Spoken typology

Incorporating the spoken signal into grammatical typology

Laura Becker

CNRS, LLF, Paris

Oct 16, 2025

Approaches in typology

Typology: the study of variation and trends of linguistic phenomena in the world's languages

Approaches in typology

Typology: the study of variation and trends of linguistic phenomena in the world's languages

field data

grammars

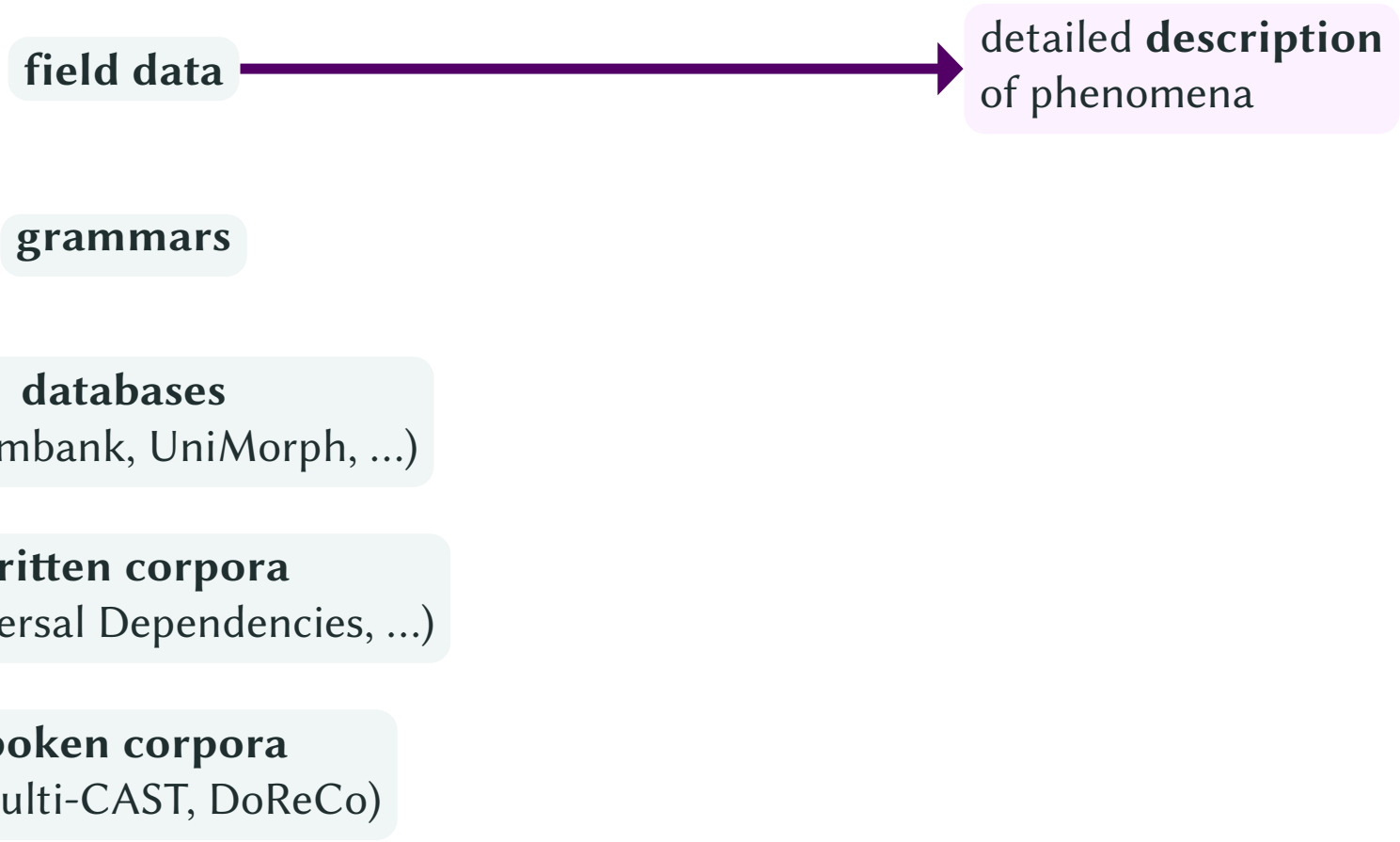
databases
(e.g. Grambank, UniMorph, ...)

written corpora
(e.g. Universal Dependencies, ...)

spoken corpora
(e.g. Multi-CAST, DoReCo)

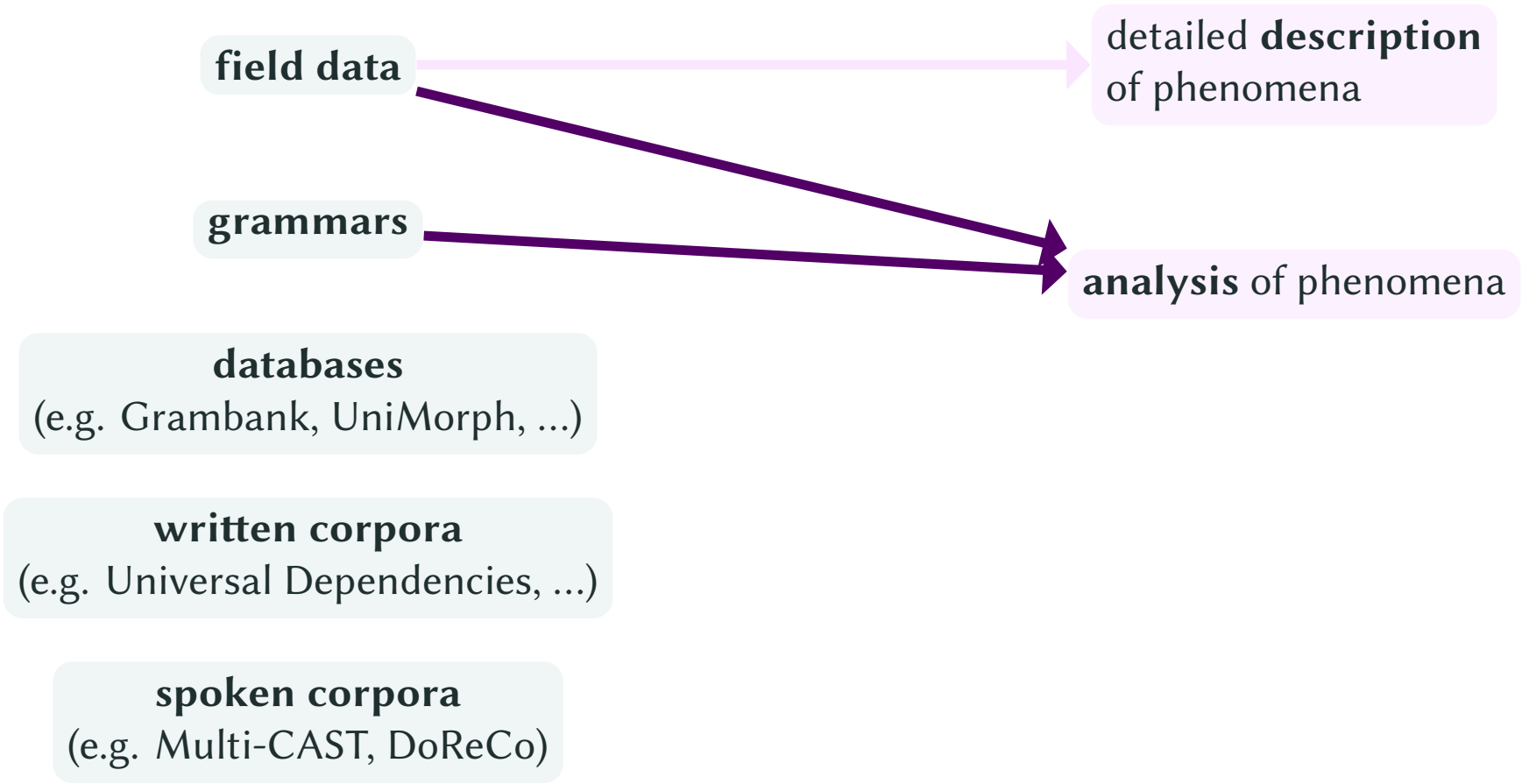
Approaches in typology

Typology: the study of variation and trends of linguistic phenomena in the world's languages



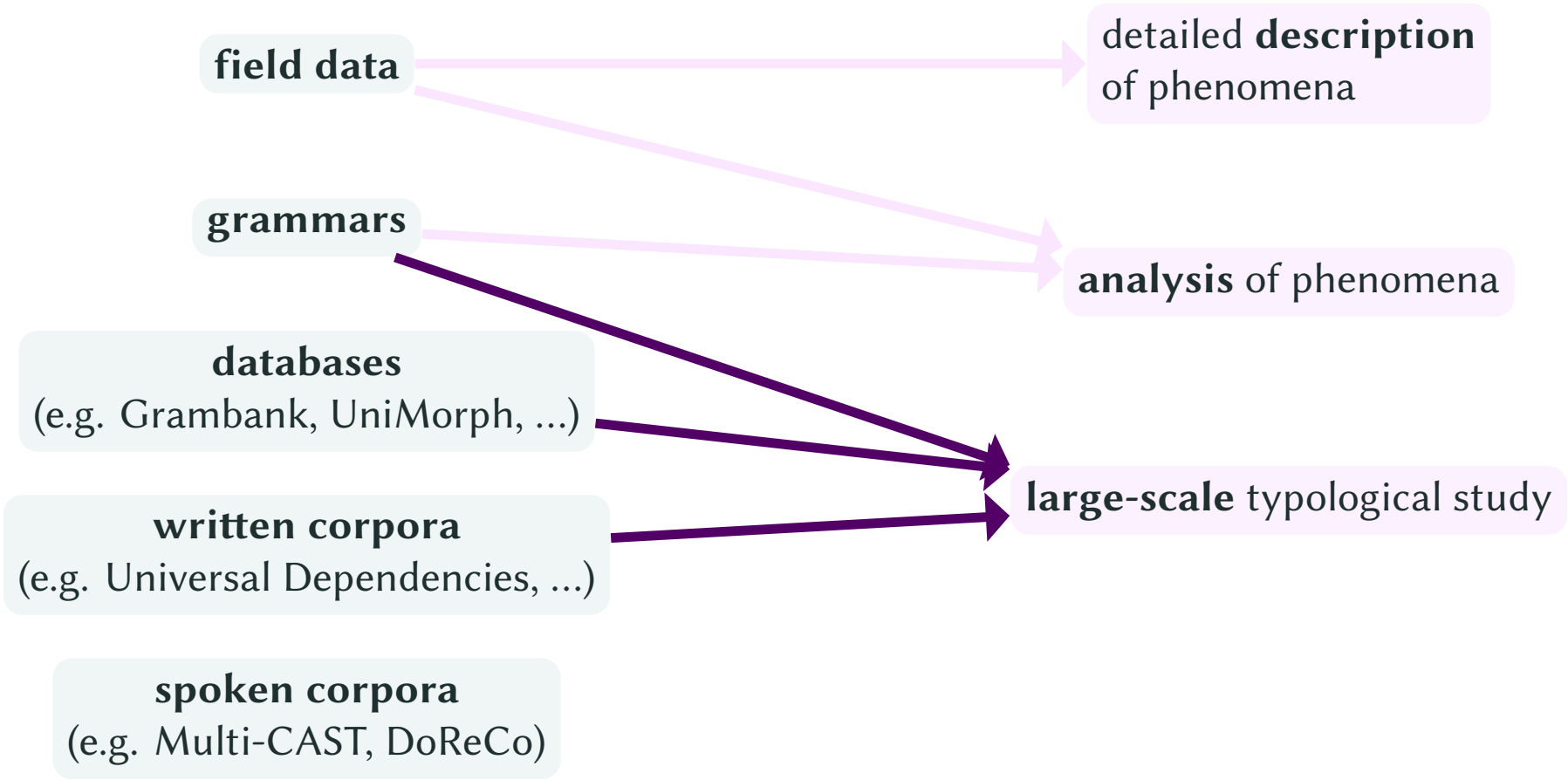
Approaches in typology

Typology: the study of variation and trends of linguistic phenomena in the world's languages



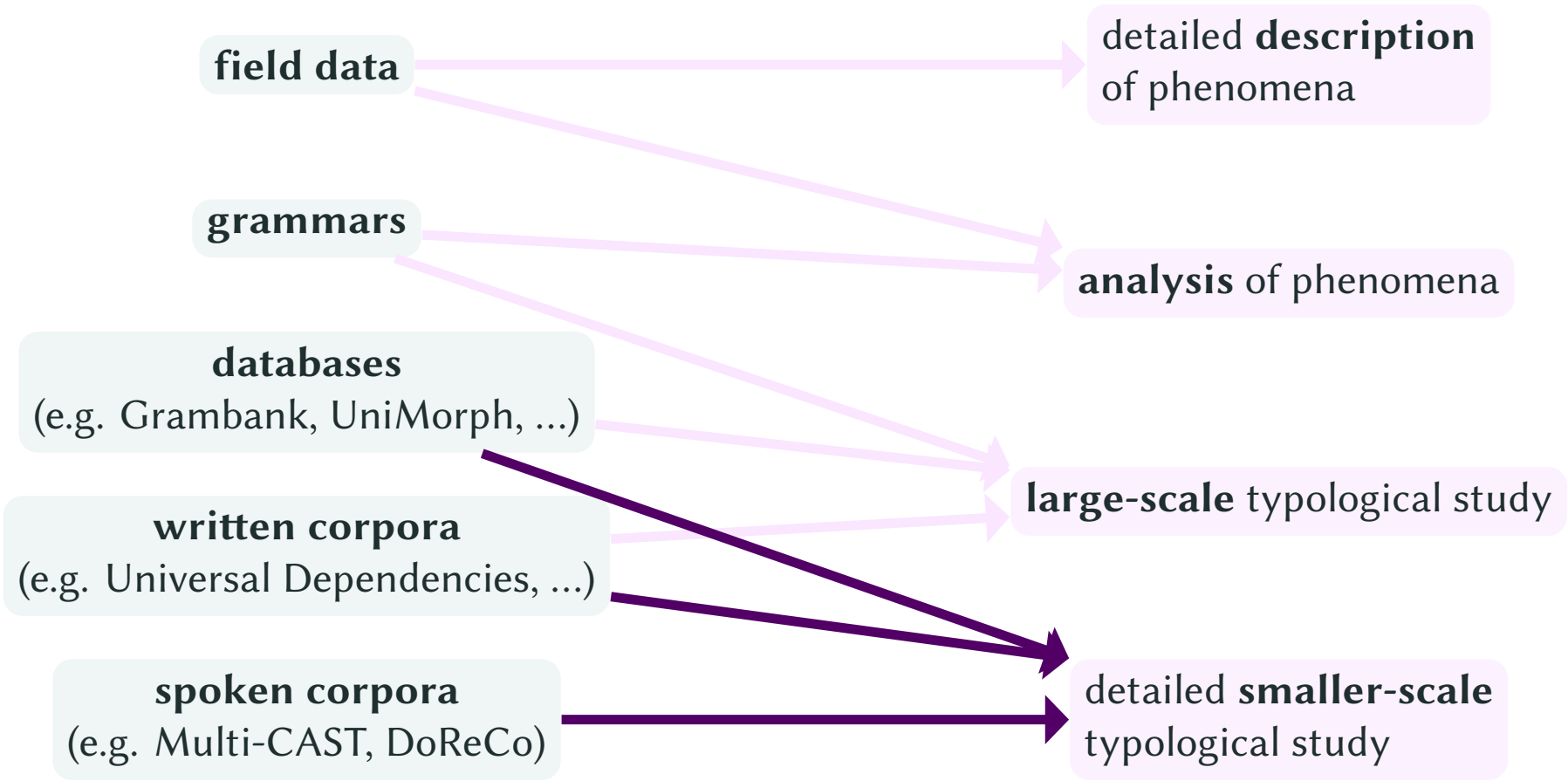
Approaches in typology

Typology: the study of variation and trends of linguistic phenomena in the world's languages



Approaches in typology

Typology: the study of variation and trends of linguistic phenomena in the world's languages



Approaches in typology

Typology: the study of variation and trends of linguistic phenomena in the world's languages

Spoken typology

- analyze spontaneous speech data from typologically diverse languages
  - ☺ capture crosslinguistic **diversity** ⟷ better **control** of the data
  - ☺ capture **variation** in language use **within** languages
  - ☺ include **phonetic & prosodic** information

Today's talk

1. frequency effects with verbal person marking
2. pausing and syntactic boundaries: prosody-syntax mismatches
3. clitics vs. affixes: testing for phonetic differences

❶ Frequency effects with verbal person marking

Frequency effects in grammar

more frequent grammatical expressions tend to be **shorter** than comparable less frequent grammatical expressions

frequency		**length**
more frequent	>	shorter
less frequent	>	longer

examples of frequency effects in grammar:

- singular < plural (Greenberg 1966)
- present/past tense < future (Greenberg 1966; Bybee et al. 1994)
- allative < ablative (Stolz et al. 2014)
- dependent POSS < independent POSS (Ye 2020)
- REFL with introverted verbs < extroverted verbs (Haiman 1983)
- direct < indirect causation (Levshina 2022)

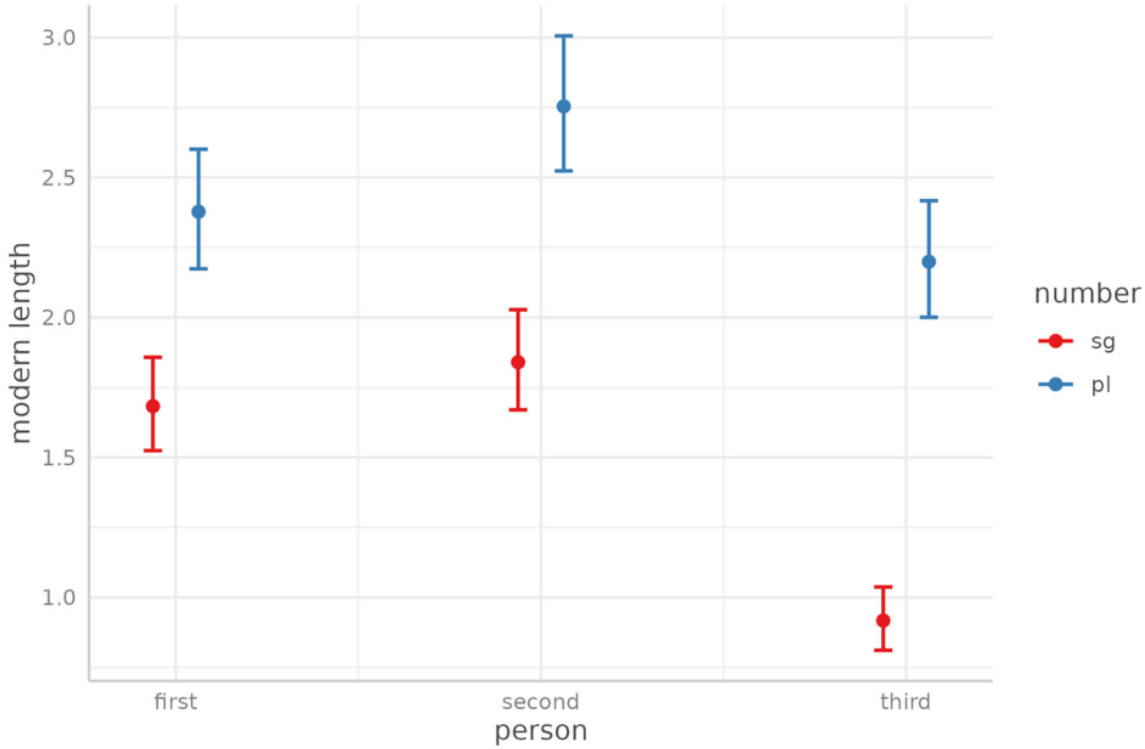
Frequency effecs in verbal person marking

Length of verb forms (Greenberg 1966)

- 3 < 1, 2
- SG < PL (and DU)

Length of verbal person markers (Seržant and Moroz 2022)

- verbal subject markers in 383 languages



Open questions

- Is the number of segments a sufficient approximation for “length”?
- What about other systems / types of person markers?
- What about variation within languages?
- How strong is the association between frequency and marker length?

Data

- DoReCo (Seifart et al. 2022): spontaneous speech data from 51 languages
- time-aligned grammatical annotation
- today: data from **8 languages**

Examples

Some languages have very consistent exponents for person marking.

Kamas, Uralic (Gusev et al. 2022)

(1) /ertən uptəbiam/ ▶

Erte-n uʔbdə-bi-**am**.
morning-ADV get.up-PST-**1SG**
'In the morning I got up.'

(2) /mɐn mbinʲibuʒʲ tnan detl**em**/ ▶

Măn ĭmbi-n'ibud' tănan det-le-**m**.
I.NOM what-INDEF you.DAT bring-FUT-**1SG**
'I will bring you something.'

Examples

Verbal person markers can express combinations of person values (scenarios).

Texistepec Popoluca, Mixe-Zoque (Wichmann 2022)

(3) /ɛɲtʃ ɛːtʃɛ **k**wata?/ ▶

eñch eeche **kn**-wat-a?
no nothing **1>2**-do-APPL
'I am not doing anything to you.'

(4) /ma? **kŋ**gjɔkska? djim/ ▶

Ma? **knky**-oks-ka? dy-im.
PFV **2>1**-punch-INT 3:A-say
'You punched me, he said.'

Examples

Verbal person markers can co-express values of other grammatical categories.

Nafsan, Austronesian (Thieberger 2022)

(5) /**i**mu:r na **ke**kinki tefla/ ▶

i=muur nag **ke**=kin-ki tefla

3SG.REAL=want REL **3SG.IRR**=pinch-TR similar

'he wanted to pinch us like that'

Measures

Length

- duration in ms
- segmental length (number of phones)
- segmental shape (10 most common types of vowel-consonant sequences)

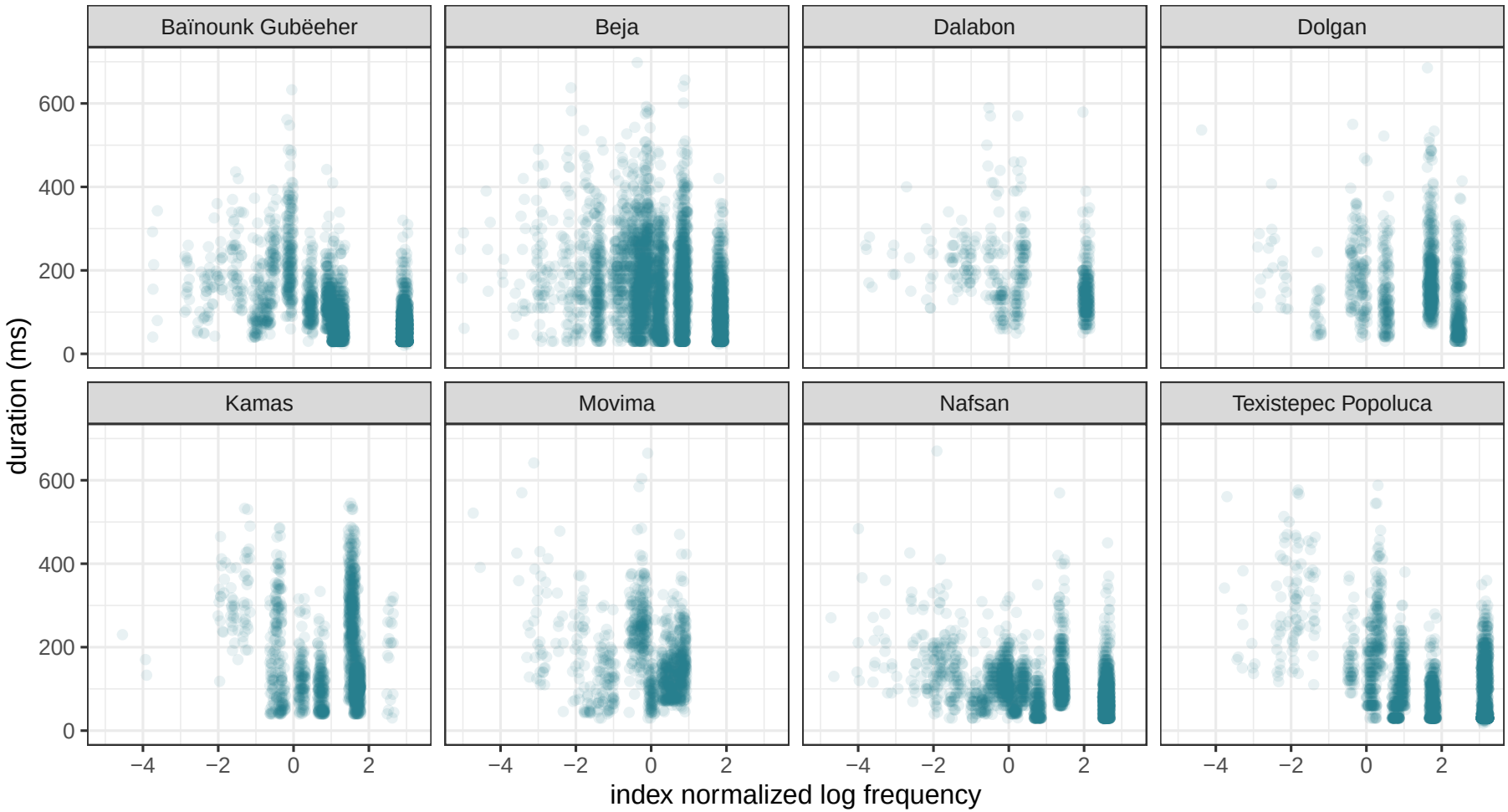
Distribution

- token frequency, normalized across corpora
  (log frequency of the markers per 100 words)

Controls

- speech rate (phonemes / sec)
- position of the marker in the word (and in relation to silent pauses)
- semantic complexity (number of grammatical values expressed)

Marker frequency and duration



Modeling the association between frequency and duration

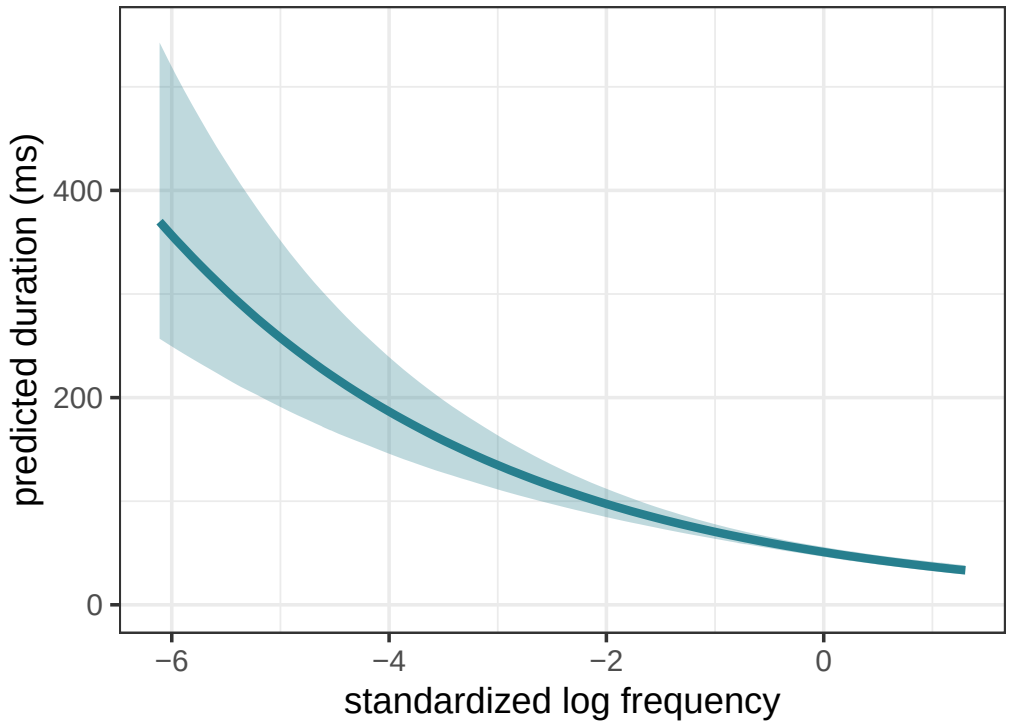
We want to predict the **duration** of person markers from their **frequency**.

We also want to account for the effects of

- frequency by language (and variation across speakers)
- speech rate
- marker position
- marker semantic complexity
- individual speakers

I fitted Bayesian gamma regression models using *Stan* and *brms* in *R*.

Conditional effects of frequency on marker duration



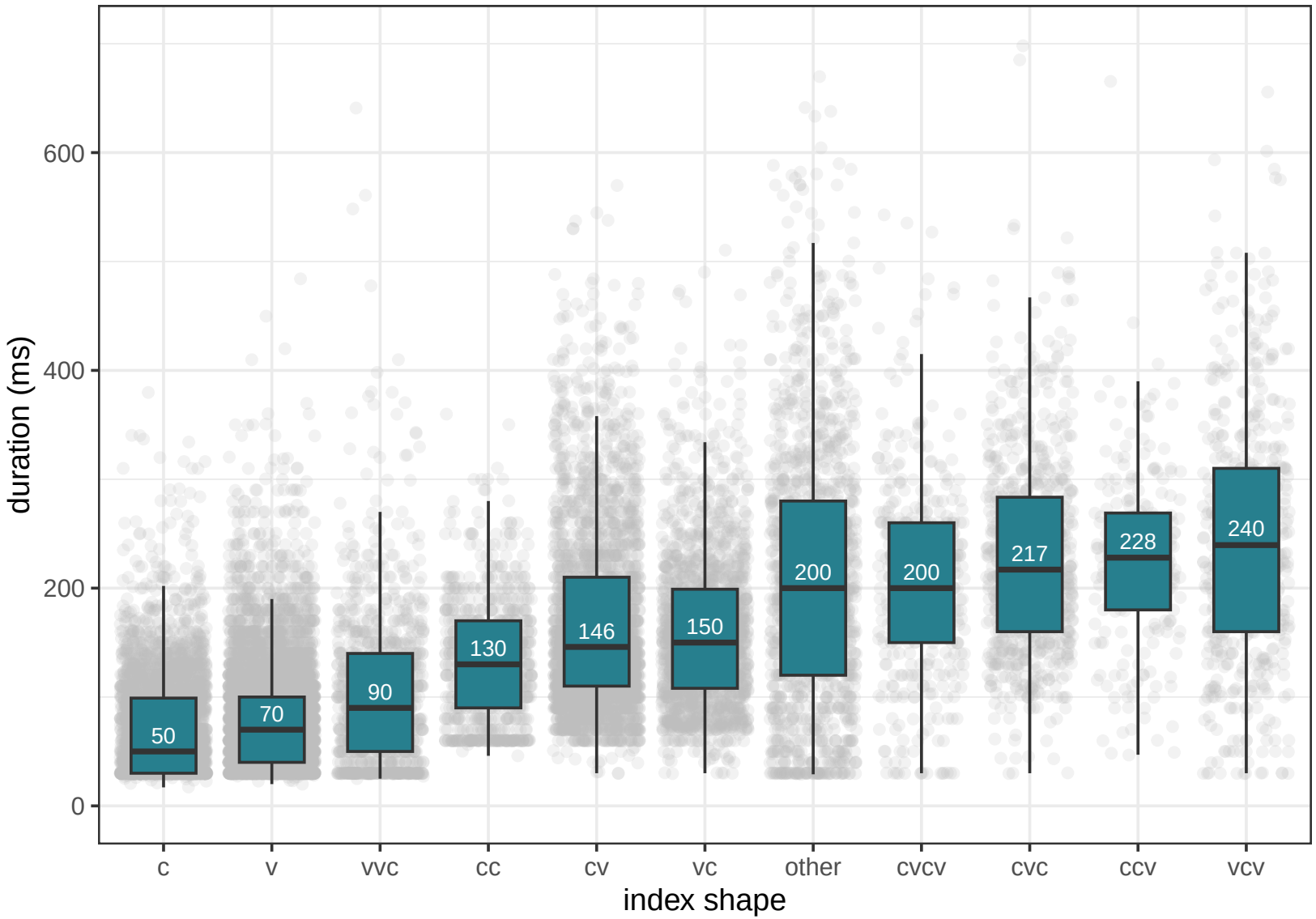
☺ more frequent markers tend to have shorter durations

Segmental shapes

- I annotated the markers for their segmental shapes:
  length of the marker + information on vowels and consonants

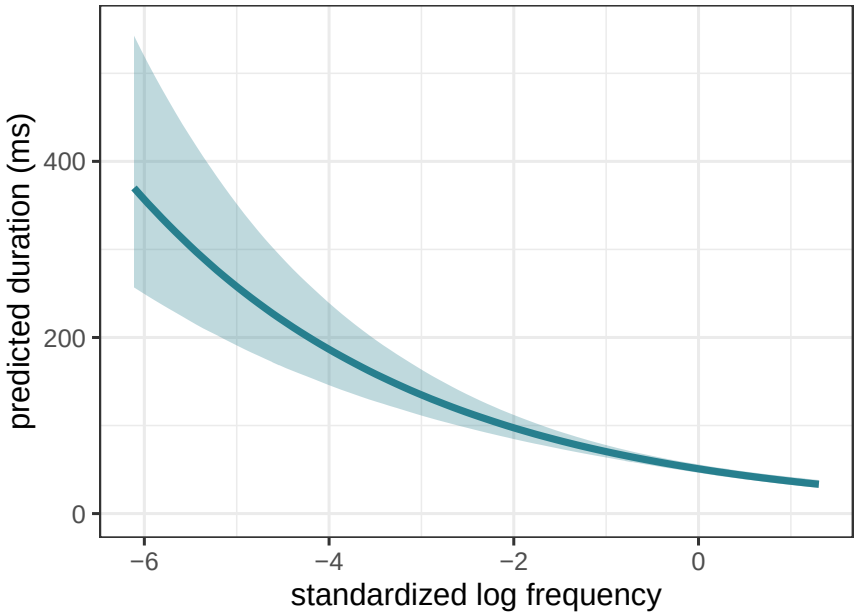
shape	N occurrences
v	5,939
c	3,572
cv	2,643
vc	1,694
other	1,334
vvc	957
cc	813
cvc	743
vcv	518
cvcv	408
ccv	223

Segmental shapes and duration

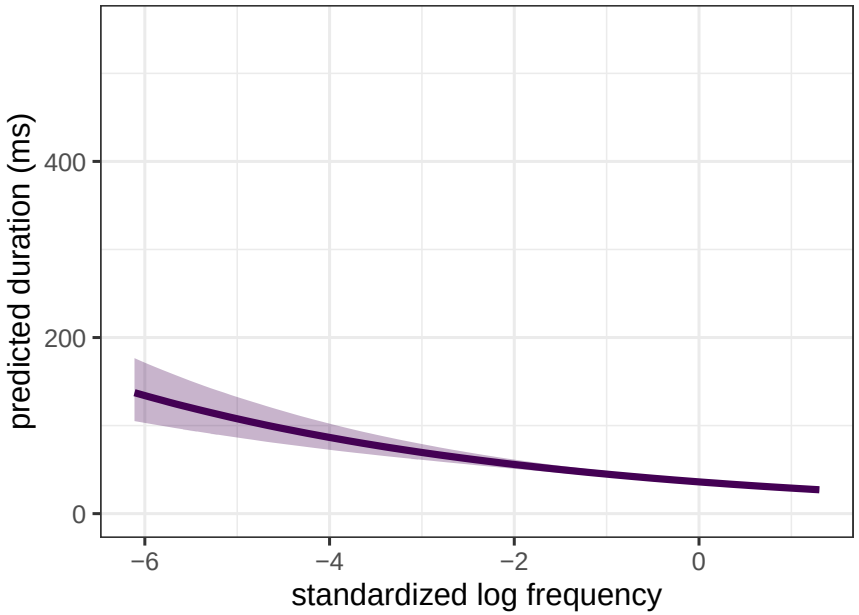


Conditional effects of frequency on marker duration

no control for segmental shapes



controlling for segmental shapes



- ✔ frequency still has a weak effect on marker duration
- ➔ direct frequency effects in terms of probabilistic variation in language use

❷ Pausing and syntactic boundaries

(joint work with Naomi Peck)

Pauses

What is a pause?

- a **temporary break** in the flow of **speech** or sign
- **silent** and/or **filled**
- a mechanism to enable **processing**, **planning** and **structuring** within interaction

Where do pauses occur?

- at higher syntactic boundaries
  (Goldman-Eisler 1968; Cooper and Paccia-Cooper 1980; Nespor and Vogel 2007)

- at higher prosodic boundaries / signal of prosodic boundaries
  (Gee and Grosjean 1983; Ferreira 1993; Himmelmann et al. 2018; Himmelmann 2022; Ladd 2008)

- between relevant semantic units
  (O'Connell et al. 1969; Ferreira 1993)

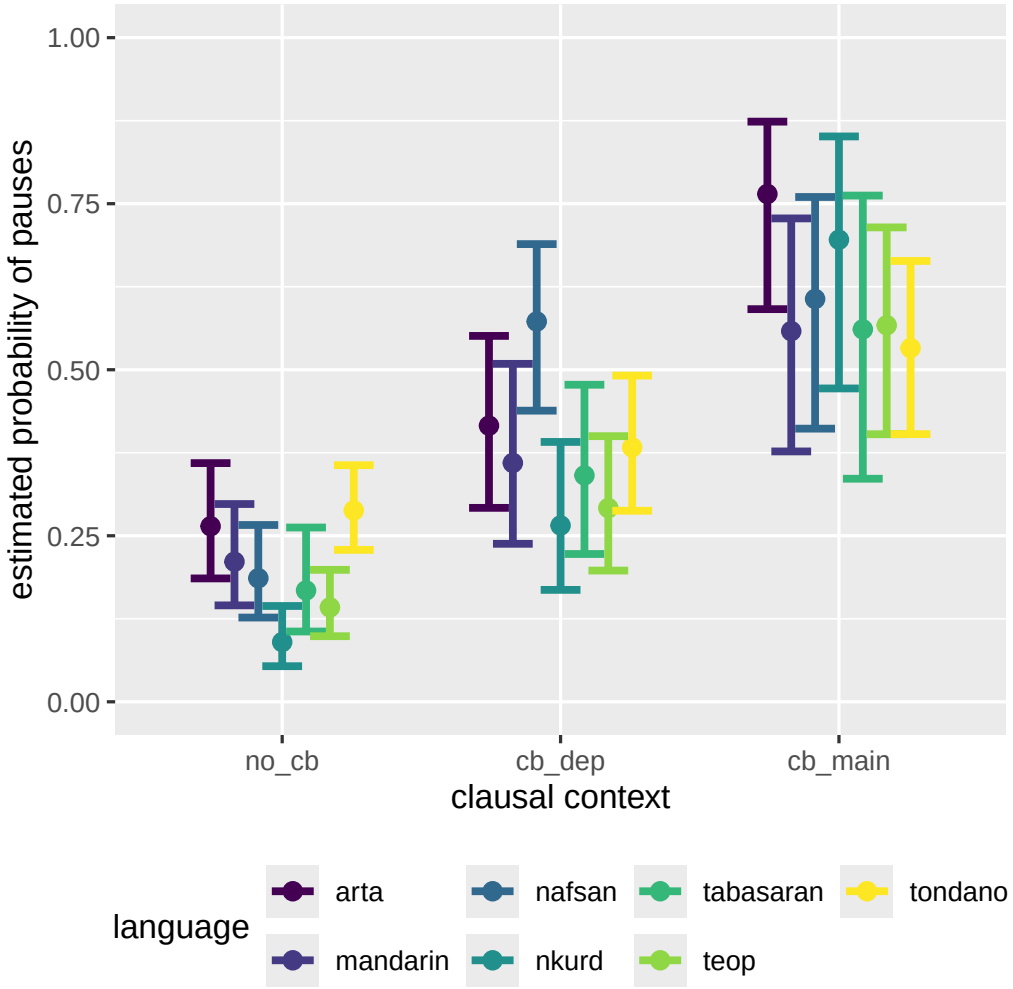
- word internally
  (Paschen 2025; Evans et al. 2008; Bundgaard-Nielsen and Baker 2020; Mansfield 2023)

Pausing at clause boundaries in spontaneous speech

Peck and Becker (2024)

probability of pauses

clause-internal boundary $<$ dep. clause boundary $<$ main clause boundary



Pausing at clause boundaries in spontaneous speech

Examples from **Teop**, Austronesian (Mosel and Schnell 2015)

- **Pause at the relative clause boundary**

(6) [a taba na karakaras nana] **<p>** [**to** kokopo nana nahu,]
[ART thing TAM rustle 3SG.IPFV] **<p>** [**REL** cover 3SG.IPFV pot]
'The thing was rustling ... (that thing) that was covered in the pot,'

- **No pause at the relative clause boundary**

(7) [tara vai nao bona <p> a iana] [**to** nomaa bata=na.]
[see now thither OBJ.ART <p> ART.SG fish] [**REL** come CONT=3SG.IPFV]
'They now saw the ... the fish that was coming along.'

- **Intervening pause following the relativizer**

(8) [Es roosuu] [**to** **<p>** va-mee bata=ri te=a kave vai.]
[ART giant] [**REL** **<p>** MULT-lead CONT=3PL.OBJM PREP=ART.SG net DEM]
'It was the giant who ... led (them) to this kave.'

Prosody-syntax mismatches

- some function words show mismatches between their prosodic and syntactic alignment (ditropic clitics)

(9)

SYNTAX	*I*		*'m*	=	*happy*
PROSODY	*I*	=	*'m*		*happy*

- such mismatches are known for auxiliaries, prepositions, determiners, argument indexes
  (Cysouw 2005; Gussenhoven and Jacobs 1998; Spencer and Luís 2012)

- they most likely originate from pausing between a preposed function word and the following lexical word (Himmelmann 2014)

Prosody-syntax mismatches

Intervening pauses leading to prosody-syntax mismatches are theoretically interesting.

- they may lead to a preference for encliticisation

→ preference for **suffixation** (Himmelmann 2014)

✓ **GRAM LEX → GRAM <p> LEX**

- grammatical words have higher frequencies and a lower processing load and can be uttered quickly
- uttering the grammatical word with a following pause allows the speaker to hold the floor in conversation

✗ **LEX GRAM → LEX <p> GRAM**

- no need for the pause

Evidence for GRAM <p> LEX?

Himmelmann (2014) presents evidence for intervening pauses with preposed grammatical markers:

- **English** def. article, prepositions and auxiliaries **10-20%**
- **English** def. article and prepositions **∼7%**
- **German** articles and auxiliary **9%**
- **Tagalog** article, genitive, locative and plural marker **6-27%**

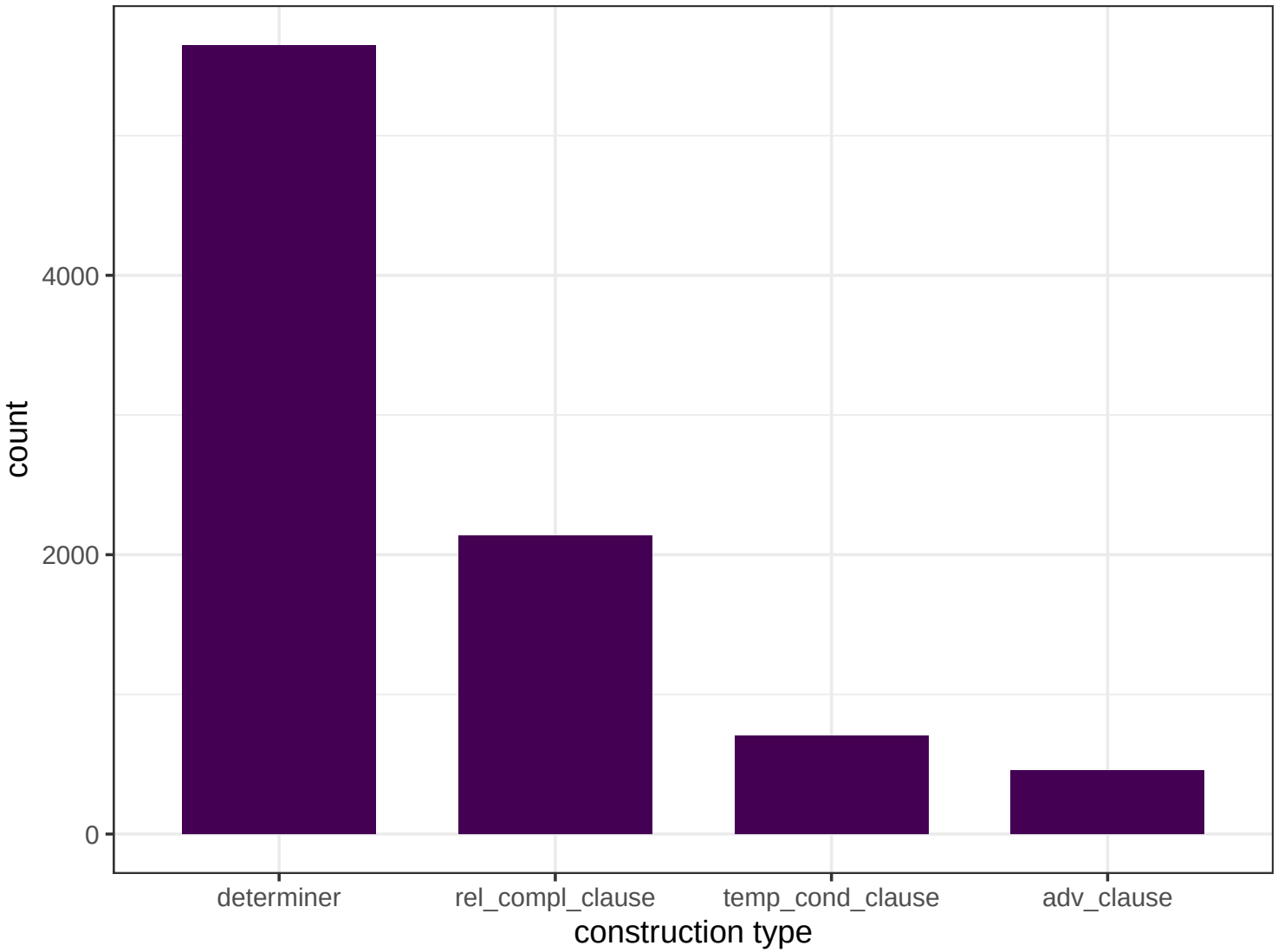
But how common are they cross-linguistically

- ? at the phrasal level?
- ? at the clause level?

Dataset

We analyse **spontaneous speech** data from **11 languages** in DoReCo (Seifart et al. 2024).

Construction types



Examples: Determiner

(10) **Arapaho**, Algic ▶ (Cowell 2024)

niiyou <p> nii-neyei-hi'-eeneti-'niiyou [**nuhu'** **<p>**
here.it.is <p> IMPERF-try-INSTR-speak-1PL.here.it.is [**this** **<p>**
hinono'ein-eenetiit]
Arapaho-language]
'We are trying to talk about this Arapaho way of speaking.'

(11) **Vera'a**, Austronesian ▶ (Schnell 2024)

vuvuluk mu-n [**e** **<p>** **'e** reñe anē=m] <p>
friend POSS.GEN-CS [**PERS.ART** **<p>** **HES** woman DEM1.A=TAM1] <p>
van ma
go hither
'a close friend of this woman came,'

Examples: Relative clause

(12) **Goemai**, Afro-Asiatic (Hellwig 2024)

ji lang ji wa n-pe [**goe-pe**
LOGS.SGM.S hang/move LOGS.SGM.S return_home LOC-place **NOMZ-COM**
<p> t'ong ji kat long yi]
<p> IRR LOGS.SGM.S find chief SUB]
'He moves (and) returns home to the place where he would get the chieftancy.'

(13) **Komnzo**, Yam ▶ (Döhler 2024)

zane <p> kwosi [**mane** **<p>** zfrä]
DEM:PROX <p> dead [**which** **<p>** 3SG.FEM:SBJ:RPST:IPFV.be]
'the one that died'

Examples: Temporal and conditional clauses

(14) **Komnzo**, Yam ▶ (Döhler 2024)

yəmarwəre [**fəthə** **<p>** fä ŋaritakwər nima]
1PL>3SG.M.see.NPST.IPFV [**when** **<p>** DIST 3SG.cross.NPST.IPFV like.this]
'we see him when he walks across there'

(15) **Movima** (Haude 2024)

do:koy [**di'** **<p>** te:lo] <p> it te:lo łat jankwa
good [**HYP** **<p>** dance] <p> 1INTR dance EVD say
OK, if you dance, I dance, I said.

Examples: Other adverbial clauses

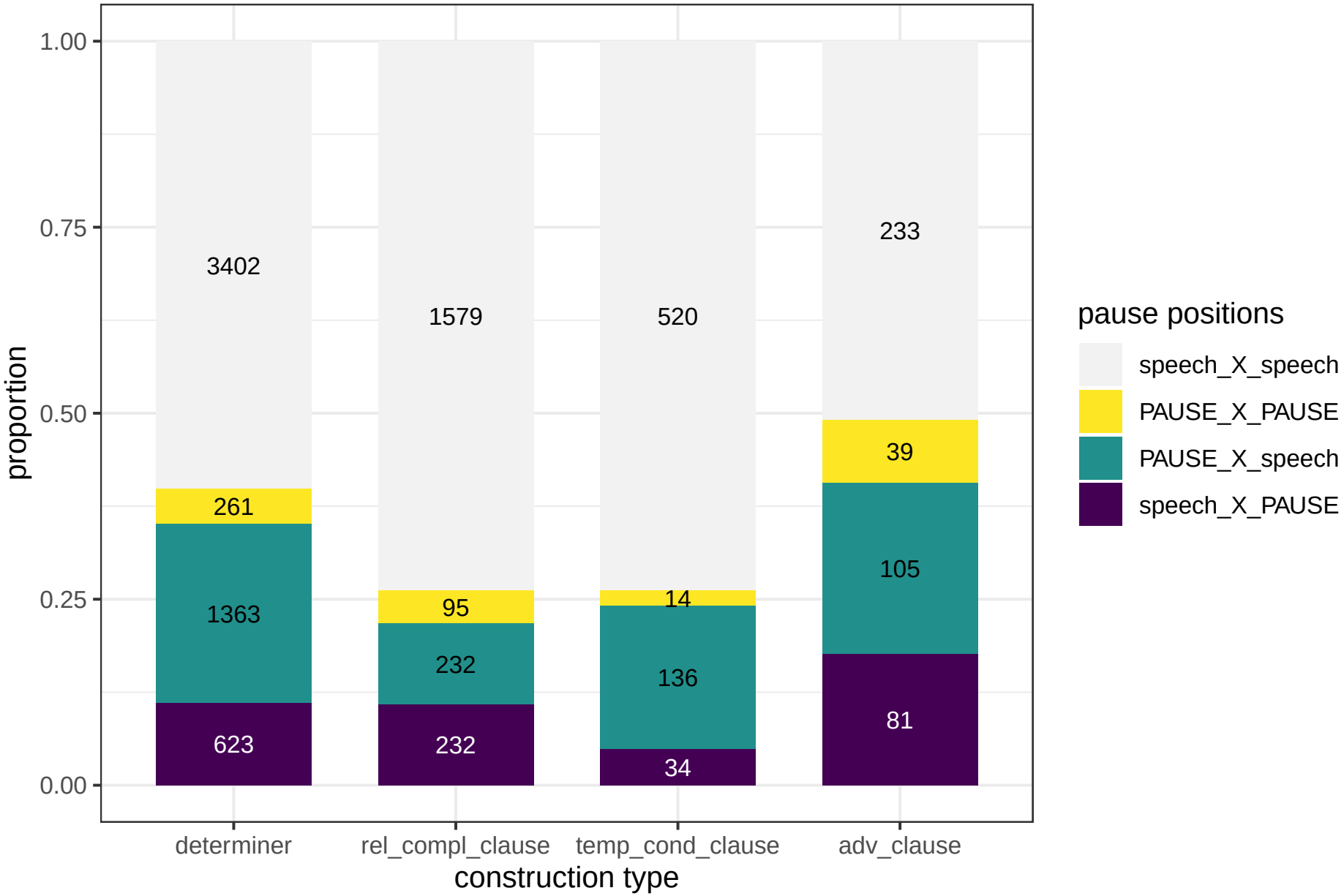
(16) **Sümi**, Sino-Tibetan ▶ (Teo 2024)

a-zükutunu=pu ghi a-zükutunu=pu shi-ve [**ku**
NRL-be.funny=CONN even NRL-be.funny=CONN happen-VM [**what**
ghengu **<p>** ningu kümtsü gha khumu mtha=mu]
because **<p>** 1PL all ADV nothing NEG.know=even.though]
'it became funny, it became funny because none of us knew anything'

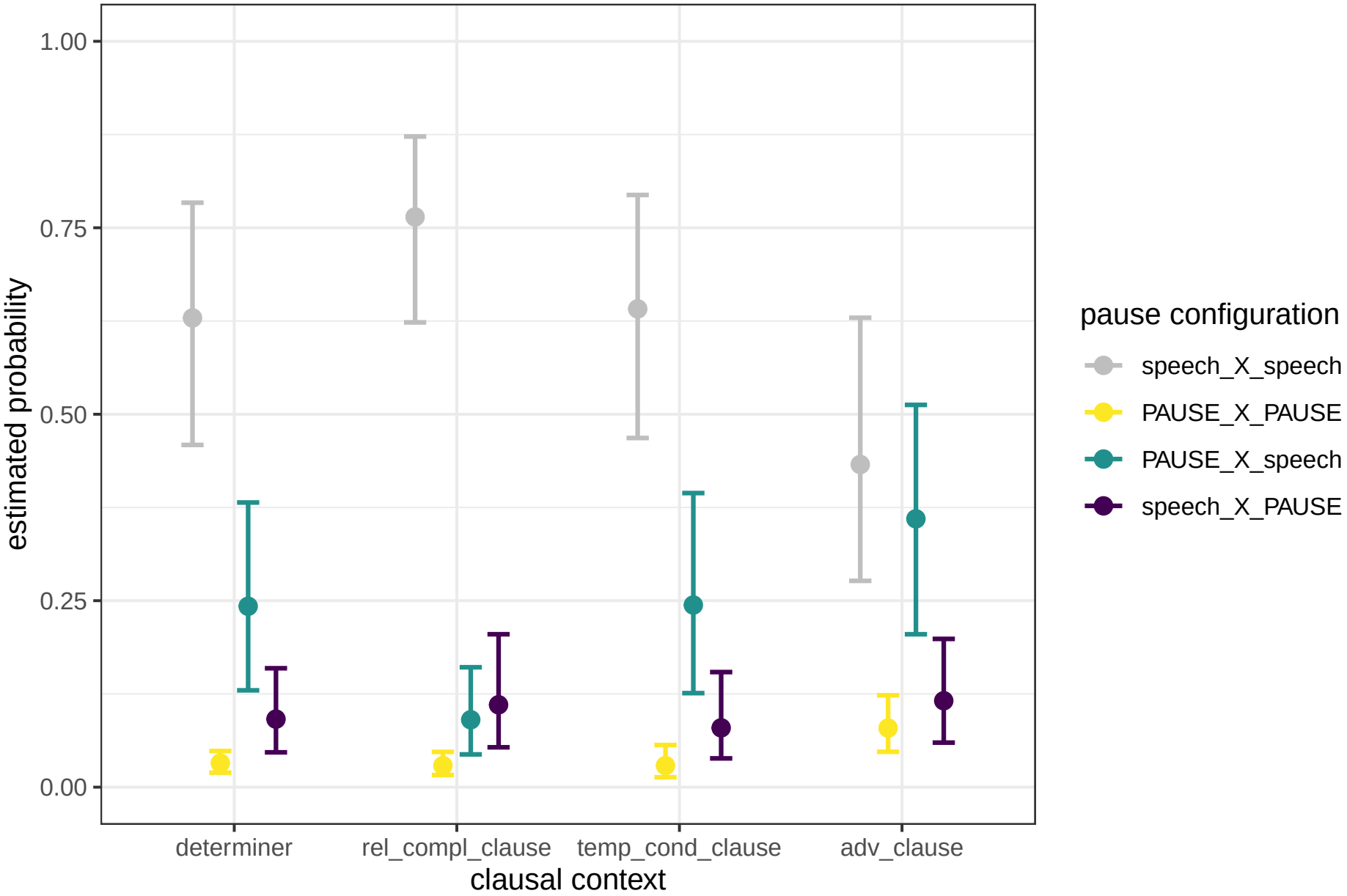
(17) **Arapaho** (Cowell 2024)

noohoot-oo [**toh-** **<p>** **toh-** **<p>** toh-nee'ees-iine'etii-t niine'eehek
see-0.IMPER [**since-** **<p>** **since-** **<p>** since-thus-live-3.s here_is.3s
nehe' hinono'ei]
this Arapaho
'See/look at it, because that is how these here Arapahos live.'

Results: Pause distribution



Model results: Pause distribution



❸ Clitics vs. affixes

(future joint work with Tim Zingler & Phillip Rogers)

Background

Assumptions in the grammaticalization literature

- clitics become affixes in the course of grammaticalization processes
- this is accompanied by phonetic and then phonological reduction

➔ clitics should be longer than affixes

Problem(s)

? should we distinguish between clitics and affixes (especially across languages)?

? if so, are clitics really systematically longer than affixes?

Zingler and Rogers (To appear)

These questions are addressed in Zingler and Rogers (To appear).

- quantitative cross-linguistic study with a sample of **36 languages** (using reference grammars)
- measured length in
  - N syllables
  - N phonemes

Findings

- **clitics** are significantly more likely to be **monosyllabic**
- **proclitics** tend to be **slightly shorter** than enclitics (not significant)
- **similar phonological length** between clitics and affixes

A spoken typology approach to clitics vs. affixes

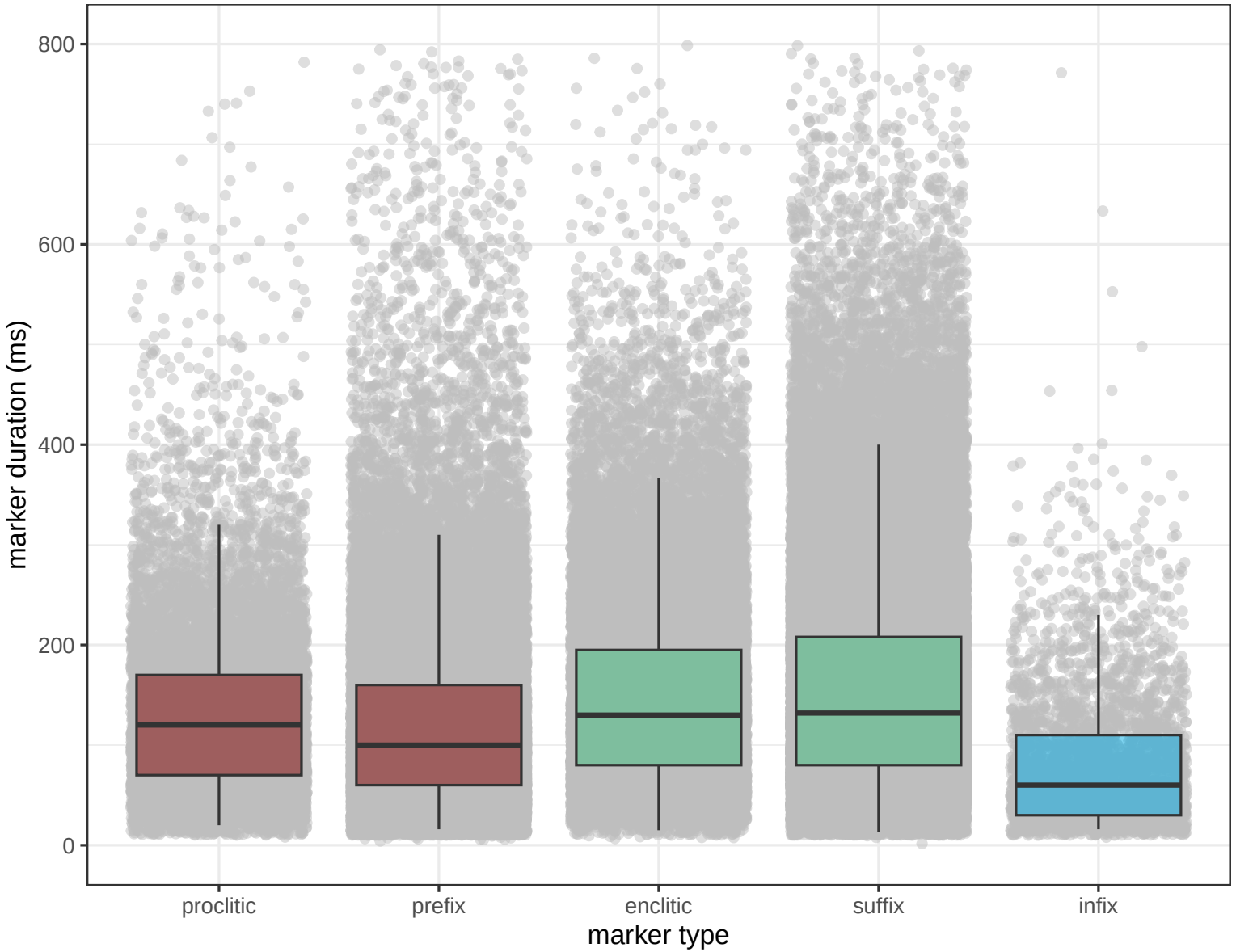
⚠ Work in progress

- check for systematic differences between clitics and affixes with spontaneous speech corpora
- use the original classification into clitics and affixes
- check for phonetic and phonological differences between
  - affixes vs. clitics
  - preposed vs. postposed elements
  - markers in the nominal vs. verbal domain
  - ...

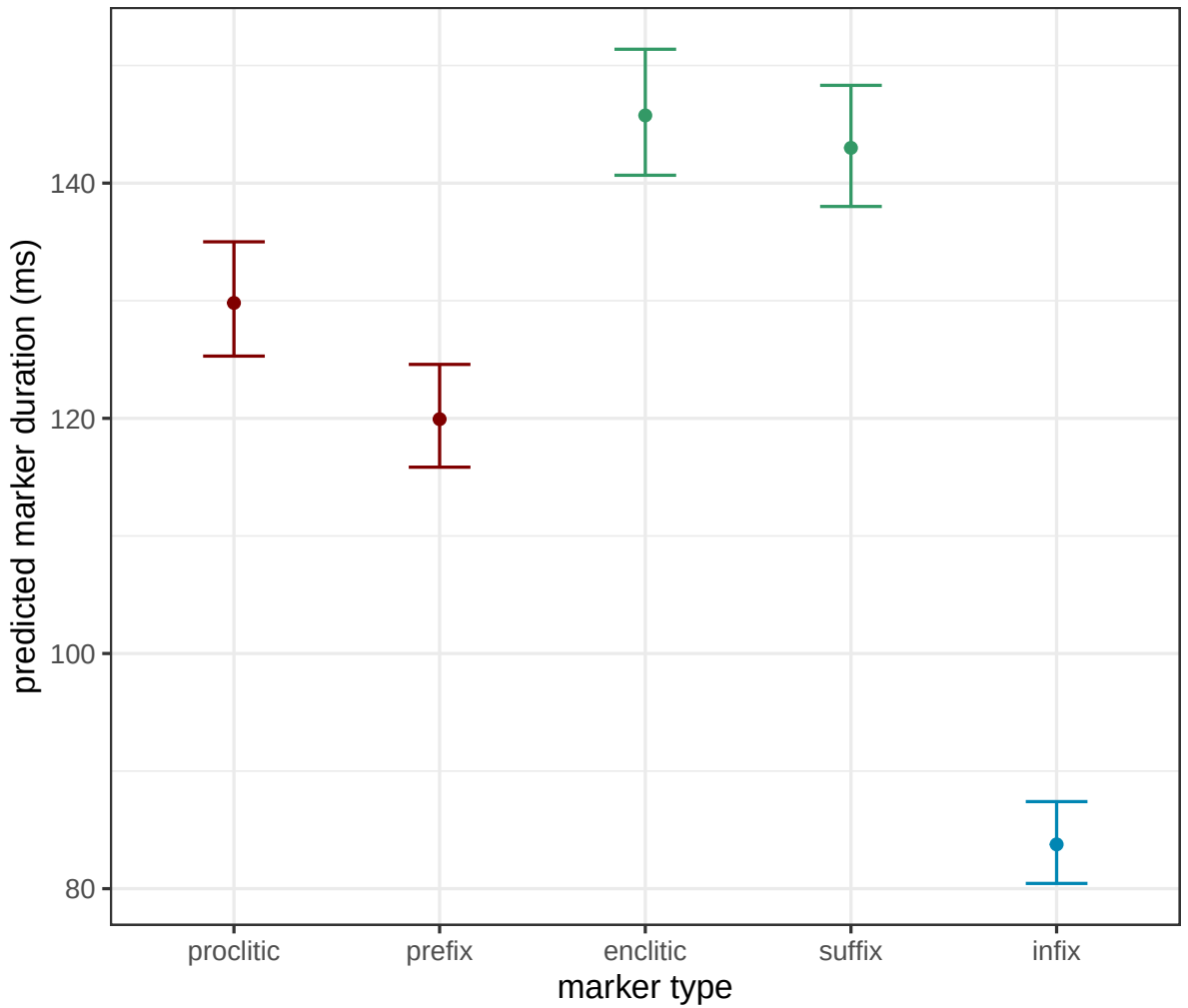
Dataset

We have started working with data from **35 languages** in DoReCo (Seifart et al. 2024).

Clitic and affix durations



Conditional effects of marker type



(controlling for speech rate, speakers and phylogenetic relations between languages)

Conclusion: Spoken typology

Smaller corpora of spontaneous speech can provide new insights to old questions.

- **token-based**

  we can capture intra-language variation

- **spoken data**

  we can integrate phonetic properties and make more fine-grained distinctions

➔ We come closer to a more realistic account of grammar in language use.

Thank you!

References I

Rikke L. Bundgaard-Nielsen and Brett J. Baker. "Pause Acceptability Indicates Word-Internal Structure in Wubuy." In: *Cognition* 198 (2020), p. 104167.

Joan Bybee, Revere Perkins, and William Pagliuca. *The Evolution of Grammar. Tense, Aspect, and Modality in the Languages of the World.* Chicago: The University of Chicago Press, 1994.

Alexander Yao Cobbinah. "Baïnounk Gubëeher DoReCo Dataset." In: *Language Documentation Reference Corpus (DoReCo) 1.2*. Ed. by Frank Seifart, Ludger Paschen, and Matthew Stave. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), 2022.

William E. Cooper and Jeanne Paccia-Cooper. *Syntax and Speech.* Cambridge, MA: Harvard University Press, 1980.

Andrew Cowell. "Arapaho DoReCo Dataset." In: *Language Documentation Reference Corpus (DoReCo) 2.0*. Ed. by Frank Seifart, Ludger Paschen, and Matthew Stave. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), 2024.

Michael Cysouw. "A Survey of Preposed Enclitics: Morphology in the Wrong Place." In: *Morphology and Its Demarcations: Selected Papers from the 11th Morphology Meeting, Vienna, February 2004*. Ed. by Wolfgang U. Dressler et al. Amsterdam: Benjamins, 2005, pp. 17–37.

Christian Döhler. "Komnzo DoReCo Dataset." In: *Language Documentation Reference Corpus (DoReCo) 2.0*. Ed. by Frank Seifart, Ludger Paschen, and Matthew Stave. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), 2024.

Nicholas Evans, Janet Fletcher, and Belinda Ross. "Big Words, Small Phrases: Mismatches between Pause Units and the Polysynthetic Word in Dalabon." In: 46.1 (2008), pp. 89–129.

References II

Fernanda Ferreira. "Creation of Prosody during Sentence Production." In: *Psychological Review* 100.2 (1993), pp. 233–253.

James Paul Gee and François Grosjean. "Performance Structures: A Psycholinguistic and Linguistic Appraisal." In: *Cognitive Psychology* 15 (1983), pp. 411–458.

Frieda Goldman-Eisler. *Psycholinguistics: Experiments in Spontaneous Speech.* London: The Academic Press, 1968.

Joseph Greenberg. *Language Universals: With Special Reference to Feature Hierarchies.* The Hague: Mouton, 1966.

Valentin Gusev et al. "Kamas DoReCo Dataset." In: *Language Documentation Reference Corpus (DoReCo) 1.2*. Ed. by Frank Seifart, Ludger Paschen, and Matthew Stave. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), 2022.

Carlos Gussenhoven and Haike Jacobs. *Understanding Phonology.* London: Arnold, 1998.

John Haiman. "Iconic and Economic Motivation." In: *Language* 59.4 (1983), pp. 781–819.

Katharina Haude. "A Grammar of Movima." Nijmegen: Radboud Universiteit, 2006.

Katharina Haude. "Movima DoReCo Dataset." In: *Language Documentation Reference Corpus (DoReCo) 2.0*. Ed. by Frank Seifart, Ludger Paschen, and Matthew Stave. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), 2024.

Birgit Hellwig. "Goemai DoReCo Dataset." In: *Language Documentation Reference Corpus (DoReCo) 2.0*. Ed. by Frank Seifart, Ludger Paschen, and Matthew Stave. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), 2024.

References III

Nikolaus Himmelmann. "Asymmetries in the Prosodic Phrasing of Function Words: Another Look at the Suffixing Preference." In: *Language* 90.4 (2014), pp. 927–960.

Nikolaus Himmelmann. "Prosodic Phrasing and the Emergence of Phrase Structure." In: *Linguistics* 60.3 (2022), pp. 715–743.

Nikolaus Himmelmann et al. "On the Universality of Intonational Phrases: A Cross-Linguistic Interrater Study." In: *Phonology* 35.2 (2018), pp. 207–245.

Robert Ladd. *Intonational Phonology*. Cambridge: Cambridge University Press, 2008.

Natalia Levshina. *Communicative Efficiency: Language Structure and Use*. Cambridge: Cambridge University Press, 2022.

John Mansfield. "The Prosodic Structure of Australian Polysynthetic Verbs: Bininj Gun- Wok, Murrinhpatha and Ngalakgan." In: *Word Prominence in Morphologically Complex Languages*. Ed. by Ksenia Bogomolets and Harry van der Hulst. Oxford: Oxford University Press, 2023, pp. 411–438.

Ulrike Mosel and Stefan Schnell. "Multi-CAST Teop." In: *Multi-CAST. Multilingual Corpus of Annotated Spoken Texts*. Ed. by Geoffrey Haig and Stefan Schnell. Bamberg: University of Bamberg, 2015.

Marina Nespor and Irene Vogel. "Prosodic Phonology: With a New Foreword." In: *Prosodic Phonology*. De Gruyter Mouton, 2007.

Daniel C. O'Connell, Sabine Kowal, and Hans Hörmann. "Semantic Determinants of Pauses." In: *Psychologische Forschung* 33.1 (1969), pp. 50–67.

Ludger Paschen. "The Ubiquity of Word-Internal Pauses." Berlin, 2025.

Naomi Peck and Laura Becker. "Syntactic Pausing? Re-examining the Associations." In: *Linguistics Vanguard* 10.1 (2024), pp. 223–237.

References IV

Uta Reinöhl and Antje Casaretto. "When Grammaticalization Does NOT Occur: Prosody-syntax Mismatches in Indo-Aryan." In: *Diachronica* 35.2 (2018), pp. 238–276.

Stefan Schnell. "Vera'a DoReCo Dataset." In: *Language Documentation Reference Corpus (DoReCo) 2.0*. Ed. by Frank Seifart, Ludger Paschen, and Matthew Stave. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), 2024.

Frank Seifart, Ludger Paschen, and Matthew Stave, eds. *Language Documentation Reference Corpus (DoReCo) 1.2*. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), 2022.

Frank Seifart, Ludger Paschen, and Matthew Stave, eds. *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), 2024.

Ilja Seržant and George Moroz. "Universal Attractors in Language Evolution Provide Evidence for the Kinds of Efficiency Pressures Involved." In: *Humanities and Social Sciences Communications* 9.1 (2022), pp. 1–9.

Andrew Spencer and Ana Luís. *Clitics: An Introduction*. Cambridge: Cambridge University Press, 2012.

Thomas Stolz, Sander Lestrade, and Christel Stolz. *The Crosslinguistics of Zero-Marking of Spatial Relations*. Berlin: De Gruyter, 2014.

Amos Teo. "Sümi DoReCo Dataset." In: *Language Documentation Reference Corpus (DoReCo) 2.0*. Ed. by Frank Seifart, Ludger Paschen, and Matthew Stave. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), 2024.

References V

Nick Thieberger. "Nafsan (South Efate) DoReCo Dataset." In: *Language Documentation Reference Corpus (DoReCo) 1.2*. Ed. by Frank Seifart, Ludger Paschen, and Matthew Stave. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), 2022.

Søren Wichmann. "Texistepec Popoluca DoReCo Dataset." In: *Language Documentation Reference Corpus (DoReCo) 1.2*. Ed. by Frank Seifart, Ludger Paschen, and Matthew Stave. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), 2022.

Søren Wichmann. *The Relationship among the Mixe-Zoque Languages of Mexico*. Salt Lake City, UT: University of Utah Press, 1995.

Jingting Ye. "Independent and Dependent Possessive Person Forms: Three Universals." In: *Studies in Language* 44.2 (2020), pp. 363–406.

Tim Zingler and Phillip Rogers. "The Length of Clitics and Affixes: A Phonological Approach to the Grammaticalization Cline." In: *Studies in Language* (To appear).