Noun-to-verb ratio and word order

Jan Strunk[1], Balthasar Bickel[2], Swintha Danielsen[3], Iren Hartmann[4], Brigitte Pakendorf[5], Søren Wichmann[4], Alena Witzlack-Makarevich[6], Taras Zakharko[2], Frank Seifart[1,4]

[1]U Amsterdam, [2]U Zürich, [3]U Leipzig, [4]MPI for Evolutionary Anthropology Leipzig, [5]CNRS & U Lyon-Lumière 2, [6]U Kiel

Also involved in data processing and annotation: Helen Geyer, Olga Khorkhova, Kristina Labs, Marlous Oomen, Robert Schikowski, Lena Sell, Lisa Steinbach, Laura Wägerle, and Evgeniya Zhivotova

Initial observation

Variation in noun vs. verb availability and/or usage across

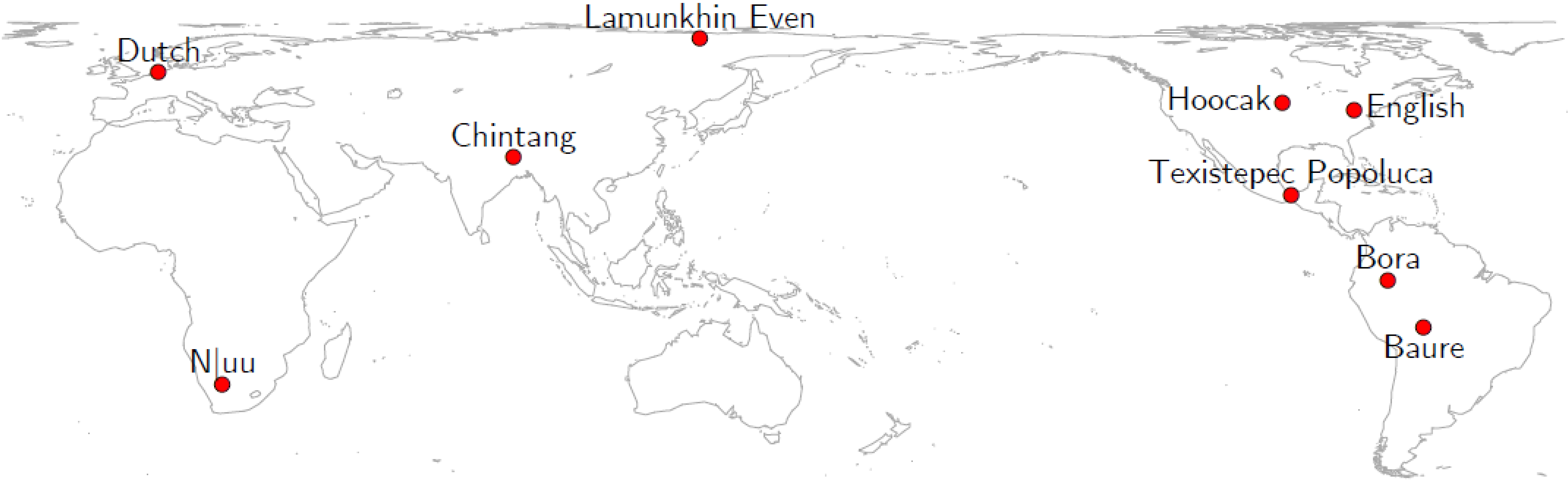
- the lifespan (Tardif et al. 1997[†], Bornstein et al. 2004[§], Stoll et al. 2010[†])
- brain health status (Bird et al. 2000[‡], Thompson et al. 2002[º])
- genres, registers, styles (Biber et al. 1998[+], Gaenszle et al. 2010[%])
- cultures and languages (Bickel 2003[*], Stoll & Bickel 2009[#])

or across combinations of these

Similar observations in our project *The relative frequencies of nouns, pronouns, and verbs cross-linguistically (NTVR)*

[*] *Language*, [#] *Crossling Appr Psych*, [+] *Corpus Ling*, [&] *Anth Ling*, [†] *J Child Lang*, [§] *Child Dev*, [‡] *Brain and Lang*, [º] *J Neuroling*

NTVR project: spoken corpora of 9 languages

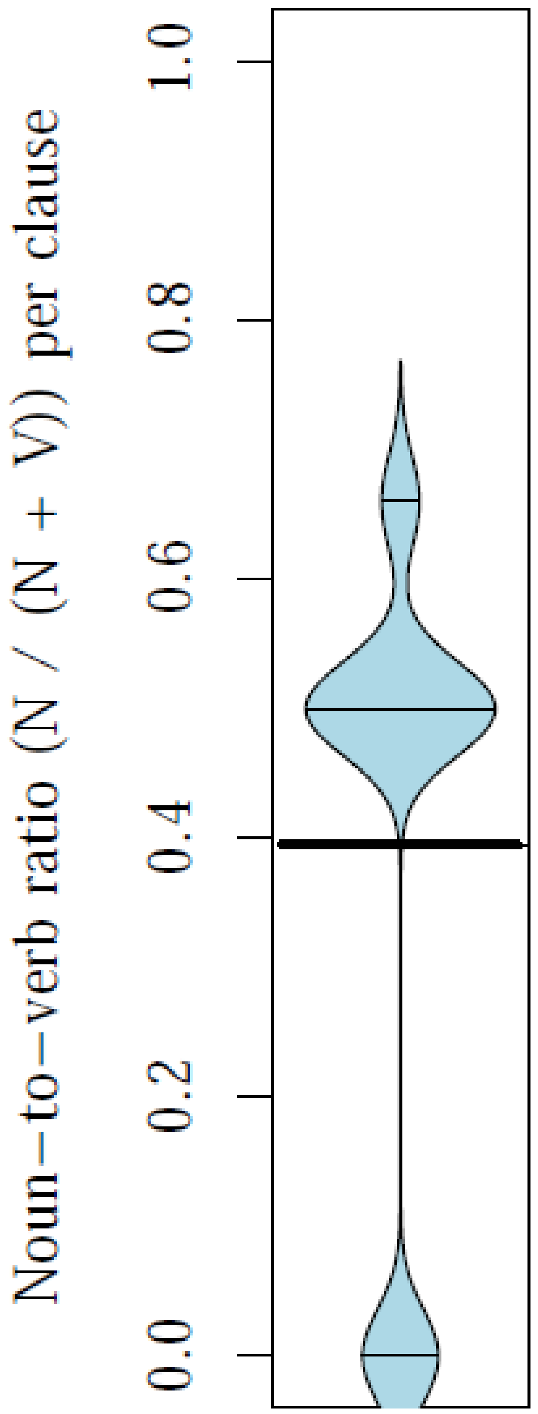


	Speakers	*Texts*	*Annotation Units*	*Words*
Baure (Arawakan; Danielsen et al. 2009)	15	45	4,925	**19,911**
Bora (Boran; Seifart 2009)	46	37	4,037	**29,997**
Chintang (Sino-Tibetan; Bickel et al. 2011)	74	40	9,378	**37,823**
Dutch (Indo-European; CGN; CGN-Consortium, Language and Speech Nijmegen & ELIS Gent 2003)	42	17	5,822	**39,720**
English (NXT-Switchboard Corpus; Godfrey & Holiman 1993; Calhoun et al. 2009)	80	47	6,942	**56,143**
Hoocak (Siouan; Hartmann 2013)	30	62	2,961	**23,207**
Lamunkhin Even (Tungusic; Pakendorf & Aralova 2010)	32	67	4,755	**34,294**
N|uu (!Ui-Taa; Güldemann et al. 2010)	8	33	8,257	**25,897**
Texistepec Popoluca (Mixe-Zoquean; Wichmann 1996)	1	9	6,453	**24,602**

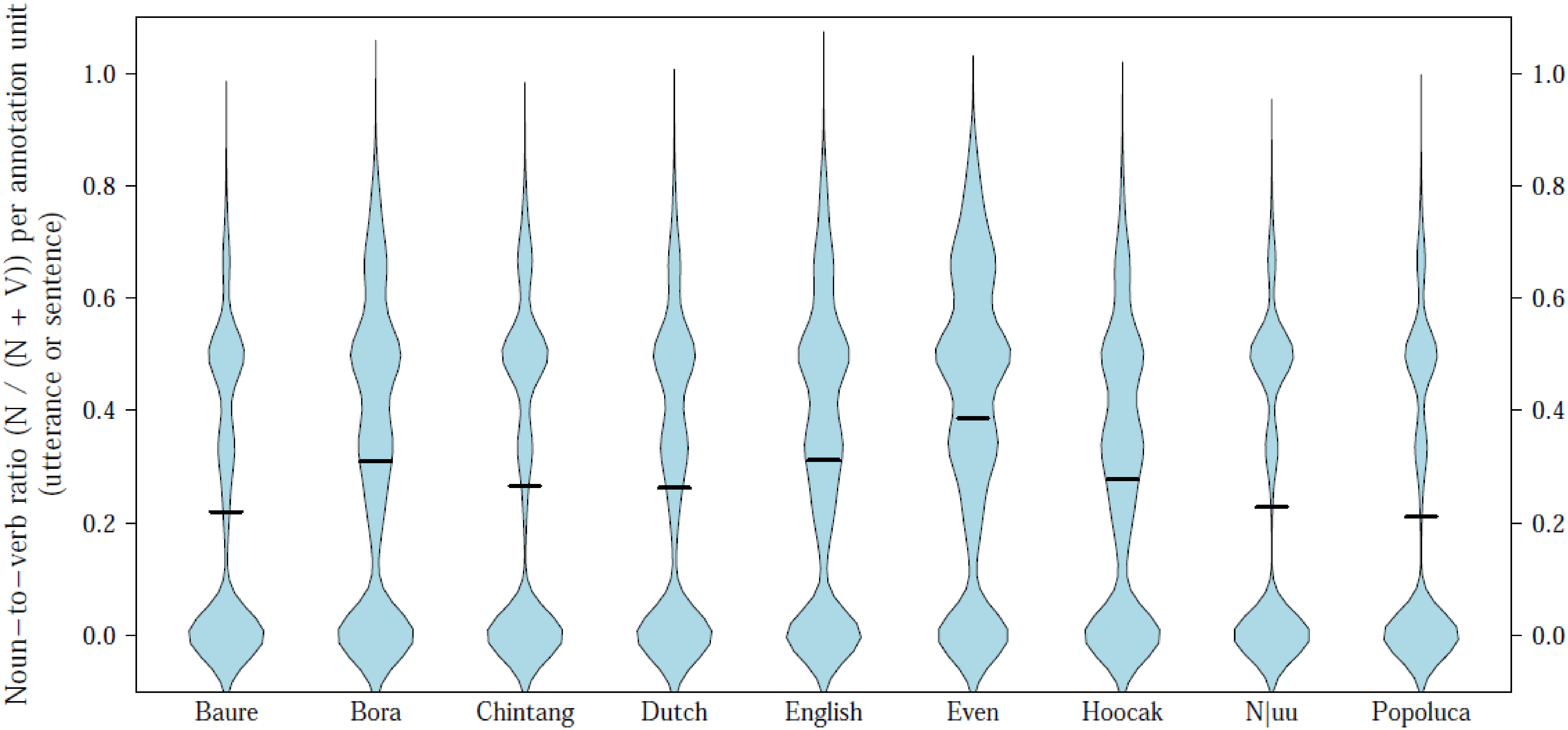
A simple example: NTVR = N/(N+V)

clause	N	V	NTVR
*A **man stayed** on a **farm**.*	2	1	0.67
*He **got** hungry.*	0	1	0
*He **says** to his **father-in-law**:*	1	1	0.5
*“**Give** me some **meat**!"*	1	1	0.5
*His **father-in-law says**:*	1	1	0.5
*“I **have** no **meat**,*	1	1	0.5
***go** to the **dune**,*	1	1	0.5
*and **hunt**!”*	0	1	0

(English translation of a N|uu story)



NTVR variation in our corpora



How to explain differences in noun vs. verb usage

- Earlier research: focused on nouns in **argument positions** and found explanations in types of agreement systems (Bickel 2003[*] on referential density)

- NTVR project: focus on noun and verb usage across the board
  - unlikely to be affected by type of agreement system (Bickel et al. 2013[#])
  - possible explanation: processing effects resulting from **word order**
  - for this study, we focus on the simple proportion of **nouns** rather than **nouns vs. verbs**
    (relative frequency: nouns / words)

[*] *Language*, [#] *Ass Ling Typ*

Theory: noun usage dependent on word order?

- Incremental production (for recent review, MacDonald 2013*)
  → alternation of partial utterance planning, execution, and subsequent planning
  → pressure to start and complete plans early

- Good for V-early structures, with early display of plan for proposition (predicate, argument structure, tense, mood, settings, etc.)

- Predictions from this for V-final structures ...

[*] *Front Psych*

Theory: noun usage dependent on word order?

Possible predictions for V-final structures:

- Increased usage of **non-verb tokens,** especially **nouns** as content words, in order to compensate for the delay in getting to the core information about the proposition
  - ***perhaps*** also more noun ***type*** variation (as observed in a correlational study of dictionaries by Polinsky 2012[+]), for more information load
  - but this may be counterbalanced by increased access cost that comes with lexical variation

[+] Polinsky, M. 2012. Headedness, again. In: *Theories of Everything. In Honor of Ed Keenan*. Los Angeles: UCLA.

Possible counter-hypothesis

- Noun usage is costly/harder to process in pre-verbal ***argument*** position (Ueno & Polinsky 2009[*]):
  - increased pro-drop
  - increased use of intransitives

- Other options:
  - production costs can also be avoided by right-dislocation (Pastor & Laka 2013[#])
  - production costs can be compensated for by optimizing lexicon shapes/the way semantic space is divided between verbs and nouns (Sauppe et al. 2013[%], in prep.)
  - speakers may just live with a slight speed loss (Seifart et al. 2014, in prep: higher N-to-V ratios result in lower production speeds)

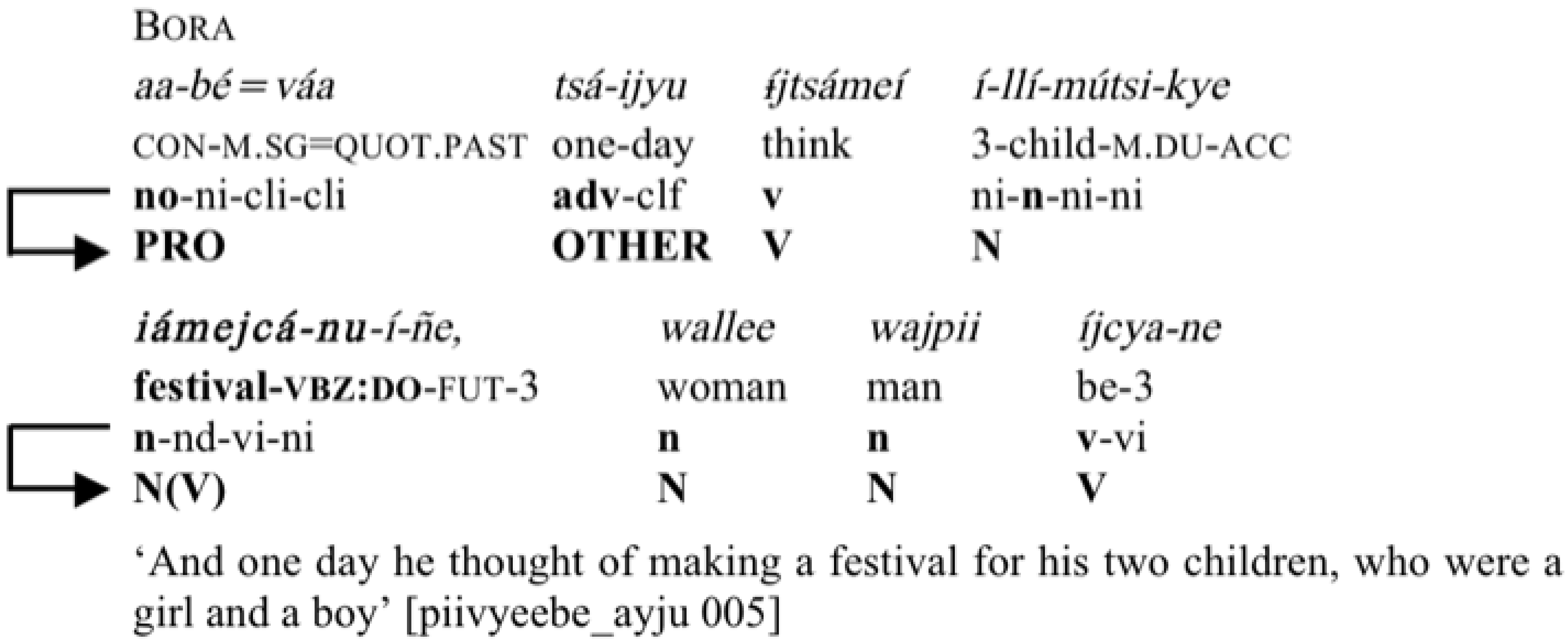
[*] *J of Ling*, [+]*Ann Rev Ling*, [#]*Open J Mod Ling*, [%]*Cog Sci*

Corpus Study

- Test the research hypothesis:
  - **Verb final languages** exhibit **increased noun usage** (in comparison to **verb non-final languages**),
  - expect weak signals for tokens
  - and perhaps also for lexical types

Data

- Mapping of language-specific PoS-tags to tags of {N, V, PRO, OTHER} per ***lexical root***



BORA

aa-bé=váa	*tsá-ijyu*	*íjtsámeí*	*í-llí-mútsi-kye*
CON-M.SG=QUOT.PAST	one-day	think	3-child-M.DU-ACC
no-ni-cli-cli	**adv**-clf	**v**	ni-**n**-ni-ni
PRO	**OTHER**	**V**	**N**

iámejcá-nu-í-ñe,	*wallee*	*wajpii*	*íjcya-ne*
festival-VBZ:DO-FUT-3	woman	man	be-3
n-nd-vi-ni	**n**	**n**	**v**-vi
N(V)	**N**	**N**	**V**

'And one day he thought of making a festival for his two children, who were a girl and a boy' [piivyeebe_ayju 005]

- Why roots?
- Our hypothesis concerns units with propositionally relevant content; in our corpus, PoS derivation like nominalization usually doesn't add information (e.g. nominalization for embedding)
- In more than 90% of cases, root and word category are identical

Methods: Linear mixed-effect models

- Linear mixed-effects models* predicting the **proportion of**
  1. **noun tokens** per **annotation unit** (utterance or sentence)
  2. **noun types** per **recording session / text**

- An extension of ordinary linear regression models that can account for random idiosyncrasies of natural groups in the data (e.g., texts of the same speaker, register, or language)

- P(nouns) ~ word order + plannedness + (1|session)

- Reads as: The proportion of nouns is predicted on the basis of the two predictors word order and plannedness (fixed effects) while accounting for random variation between recording sessions (random factor).

*lme4::lmer (Bates et al. 2014, CRAN)

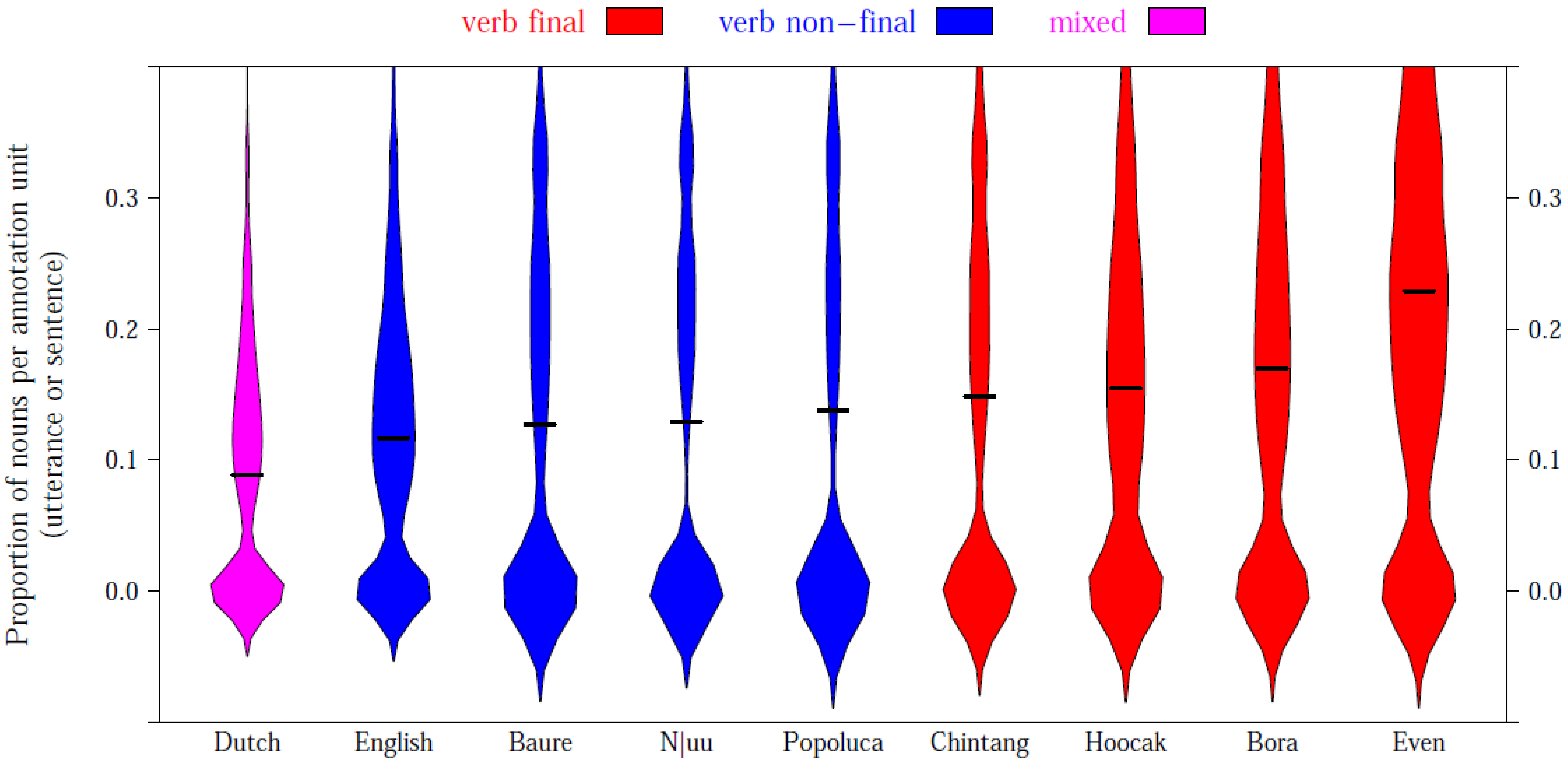
Methods

- **Fixed factors (predictors):**
  - **basic word order:**
    verb final vs. verb non-final (vs. mixed)
  - **speech setting:**
    monologue vs. dialogue vs. multi-party conversation, estimated on the basis of the number of speakers in a recording session
  - **plannedness:**
    - planned: (almost) memorized traditional narratives
    - semi-spontaneous: personal narratives, life stories, procedurals, etc.
    - spontaneous: open conversation

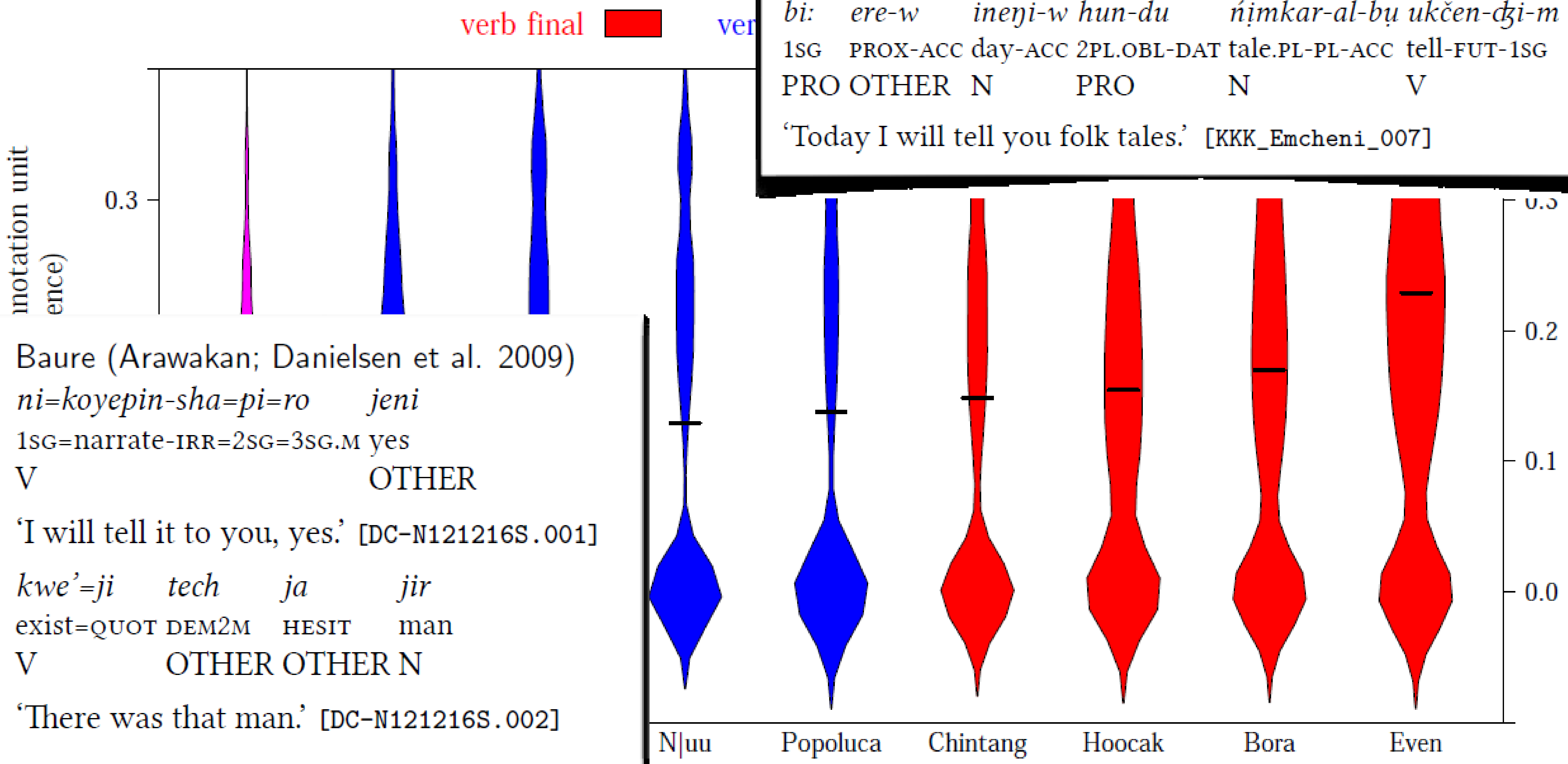
Methods

- **Random factors (for intercepts):**
  - **recording session**, capturing genre, topic choice, style, register, speakers and their social relations and interactions
  - **language**, capturing other aspects of grammar that might influence noun and verb usage

Results: proportion of nouns depending on word order



Results: proportion of nouns depending on word order



Even (Lamunkhin dialect; Tungusic; Pakendorf and Aralova 2010)

bi:	*ere-w*	*ineŋi-w*	*hun-du*	*ńịmkar-al-bụ*	*ukčen-ʤi-m*
1SG	PROX-ACC	day-ACC	2PL.OBL-DAT	tale.PL-PL-ACC	tell-FUT-1SG
PRO	OTHER	N	PRO	N	V

'Today I will tell you folk tales.' [KKK_Emcheni_007]

Baure (Arawakan; Danielsen et al. 2009)

ni=koyepin-sha=pi=ro	*jeni*
1SG=narrate-IRR=2SG=3SG.M	yes
V	OTHER

'I will tell it to you, yes.' [DC-N121216S.001]

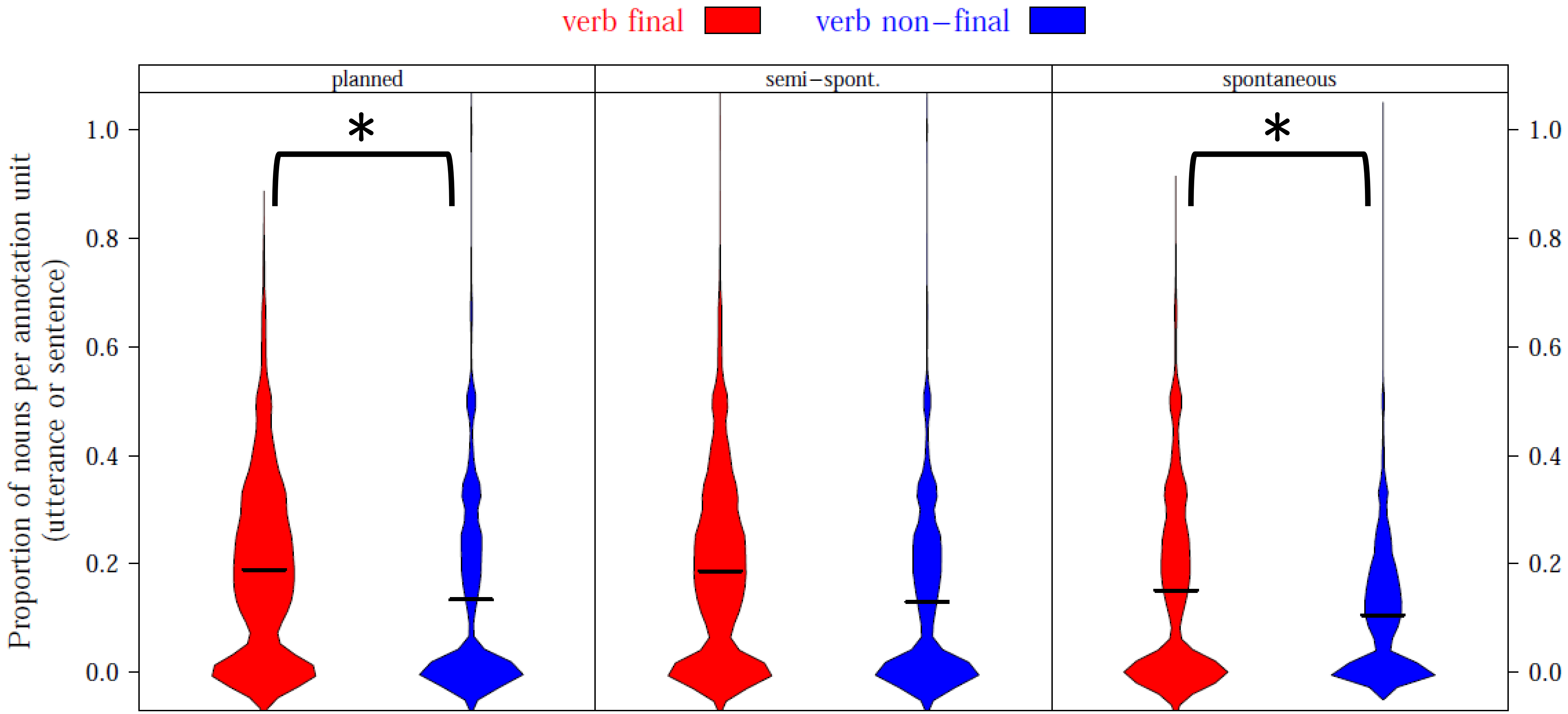
kwe'=ji	*tech*	*ja*	*jir*
exist=QUOT	DEM2M	HESIT	man
V	OTHER	OTHER	N

'There was that man.' [DC-N121216S.002]

Results: statistical model (proportion of nouns)

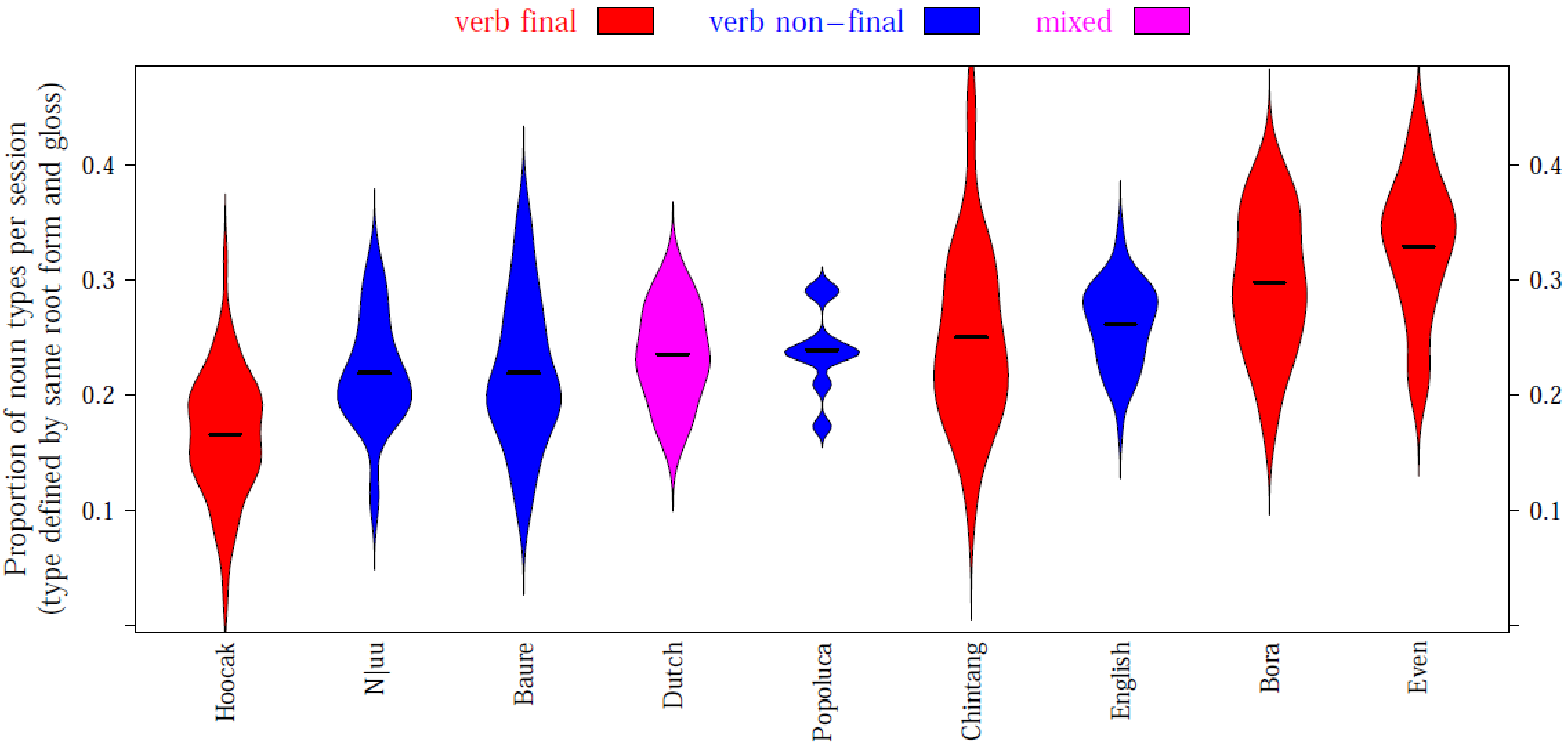
Best-fitting model: $P(\text{nouns}) \sim$ **word order × plannedness + speech setting** $+ (1|\text{session}) + (1|\text{language})$

interaction: p = .009, word order: p < .001, plannedness: p = .002, speech setting: p = .41, session: p < .001, language: p < .001



Results: lexical types (proportion of noun root types)

- Results for lexical types are much less clear
- Still a detectable overall word order effect



Discussion

- Heavier noun usage (tokens) in annotation units (sentences) of verb-final languages than in annotation units of verb-non-final languages

- Effect of word order detectable across categories of plannedness (planned, semi-spontaneous vs. spontaneous) and speech setting (monologue, dialogue vs. multi-party conversation)

- Word order effects mostly play out for the proportion of noun **tokens**, word order effects on the proportion of noun **types** (cf. Polinsky's 2012 dictionary-based approach) are still unclear

Conclusions

A small relativity effect:

The word order rules you follow also regulate the amount of noun roots you produce.

There is a higher average proportion of nouns in sentences of verb-final languages than in sentence of verb-non-final languages.

This is in line with relativity effects from other aspects of grammar (agreement systems) on noun vs. verb usage (Bickel 2003[*], Stoll & Bickel 2009[#]).

BUT the exact relationship between these effects still needs to be explored.

[*] *Language*, [#] *Crossling Appr Psych*

Thank you very much for your attention!