); Svenskt teckenspråkslexikon (2025)
https://ideophone.org/finding-continuers-across-languages-and-modalities/

Something old,
something blue,
something borrowed,
something new...

Interactional profile

Since 2023, I've been more interested in finding ways to explore **conversational and pragmatic aspects** of SL corpus data **quantitatively**

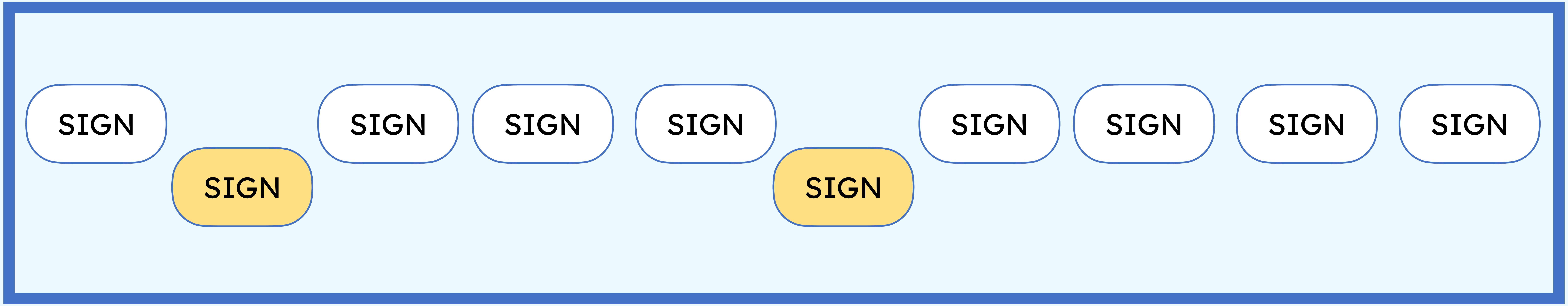
The continuers paper and the **sequential search method** worked great
– I also employed a method looking at **overlap between signers**

What could be inferred from different "**interactional profiles**" (nomenclature?) in **running text sequences** of **sign annotations**?

see also Börstell (2024b)

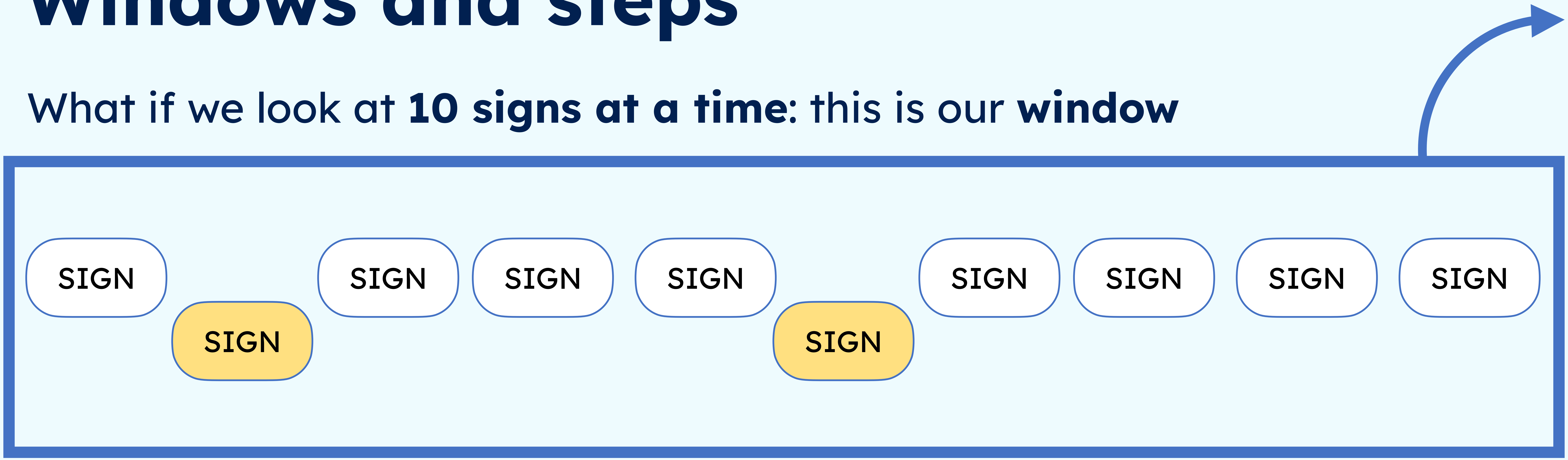
Windows and steps

What if we look at **10 signs at a time**: this is our **window**

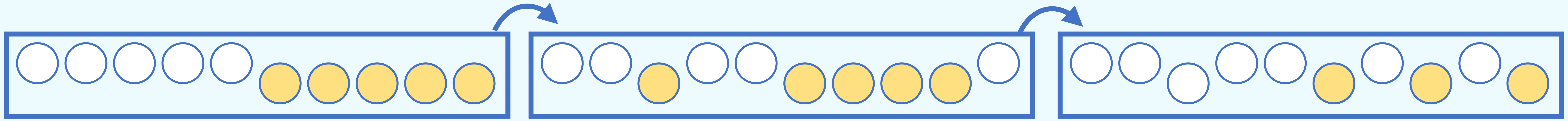


Windows and steps

What if we look at **10 signs at a time**: this is our **window**



… we then skip to the next 10 signs (no overlap): this is our **step size**



Signer entropy

Entropy is a measure of uncertainty

– in any given sequence of 10 signs, we calculate signer entropy: how surprised are we by picking a specific signer from that sequence?

Signer entropy

Entropy is a measure of uncertainty

– in any given sequence of 10 signs, we calculate signer entropy: how surprised are we by picking a specific signer from that sequence?

Entropy: 0

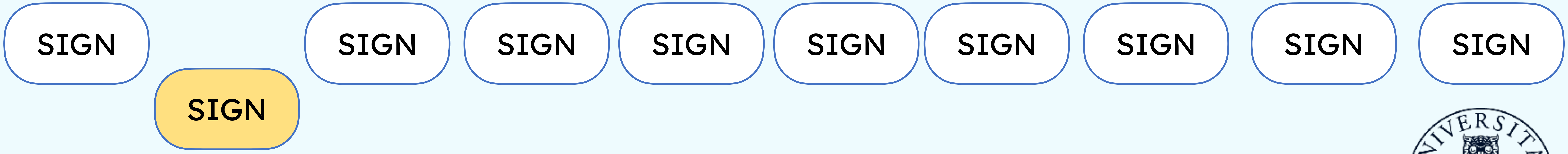
SIGN SIGN SIGN SIGN SIGN SIGN SIGN SIGN SIGN SIGN

Signer entropy

Entropy is a measure of uncertainty

– in any given sequence of 10 signs, we calculate signer entropy: how surprised are we by picking a specific signer from that sequence?

Entropy: 0.47

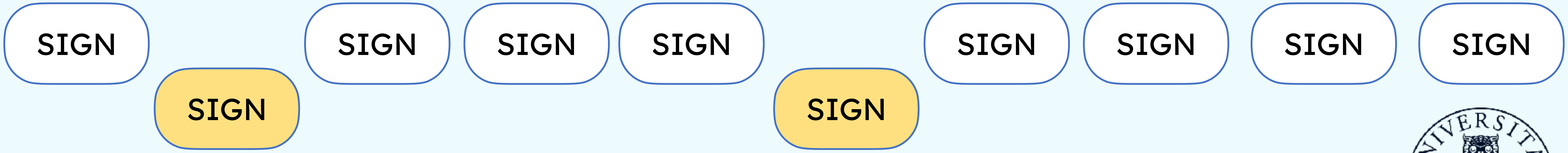


Signer entropy

Entropy is a measure of uncertainty

– in any given sequence of 10 signs, we calculate signer entropy: how surprised are we by picking a specific signer from that sequence?

Entropy: 0.72

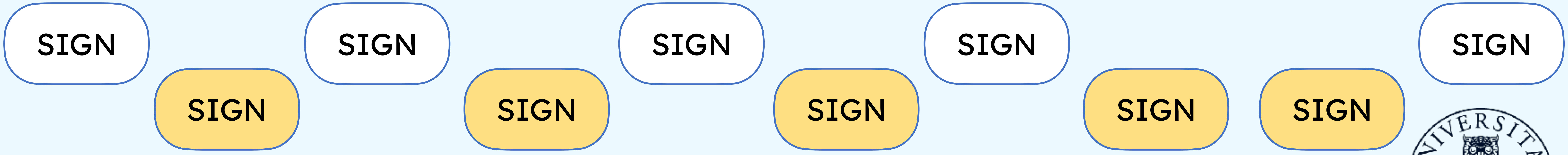


Signer entropy

Entropy is a measure of uncertainty

– in any given sequence of 10 signs, we calculate signer entropy: how surprised are we by picking a specific signer from that sequence?

Entropy: 1

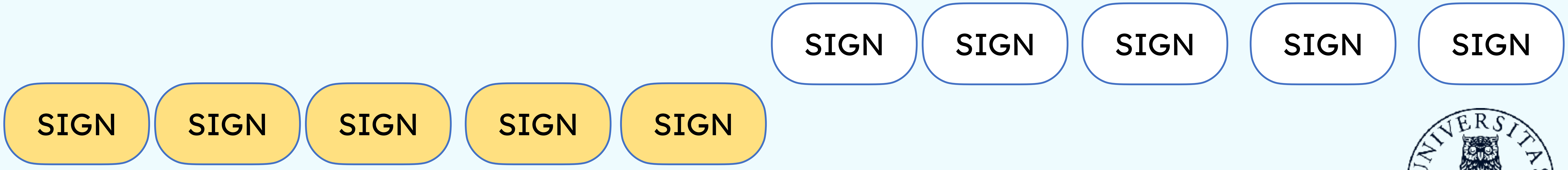


Signer entropy

Entropy is a measure of uncertainty

– in any given sequence of 10 signs, we calculate signer entropy: how surprised are we by picking a specific signer from that sequence?

Entropy: 1



Signer switches

For every window, how often does a signer switch occur

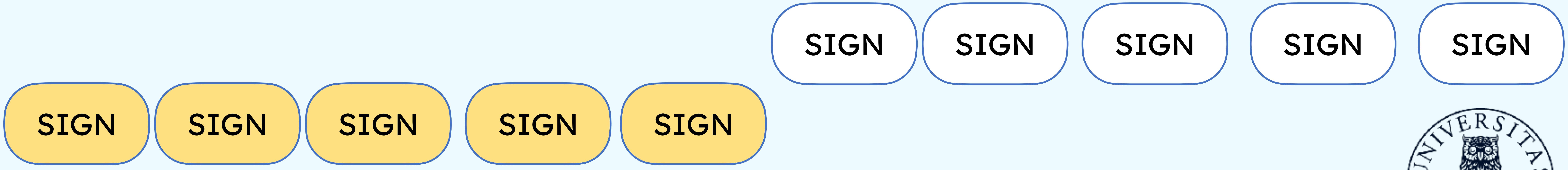
- based on time-ordered signs by start time, across both signers in a file
- 9 switches are possible within a window of 10 signs

Signer switches

For every window, how often does a signer switch occur

- based on time-ordered signs by start time, across both signers in a file
- 9 switches are possible within a window of 10 signs

Switch rate: 1/9 ≈ 0.11

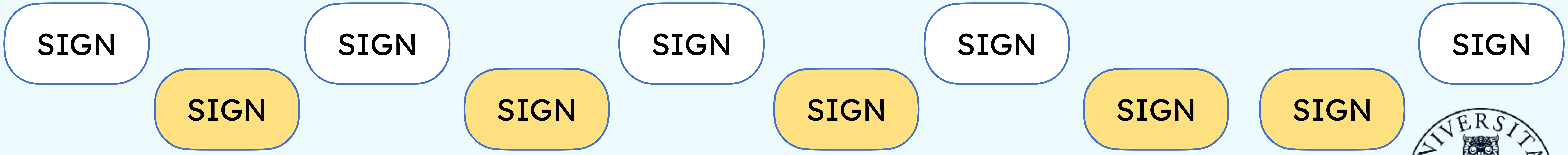


Signer switches

For every window, how often does a signer switch occur

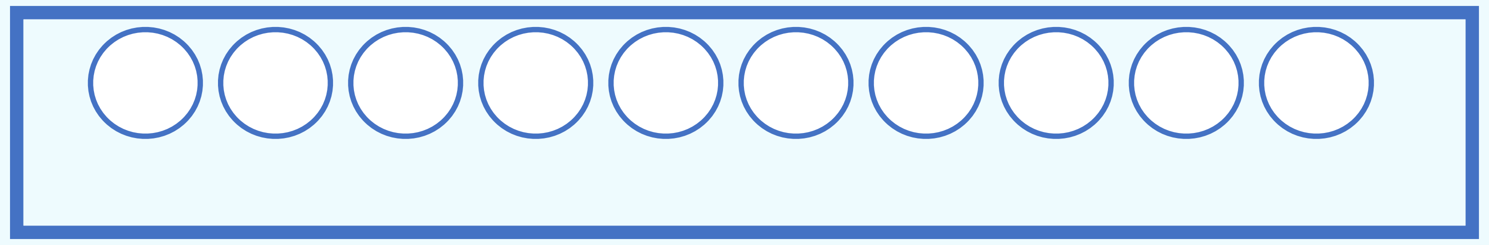
- based on time-ordered signs by start time, across both signers in a file
- 9 switches are possible within a window of 10 signs

Switch rate: 8/9 ≈ 0.89

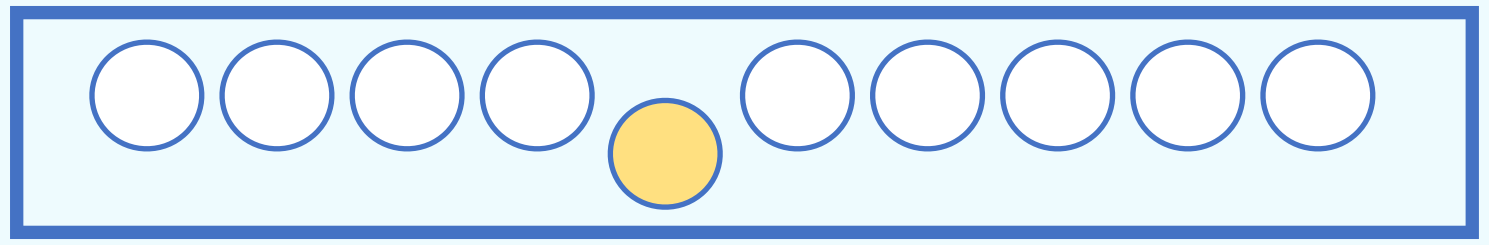


Interactional profile: hypotheses

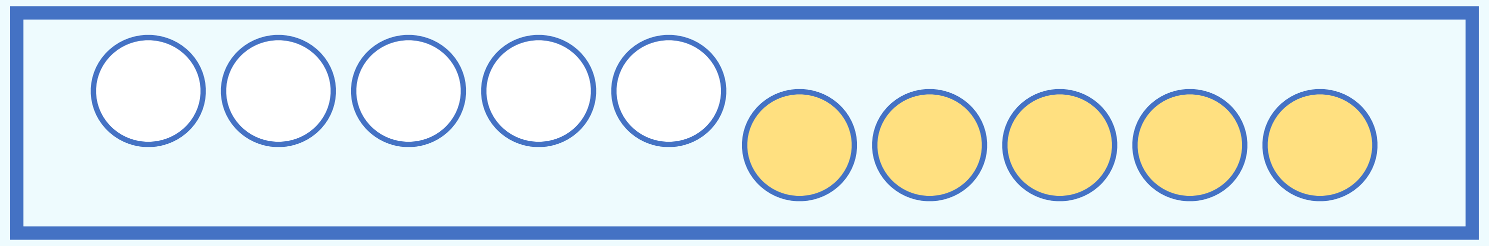
Zero entropy & Zero switch rate ≈ **main signer** (“monologue”)



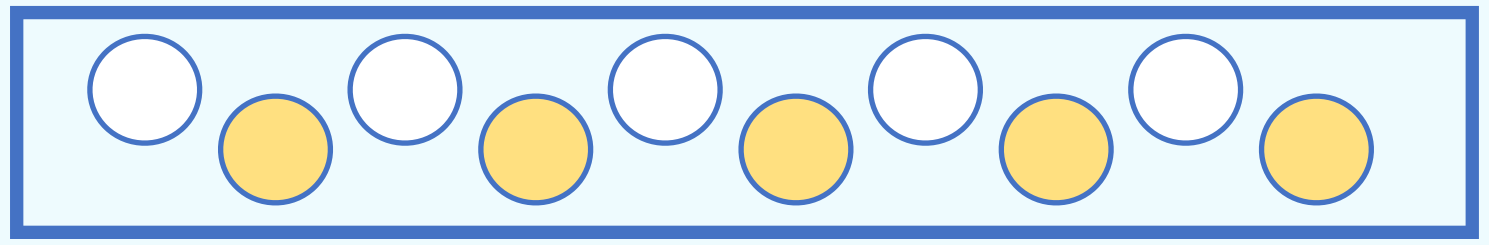
Low entropy & low switch rate ≈ **insertion**(s) (backchannels?)



High entropy & low switch rate ≈ **signer change**



High entropy & high switch rate ≈ **parallel** (negotiation?/repair?/TRP?)

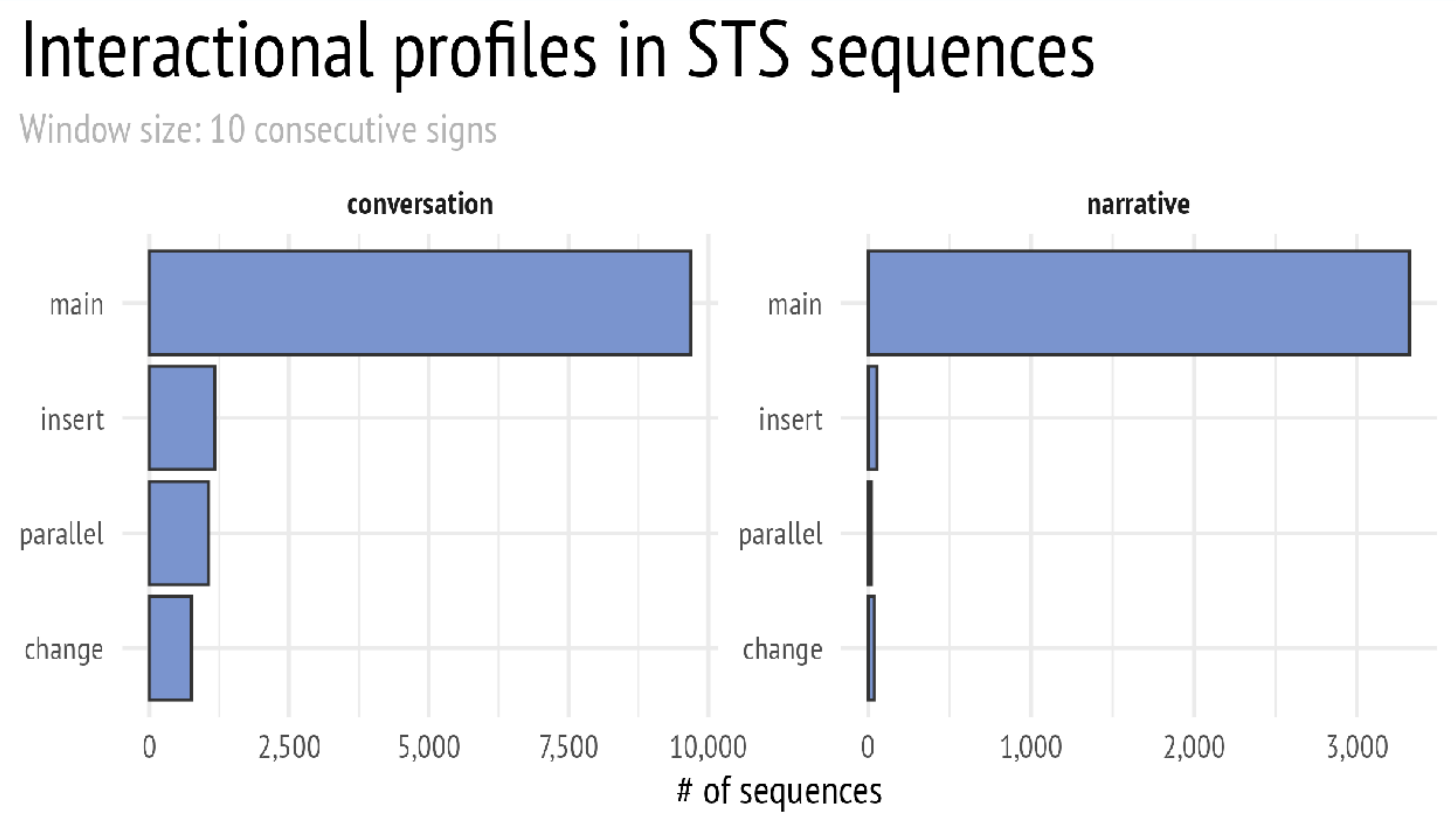


Results: distributions

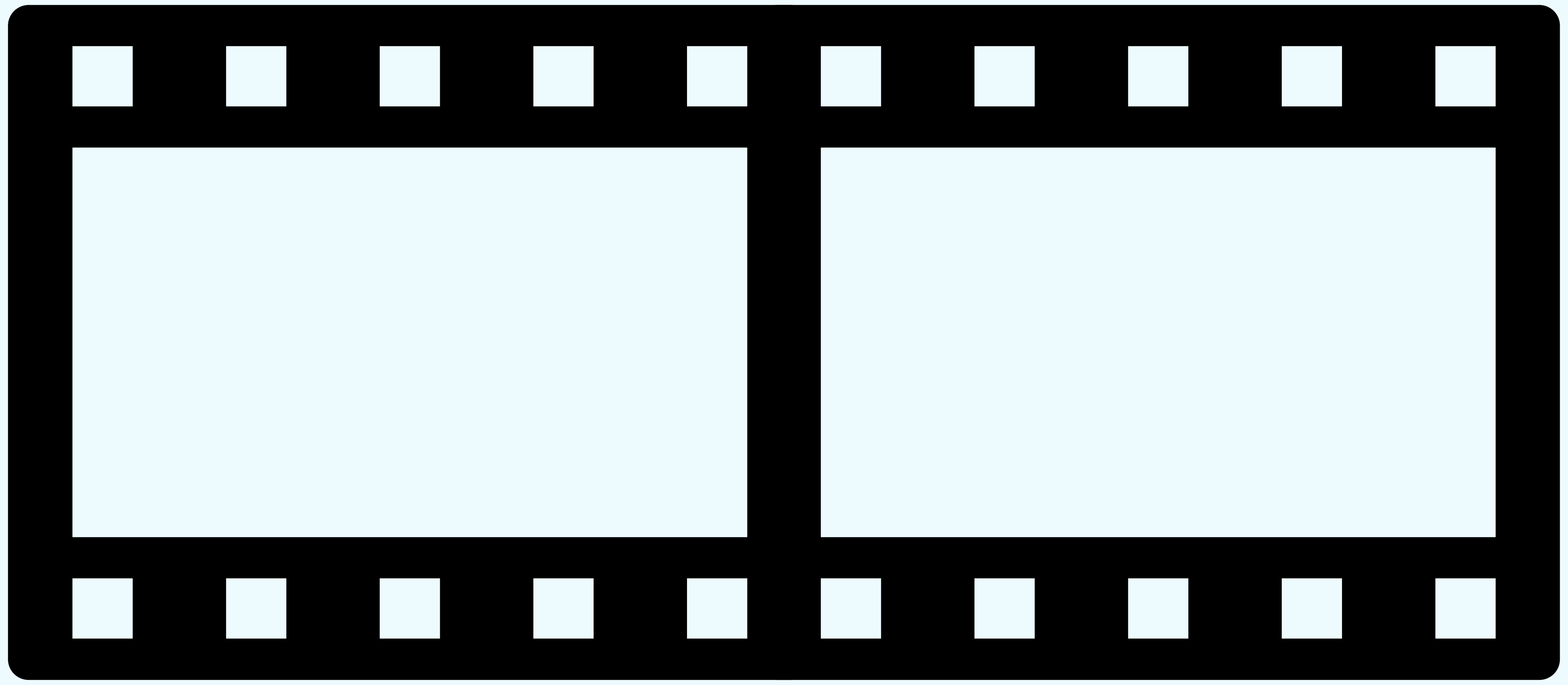
Mostly “**monologue**” sequences (69%)

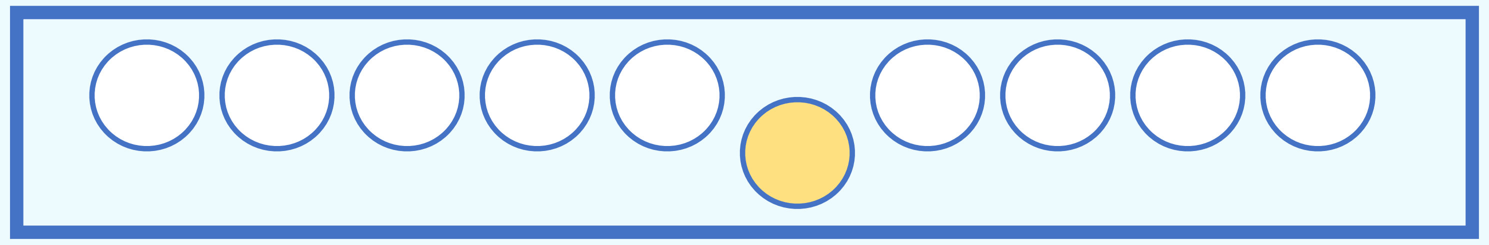
More interactional sequences in **conversation** texts

Many sequences (~15%) are **N/A**



Example: insertion

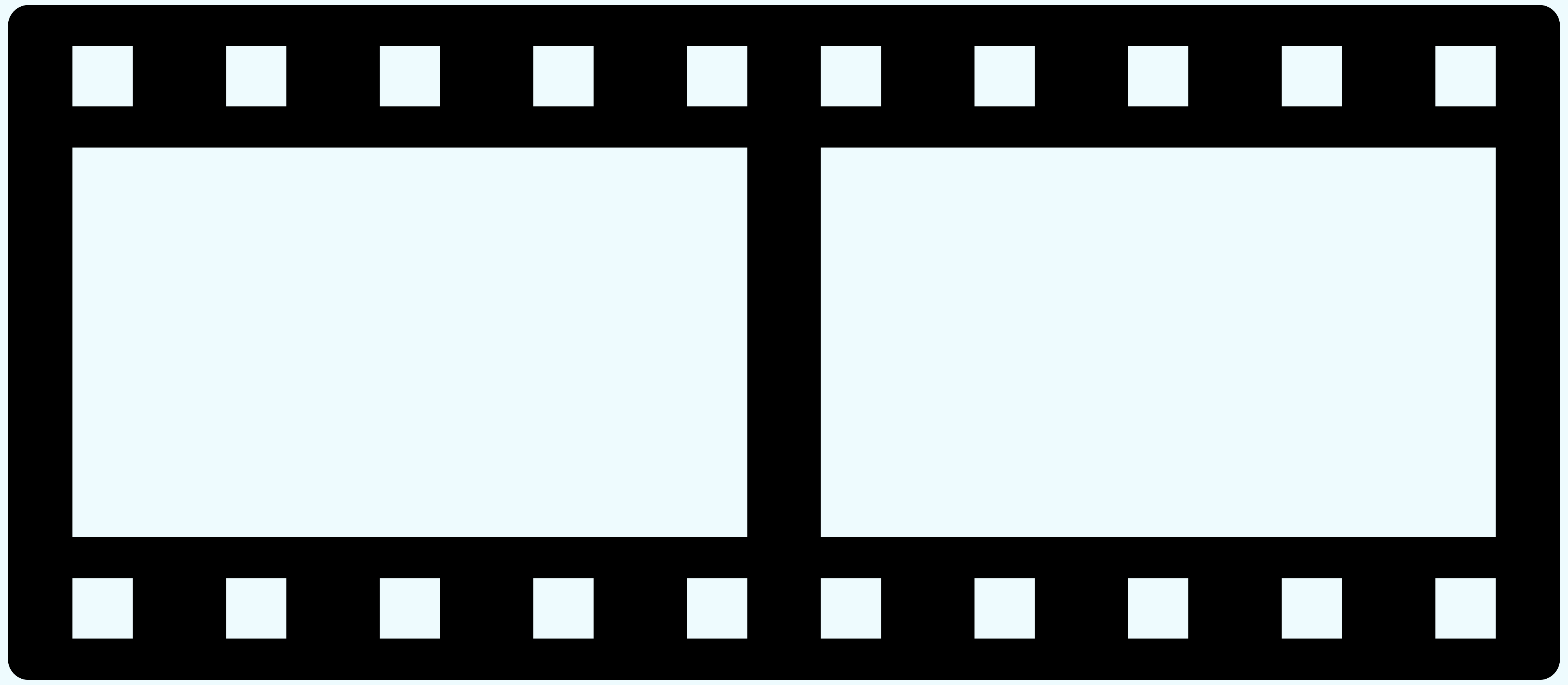




https://teckensprakskorpus.su.se/video/sslc01_264.eaf?t=203070

See also:
https://teckensprakskorpus.su.se/video/sslc01_322.eaf?t=193853
https://teckensprakskorpus.su.se/video/sslc01_203.eaf?t=431070

Example: change



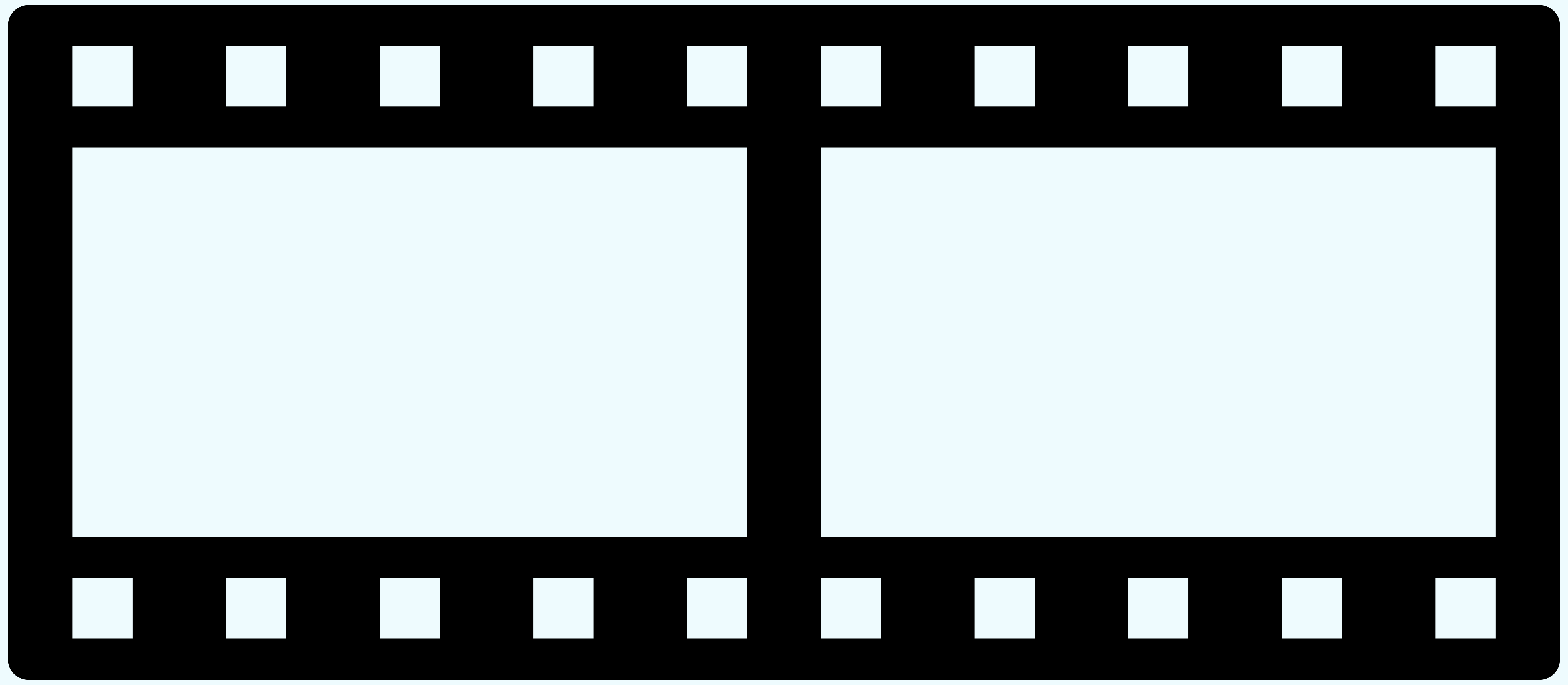
https://teckensprakskorpus.su.se/video/sslc01_021.eaf?t=476320

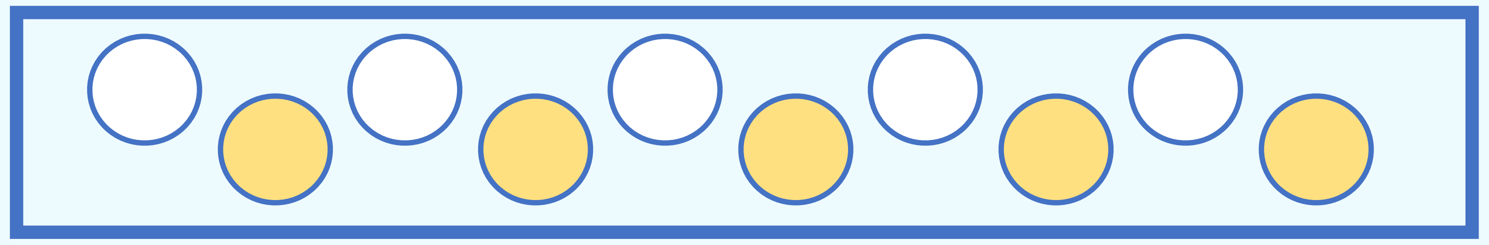
See also:

https://teckensprakskorpus.su.se/video/sslc01_307.eaf?t=228355

https://teckensprakskorpus.su.se/video/sslc01_141.eaf?t=174290

Example: parallel





https://teckensprakskorpus.su.se/video/sslc01_302.eaf?t=91580

See also:
https://teckensprakskorpus.su.se/video/sslc01_244.eaf?t=289640
https://teckensprakskorpus.su.se/video/sslc01_141.eaf?t=196862

Interactional profile: goals

We could target specific **places of interest** within files
– may **save time**; quicker than visual monitoring

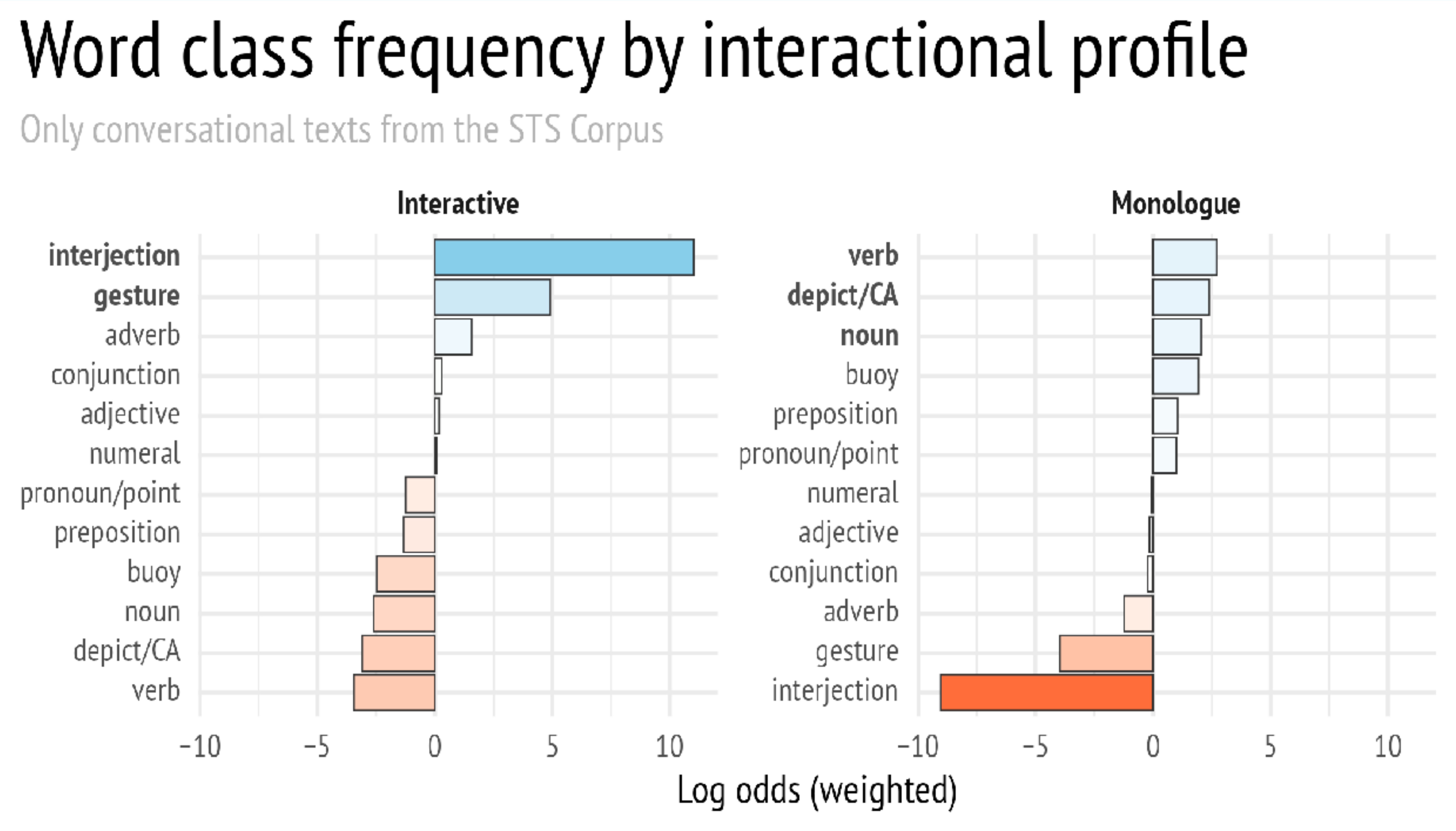
Interesting sequences could be looked at **qualitatively** in the corpus
– what is the **pragmatics** of different sequences?

Differences between the “**interactional profiles**” could be studied with regard to the **signs that occur within** them

Results: preliminary

Conversation only:

The interactional profile (distribution of signing activity) is reflected in **word class prevalence** among the signs in conversational texts

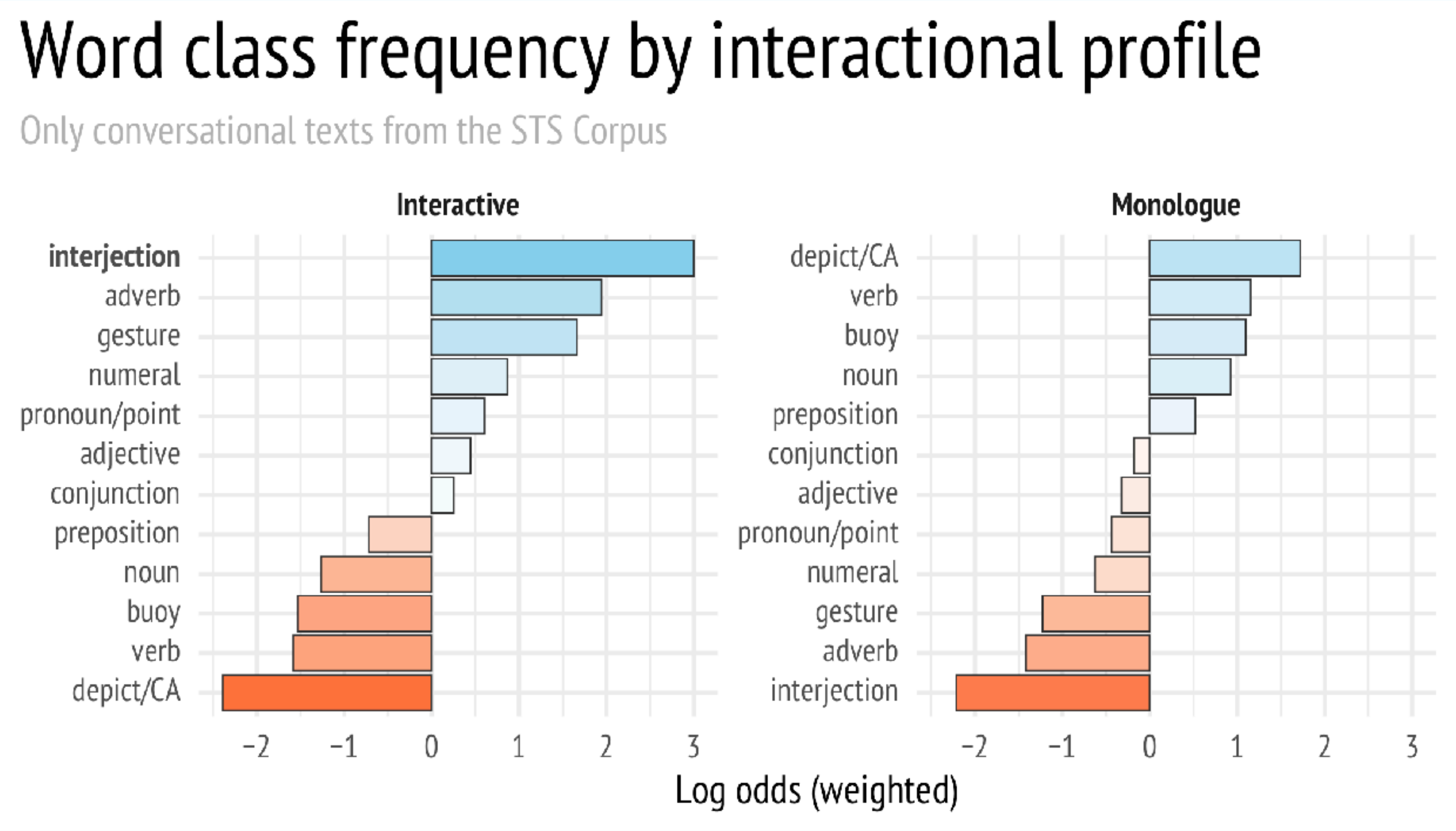


Results: preliminary

Conversation only:

The interactional profile (distribution of signing activity) is reflected in **word class prevalence** among the signs in conversational texts

… JA & PU removed

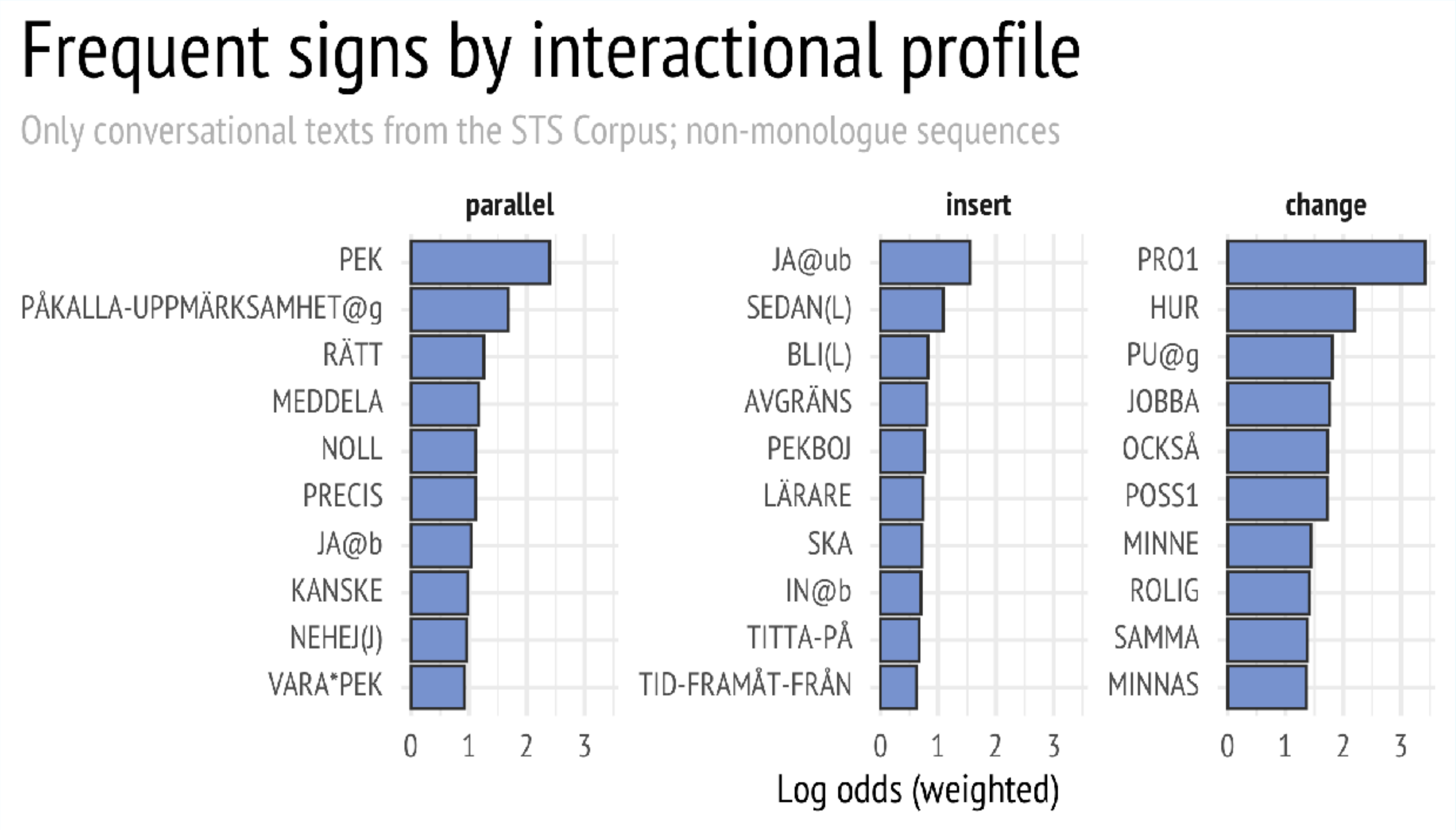


Results: preliminary

Conversation only:

Top-10 most frequent signs as distributed across the **non-monologue sequences** only

Some patterns, but also small dataset…



Feedback wanted!

Is this the **right way** to go?

Clear downside: **zero non-manual information**!

What are the **next steps**?

the WHY

Why do we use a corpus?

To **explore or confirm** theories, intuitions or anecdotal observations

Quantify claims that are made with little or artificial data
– allow for **variation** here: data is always **gradient** (= messy)

Use it as a source of **observations**: very possible
Use it as a source of **data-driven** distributions: still too small?

see, e.g., Kennedy (2014); O'Keefe (2018); Aijmer (2018)

Know your data!

Using a corpus is **not a substitute** for looking at your data!

The **context of occurrences** is important

You need to know annotation **conventions** and methodological **choices**

- BSL corpus: longer sign annotations and split files by signer
- STS corpus: two-handed signs are not doubled

Is the corpus **representative** for what you want to investigate?

Where do we go from here?

There is no sociolinguistics! ≈ **There is no corpus linguistics**!
– compare: experimental linguistics

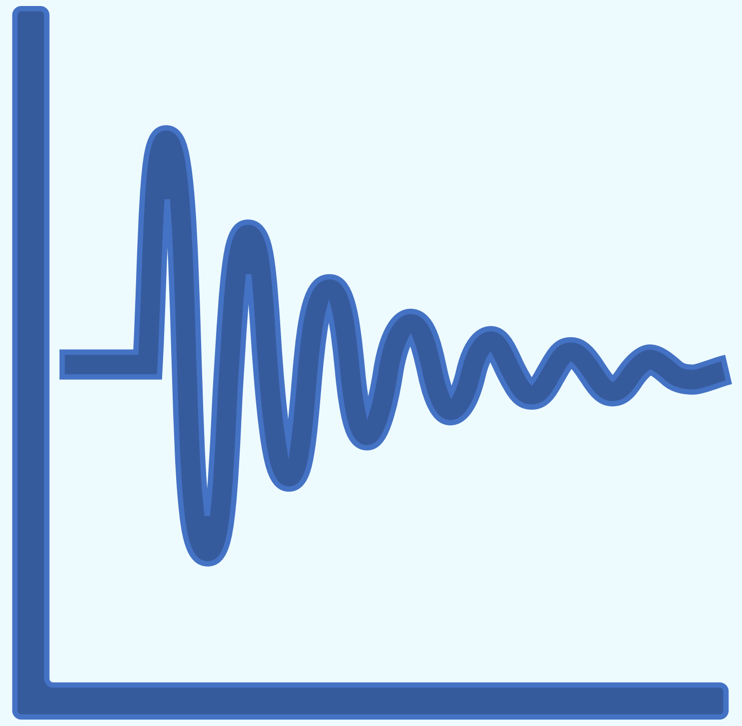
Corpus linguistics = **methods & resources** to use for Linguistics!
– assuming corpus-based rather than corpus-driven research

Focus on **systematic enhancement**, not expansion of "superficial" data
– or, maybe: ¿por qué no los dos? ('why not both?')

McEnery & Hardie (2011: xiv)

Computer vision?

Is computer vision the future for sign language corpora?



Maybe?

My view is:

– it can help **enhance current corpora** with little time and effort

– it can **capture** some **phonetic details** (but: errors!)

– it **cannot replace human** annotation/correction

see, e.g., EnvisionBOX and the NONMANUAL project

Feedback on the “interactional profiles”, please!

References

- Aijmer, Karin. 2018. Corpus pragmatics: From form to function. In Andreas H. Jucker, Klaus P. Schneider & Wolfram Bublitz (eds.), Methods in Pragmatics, 555–586. De Gruyter. https://doi.org/10.1515/9783110424928-022.
- Biber, Douglas, Susan Conrad & Randi Reppen. 1998. Corpus linguistics: investigating language structure and use (Cambridge Approaches to Linguistics). Cambridge ; New York: Cambridge University Press.
- Börstell, Carl. 2022. Searching and Utilizing Corpora. In Jordan Fenlon & Julie A. Hochgesang (eds.), Signed Language Corpora (Sociolinguistics in Deaf Communities 25), 90–127. Washington, DC: Gallaudet University Press.
- Börstell, Carl. 2024a. How to Approach Lexical Variation in Sign Language Corpora. In Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Johanna Mesch & Marc Schulder (eds.), Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources, 222–229. Torino, Italy: ELRA Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL). https://www.sign-lang.uni-hamburg.de/lrec/pub/24026.pdf.
- Börstell, Carl. 2024b. Finding continuers in Swedish Sign Language. Linguistics Vanguard. https://doi.org/10.1515/lingvan-2024-0025.
- Börstell, Carl, Thomas Hörberg & Robert Östling. 2016. Distribution and duration of signs and parts of speech in Swedish Sign Language. Sign Language & Linguistics 19(2). 143–196. https://doi.org/10.1075/sll.19.2.01bor.
- Börstell, Carl, Adam Schembri & Onno Crasborn. 2024. Sign duration and signing rate in British Sign Language, Dutch Sign Language and Swedish Sign Language. Glossa Psycholinguistics 3(1). https://doi.org/10.5070/G60111915.
- Börstell, Carl & Robert Östling. 2016. Visualizing Lects in a Sign Language Corpus: Mining Lexical Variation Data in Lects of Swedish Sign Language. In Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen & Johanna Mesch (eds.), Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining, 13–18. Portorož, Slovenia: European Language Resources Association (ELRA). https://www.sign-lang.uni-hamburg.de/lrec/pub/16004.pdf.
- Crasborn, Onno. 2015. Transcription and Notation Methods. In Eleni Orfanidou, Bencie Woll & Gary Morgan (eds.), Research Methods in Sign Language Studies, 74–88. Chichester: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118346013.ch5.

References

- Dingemanse, Mark, Andreas Liesenfeld & Marieke Woensdregt. 2022. Convergent Cultural Evolution of Continuers (mhmm). In Andrea Ravignani, Rie Asano, Daria Valente, Francesco Ferretti, Stefan Hartmann, Misato Hayashi, Yannick Jadoul, et al. (eds.), Proceedings of the Joint Conference on Language Evolution (JCoLE), 160–167. Kanazawa, Japan. https://doi.org/10.17617/2.3398549.
- Fenlon, Jordan & Julie A. Hochgesang (eds.). 2022. Signed Language Corpora (Sociolinguistics in Deaf Communities 25). Washington, DC: Gallaudet University Press. https://doi.org/10.2307/j.ctv2rcnfhc.
- Fenlon, Jordan, Adam Schembri, Ramas Rentelis, David Vinson & Kearsy Cormier. 2014. Using conversational data to determine lexical frequency in British Sign Language: The influence of text type. Lingua 143. 187–202. https://doi.org/10.1016/j.lingua.2014.02.003.
- Johnston, Trevor. 2010. From archive to corpus: Transcription and annotation in the creation of signed language corpora. International Journal of Corpus Linguistics 15(1). 106–131. https://doi.org/10.1075/ijcl.15.1.05joh.
- Johnston, Trevor. 2012. Lexical frequency in sign languages. Journal of Deaf Studies and Deaf Education 17(2). 163–193. https://doi.org/10.1093/deafed/enr036.
- Johnston, Trevor. 2014. The reluctant oracle: Adding value to, and extracting of value from, a signed language corpus through strategic annotations. Corpora 9(2). 155–189. https://doi.org/10.3366/cor.2014.0056.
- Joseph, Brian D. 2008. The Editor’s Department: Last scene of all . . . Language 84(4). 686–690. https://doi.org/10.1353/lan.0.0063.
- Kennedy, Graeme D. 2014. An Introduction to Corpus Linguistics (Studies in Language and Linguistics). Hoboken: Taylor and Francis.
- Kimchi, Inbal, Lucie Wolters, Rose Stamp & Inbal Arnon. 2023. Evidence of Zipfian distributions in three sign languages. Gesture 22(2). 154–188. https://doi.org/10.1075/gest.23014.kim.
- Kopf, Maria, Marc Schulder & Thomas Hanke. 2022. The Sign Language Dataset Compendium: Creating an Overview of Digital Linguistic Resources. In Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources, 102–109. Marseille, France: European Language Resources Association. https://aclanthology.org/2022.signlang-1.16.
- Kortmann, Bernd. 2021. Reflecting on the quantitative turn in linguistics. Linguistics 59(5). 1207–1226. https://doi.org/10.1515/ling-2019-0046.
- McEnery, Tony & Andrew Hardie. 2011. Corpus Linguistics: Method, Theory and Practice. 1st edn. Cambridge University Press. https://doi.org/10.1017/CBO9780511981395.

References

- McKee, David & Graeme Kennedy. 2006. The distribution of signs in New Zealand Sign Language. Sign Language Studies 6(4). 372–391. https://doi.org/10.1353/sls.2006.0027.
- Mesch, Johanna, Lars Wallin, Anna-Lena Nilsson & Brita Bergman. 2012. Dataset. Swedish Sign Language Corpus project 2009–2011 (version 1). Sign Language Section, Department of Linguistics, Stockholm University. https://teckensprakskorpus.su.se.
- Morford, Jill P. & James MacFarlane. 2003. Frequency characteristics of American Sign Language. Sign Language Studies 3(2). 213–226. https://doi.org/10.1353/sls.2003.0003.
- O'Keeffe, Anne. 2018. Corpus-based function-to-form approaches. In Andreas H. Jucker, Klaus P. Schneider & Wolfram Bublitz (eds.), Methods in Pragmatics, 587–618. De Gruyter. https://doi.org/10.1515/9783110424928-023.
- Öqvist, Zrajm, Nikolaus Riemer Kankkonen & Johanna Mesch. 2020. STS-korpus: A Sign Language Web Corpus Tool for Teaching and Public Use. In Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen & Johanna Mesch (eds.), Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives, 177–180. Marseille, France: European Language Resources Association (ELRA). https://www.sign-lang.uni-hamburg.de/lrec/pub/20014.pdf.
- Östling, Robert, Carl Börstell, Moa Gärdenfors & Mats Wirén. 2017. Universal Dependencies for Swedish Sign Language. In Jörg Tiedemann & Nina Tahmasebi (eds.), Proceedings of the 21st Nordic Conference on Computational Linguistics, 303–308. Gothenburg, Sweden: Association for Computational Linguistics. https://aclanthology.org/W17-0243.
- Östling, Robert, Carl Börstell & Lars Wallin. 2015. Enriching the Swedish Sign Language Corpus with Part of Speech Tags Using Joint Bayesian Word Alignment and Annotation Transfer. In Proceedings of the 20th Nordic Conference of Computational Linguistics (NoDaLiDa 2015), 263–268. Vilnius, Lithuania: Linköping University Electronic Press, Sweden. https://aclanthology.org/W15-1834.
- Stamp, Rose, Adam Schembri, Jordan Fenlon, Ramas Rentelis, Bencie Woll & Kearsy Cormier. 2014. Lexical variation and change in British Sign Language. PLoS ONE 9(4). https://doi.org/10.1371/journal.pone.0094053.
- Tognini-Bonelli, Elena. 2001. Corpus linguistics at work (Studies in Corpus Linguistics volume 6). Amsterdam Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/scl.6.
- Zipf, George K. 1935. The psycho-biology of language: An introduction to dynamic philology. New York, NY: Houghton Mifflin.
- Zipf, George Kingsley. 1949. Human behavior and the principle of least effort: An introduction to human ecology. Cambridge, MA: Addison-Wesley.

Thank you!
Kiitos paljon!
Tusen takk!
Tack så mycket!