Catching variation during fieldwork on Nakh-Daghestanian languages

George Moroz, Samira Verhees

Linguistic Convergence Laboratory, NRU HSE

24 April 2020 (Dialectology and Linguistic Geography, Institute of Linguistics, RSUH)

Introduction

Investigating variation:

- In variationism (e.g. (Labov 1963) on Martha's Vineyard /ai/ ~ /au/, (Trudgill 1974) on Norwich speech, (Wolfram 1969) on Afro-American speech from Detroit) researchers focus on social stratification, mostly urban.

Investigating variation:

- In variationism (e.g. (Labov 1963) on Martha's Vineyard /ai/ ~ /au/, (Trudgill 1974) on Norwich speech, (Wolfram 1969) on Afro-American speech from Detroit) researchers focus on social stratification, mostly urban.
- "Two equally interesting questions are at the heart of this book: how an extraordinary degree of idiosyncratic linguistic variation can coexist with an extraordinarily homogeneous speaker population, and how linguists might overlook the possibility of their coexistence." (Dorian 2010: 3)

Investigating variation:

- In variationism (e.g. (Labov 1963) on Martha's Vineyard /ai/ ~ /au/, (Trudgill 1974) on Norwich speech, (Wolfram 1969) on Afro-American speech from Detroit) researchers focus on social stratification, mostly urban.
- "Two equally interesting questions are at the heart of this book: how an extraordinary degree of idiosyncratic linguistic variation can coexist with an extraordinarily homogeneous speaker population, and how linguists might overlook the possibility of their coexistence." (Dorian 2010: 3)
- In this talk we explore variation in a small, homogeneous speaker population and the probability that an average researcher of Nakh-Daghestanian languages catches this variation.

Data

Data

Data were collected from

- 44 speakers of Andi (Nakh-Daghestanian) during fieldwork in Zilo (Botlikh district, Dagestan) in 2019

Created with `lingtypology` (Moroz 2017)

Data

Data were collected from

- 44 speakers of Andi (Nakh-Daghestanian) during fieldwork in Zilo (Botlikh district, Dagestan) in 2019

Created with `lingtypology` (Moroz 2017)

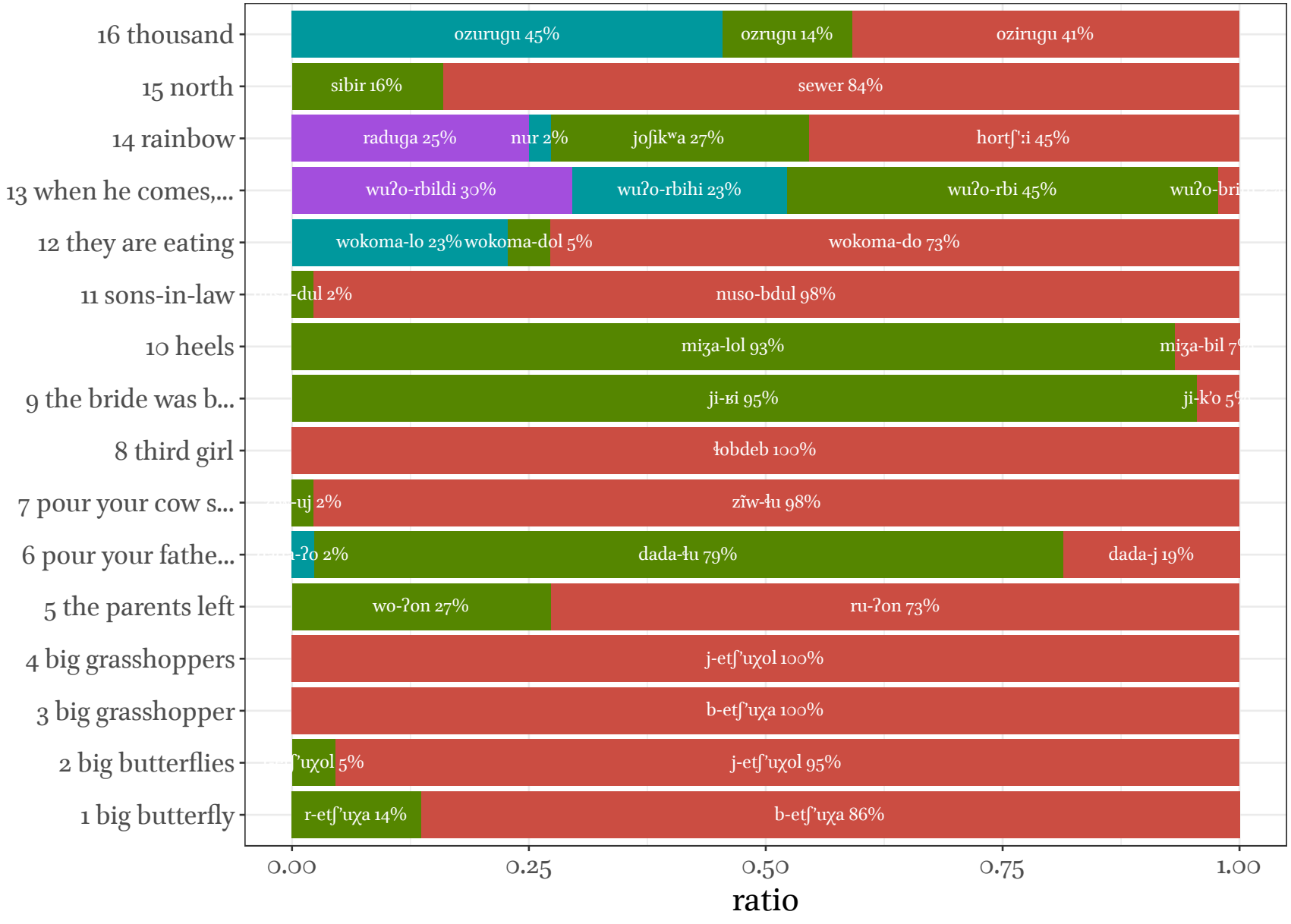
- and 23 researchers of Nakh-Daghestanian languages via an online questionary

Zilo Data

44 Zilo speakers were asked to translate 16 stimuli:

1. ‘big butterfly’
2. ‘big butterflies’
3. ‘big grasshopper’
4. ‘big grasshoppers’
5. ‘the parents left’
6. ‘pour your father some water’
7. ‘pour your cow some water’
8. ‘third girl’
9. ‘the bride was beautiful at the wedding’
10. ‘heels’
11. ‘sons-in-law’
12. ‘they are eating’
13. ‘when he comes, we will eat’
14. ‘rainbow’
15. ‘north’
16. ‘thousand’

Zilo questionary (44 speakers): results



Information entropy

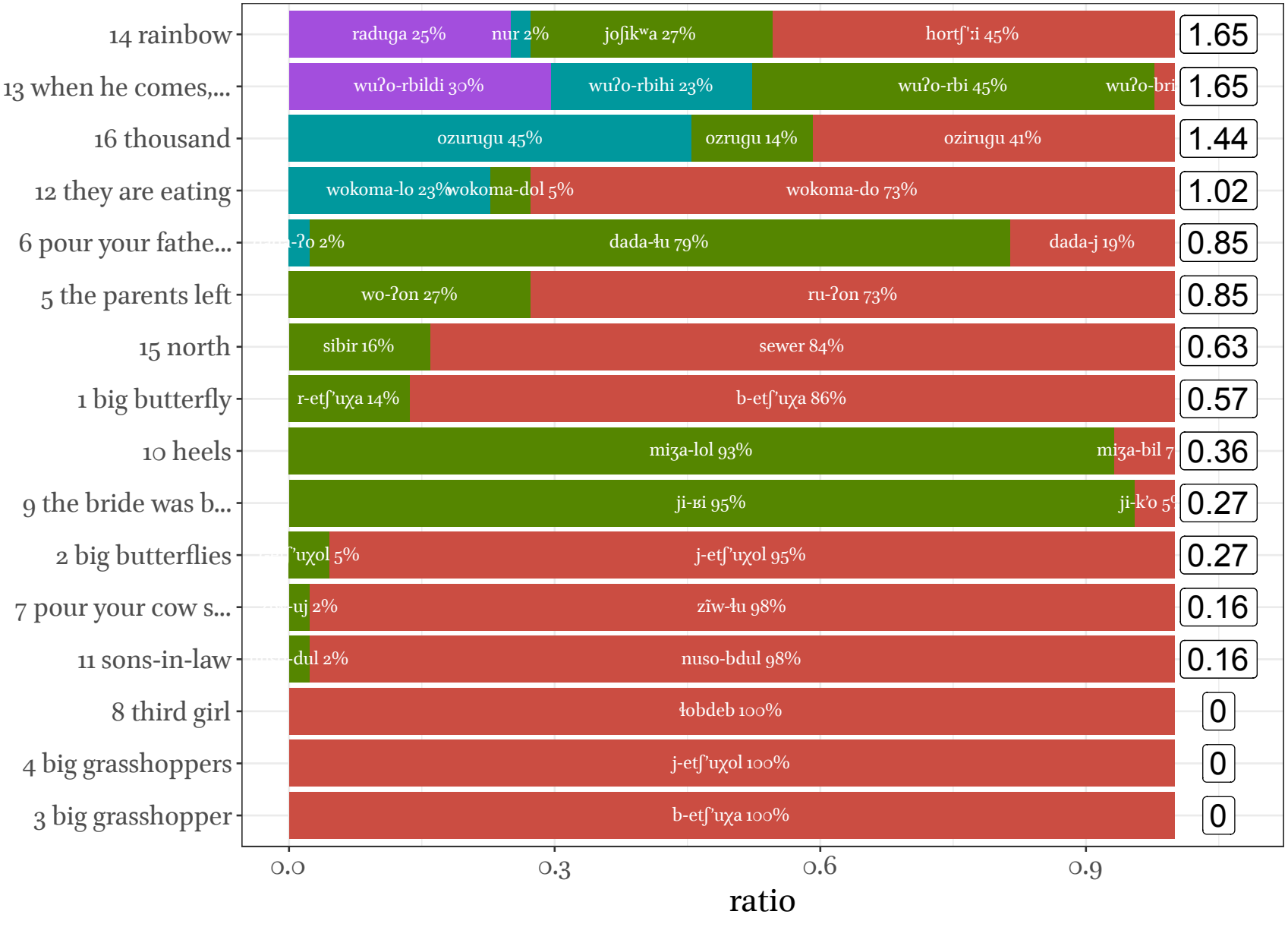
In order to measure the diversity of the questions we used the easiest measure — information entropy, introduced in (Shannon 1948):

$$H(X) = -\sum_{i=1}^{n} P(x_i) \times \log_2 P(x_i)$$

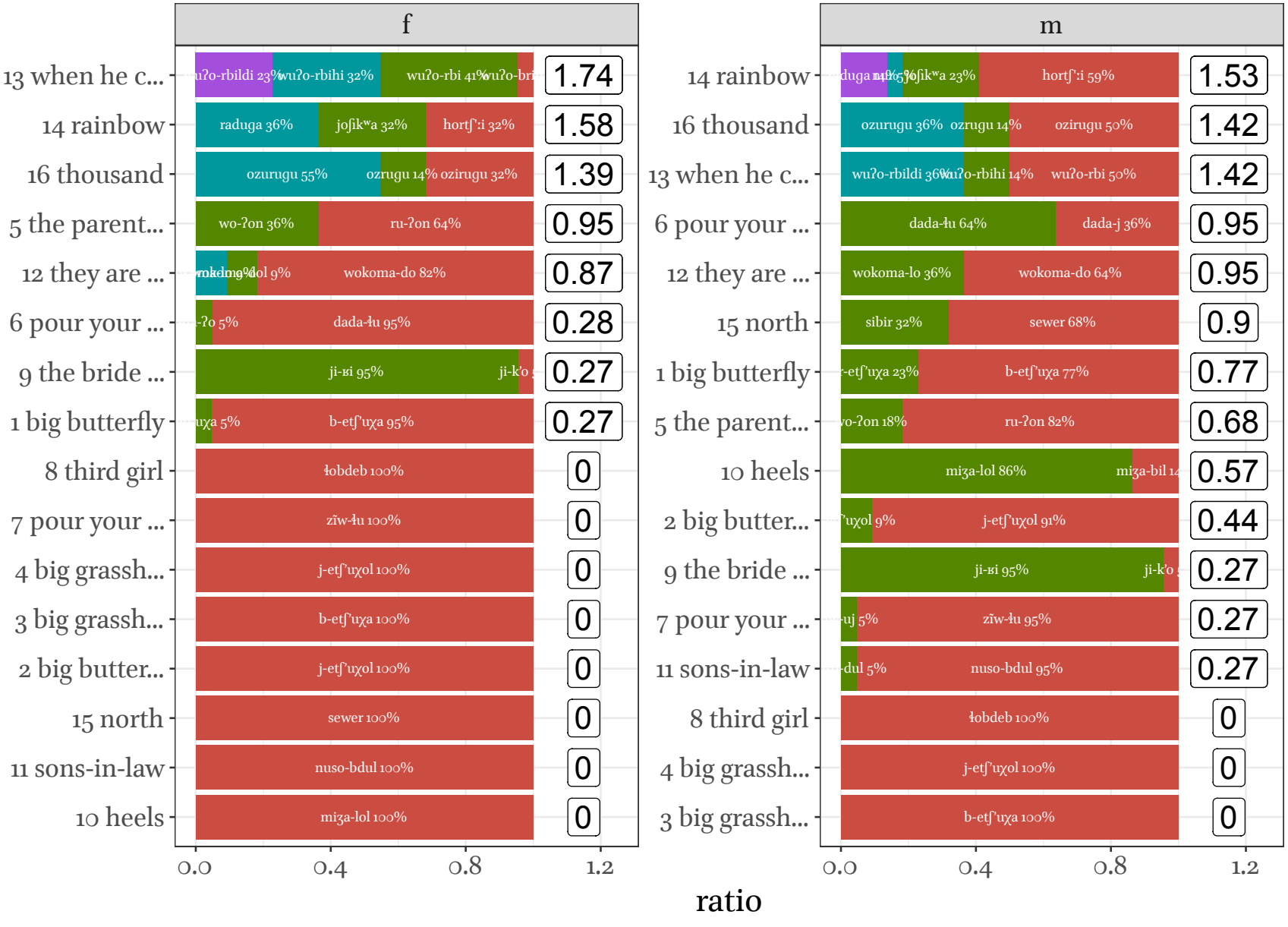
Range of the information entropy is $H(X) \in [0, +\infty]$:

data	entropy
A-A-A-A-A	0.00
A-A-A-A-B	0.72
A-A-A-B-B	0.97
A-A-B-B-B	0.97
A-A-B-B-C	1.52
A-B-C-A-B	1.52

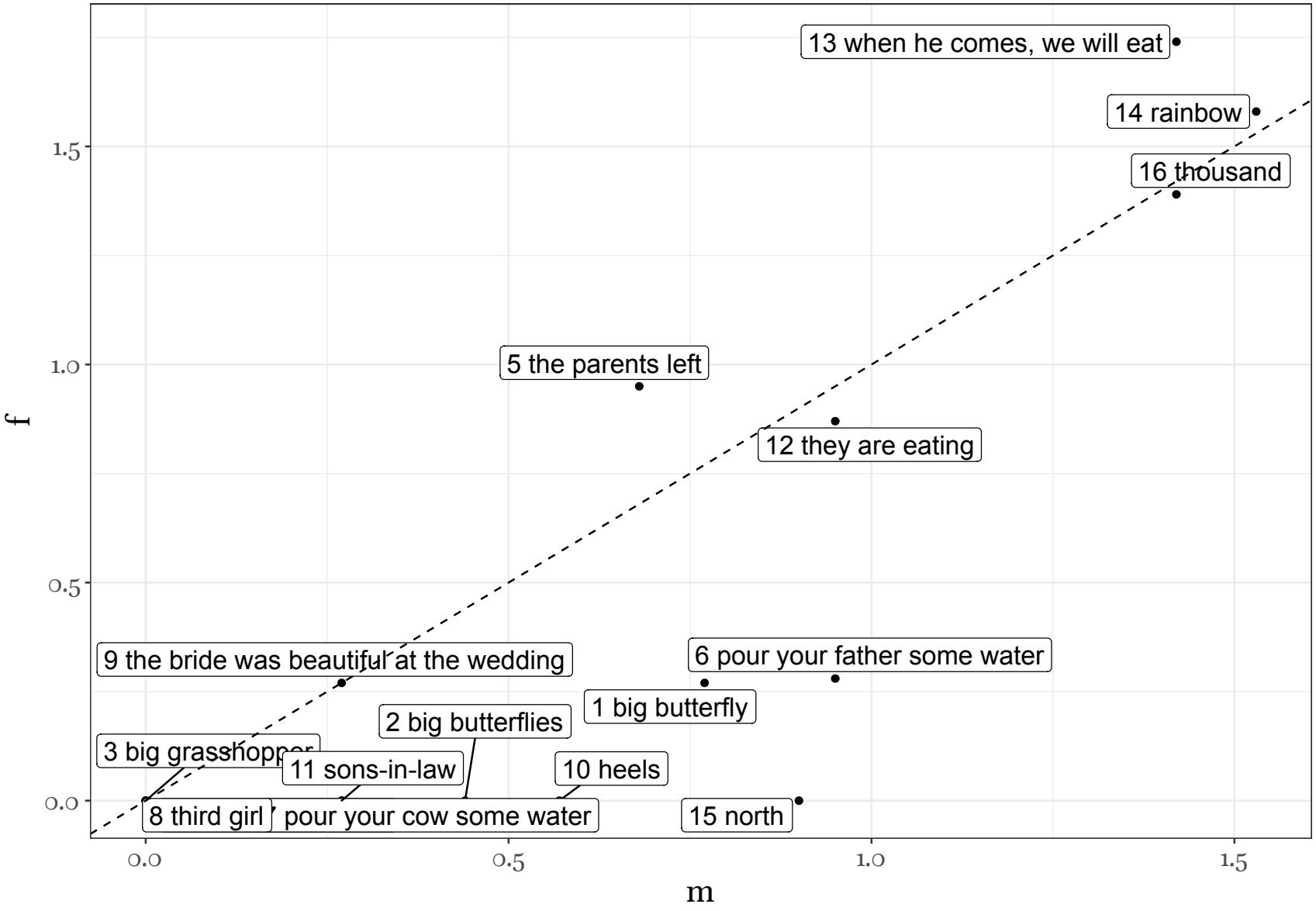
Zilo questionary (44 speakers): entropy value on the right



Zilo questionary (44 speakers): by gender



Zilo questionary (44 speakers): entropy by gender

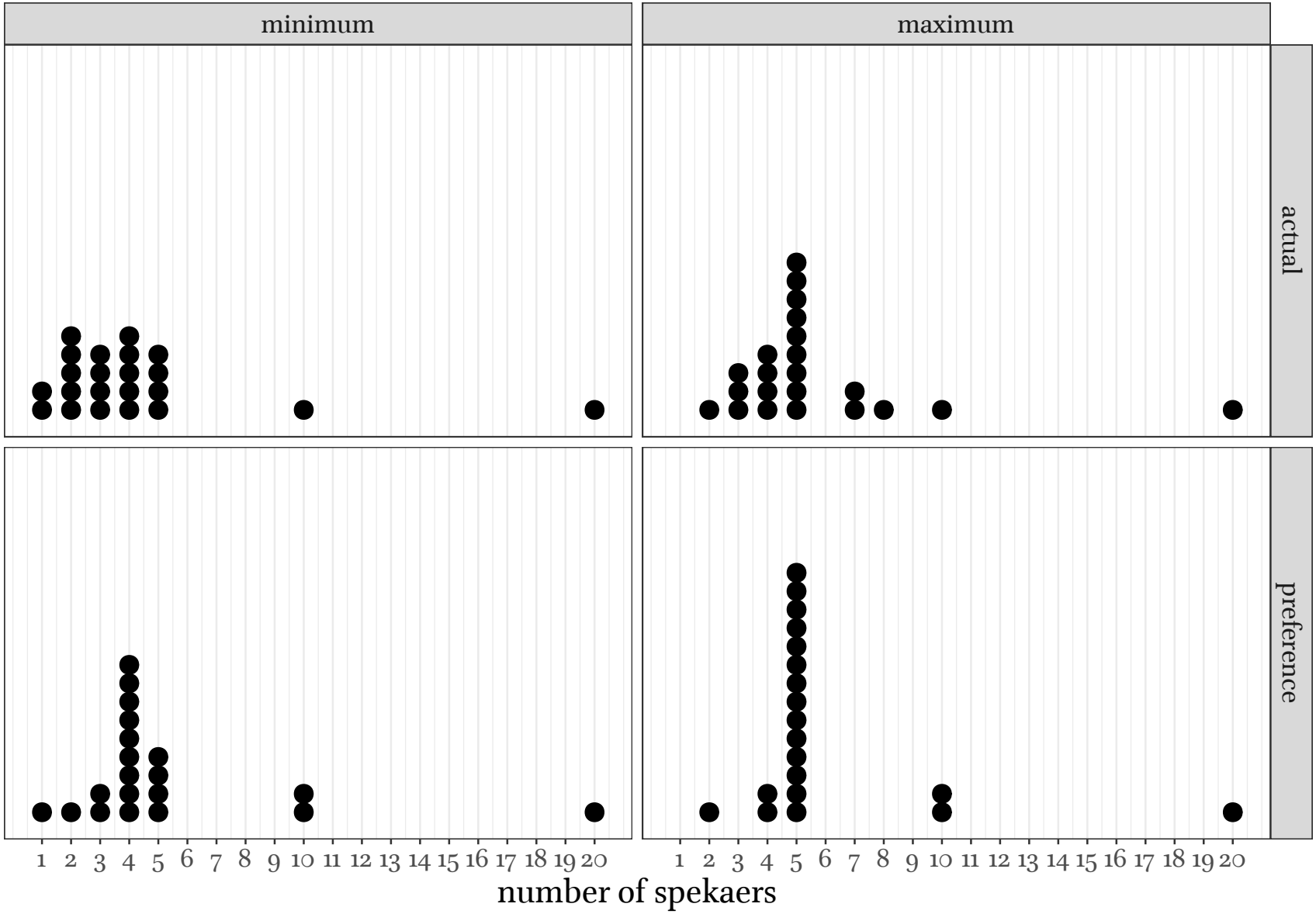


Nakh-Daghestanian Fieldwork Survey

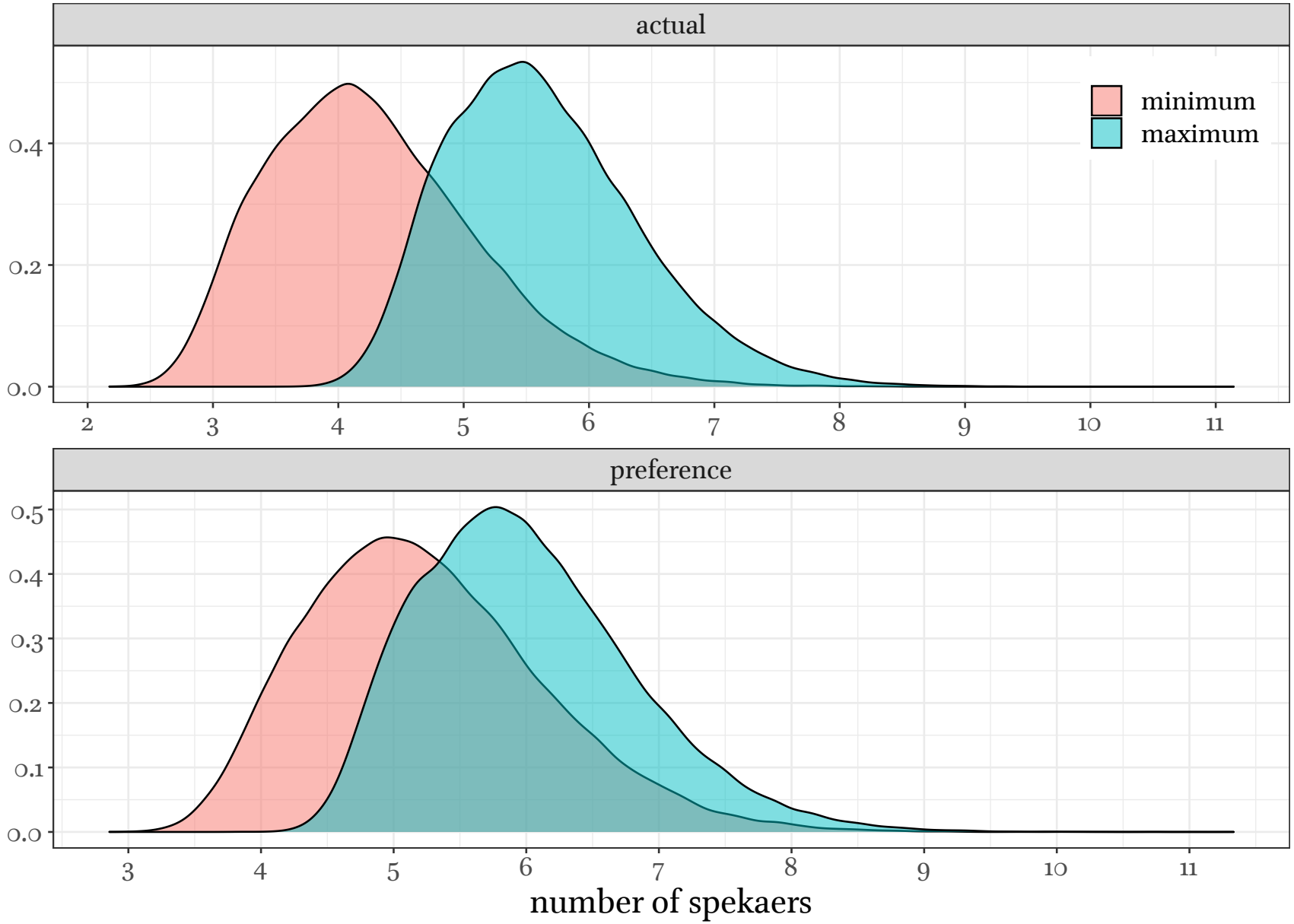
23 ND researchers were asked about:

- level of education
- linguistic interest
- studying linguistics at university
- fieldwork participation as a student
- year in which they finished their degree
- place of study and work
- number of people who participated in their fieldtrips
- preferred number of participants in fieldtrips
- goals of fieldwork
- use of elicitation and corpora
- **number of speakers a researcher *should* consult with**
- **number of speakers the researcher *usually* consults with**
- how researchers need to deal with interspeaker variability
- how researchers need to deal with intraspeaker variability
- whether speakers under the age of 13 are reliable consultants
- whether speakers older than 70 are reliable consultants
- personal (dis)preferences about the choice of consultants

Number of speakers

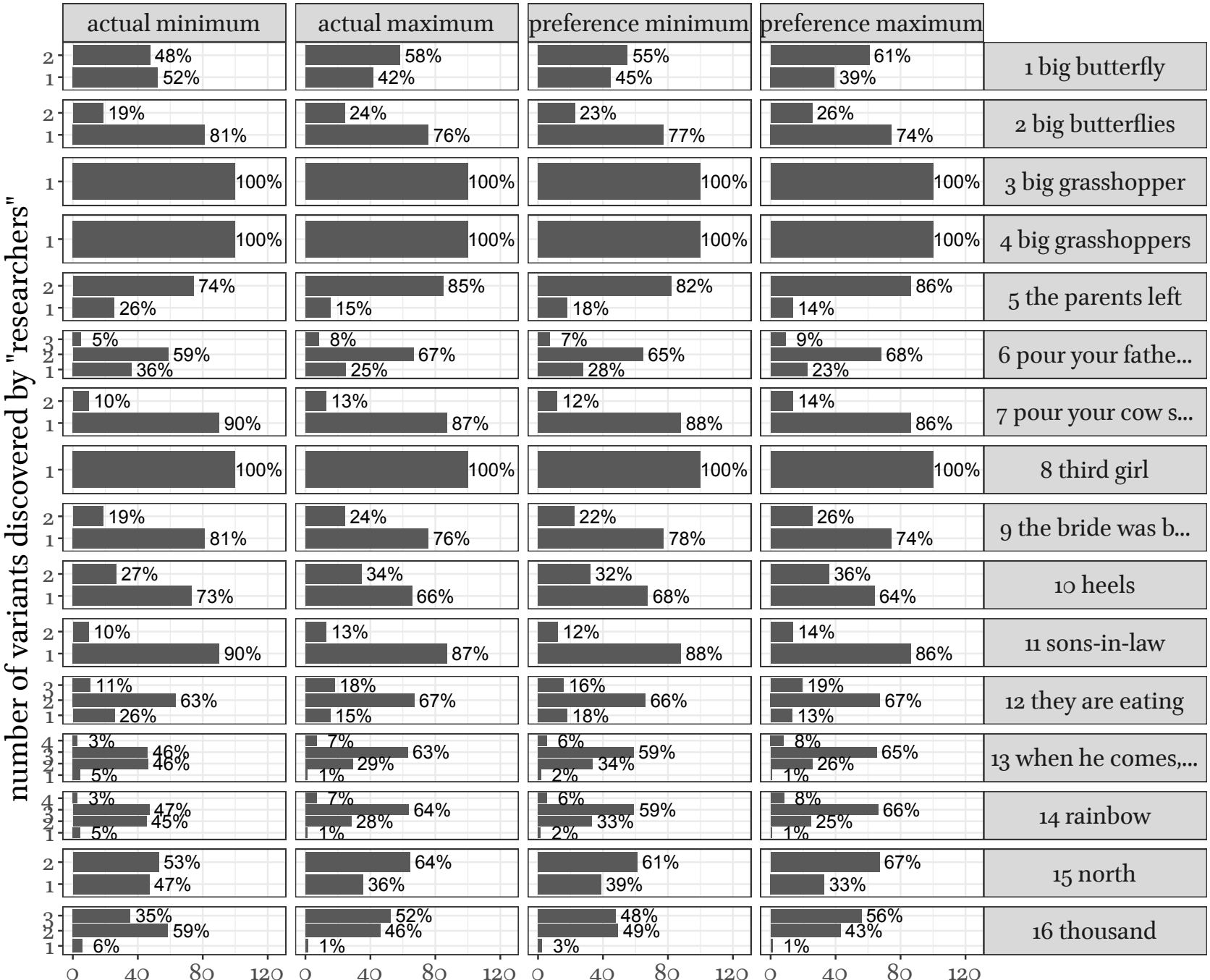


Bootstrapped mean number of speakers (10^5 iterations)

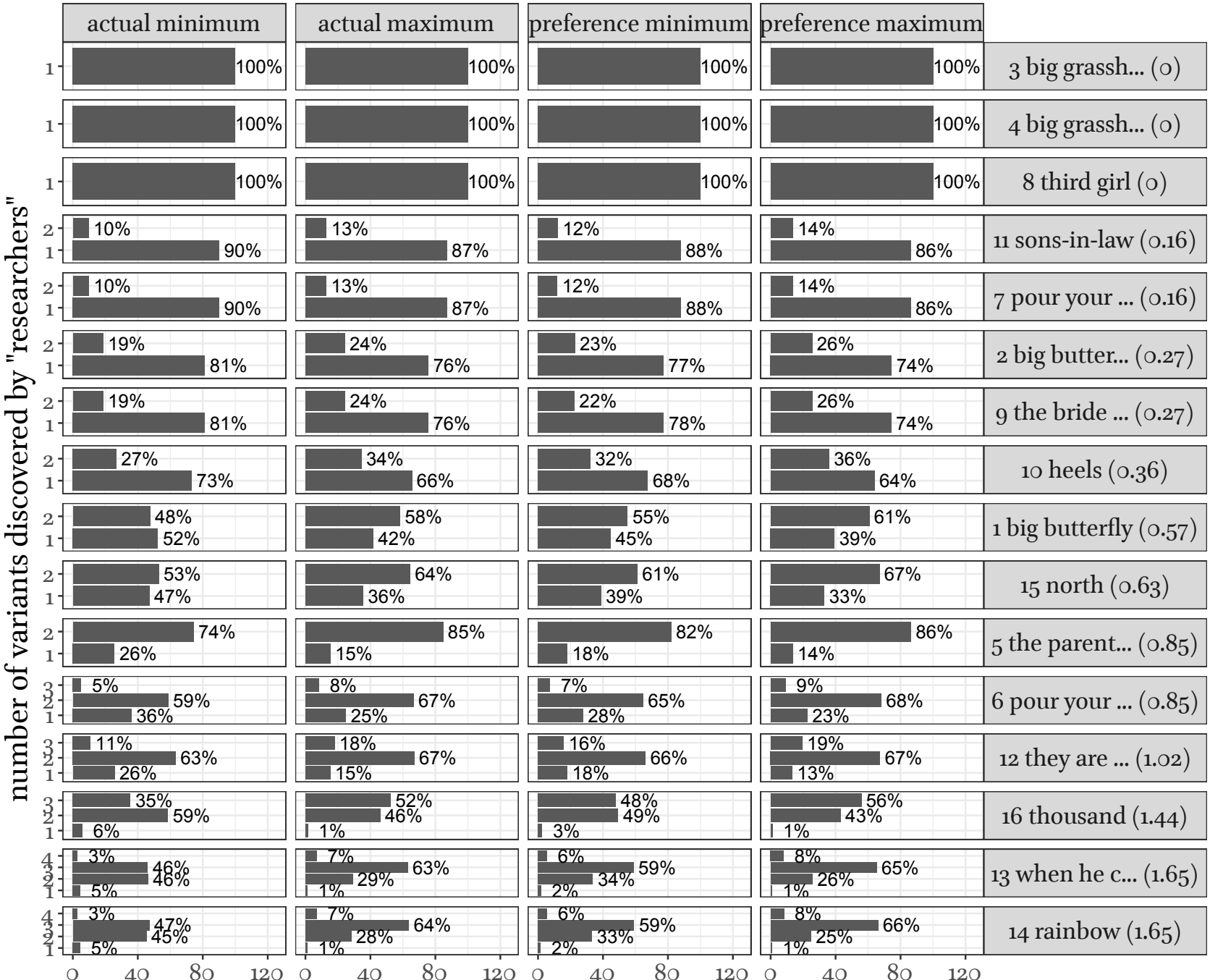


What if 10^5 “average researchers” …
come to Zilo?

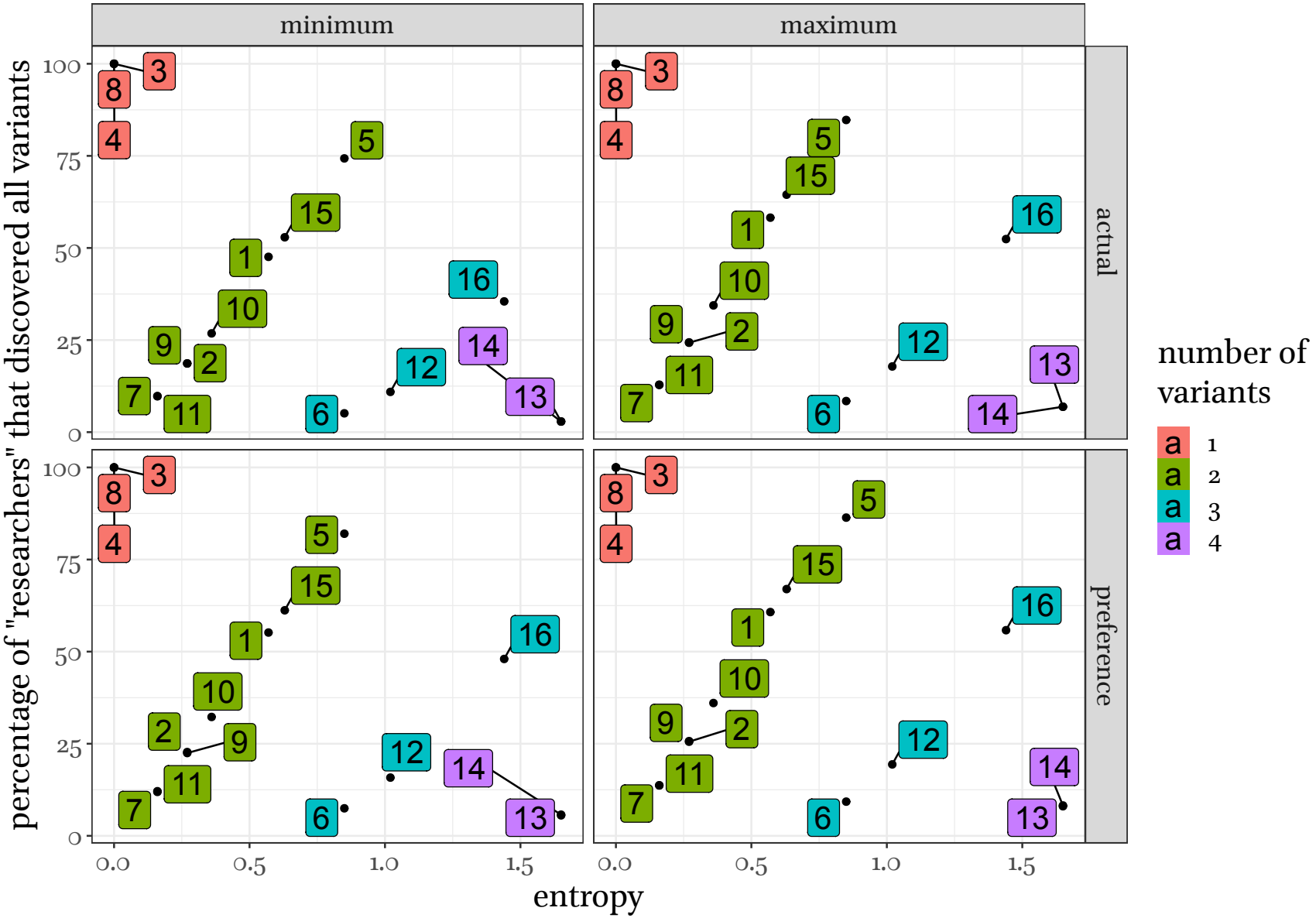
10^5 samples from experiment results



10^5 samples from experiment results sorted by entropy



When “researchers” will find less?



Number on the plot represents number id of the question.

Conclusions

Conclusions:

- Shannon information entropy helps to find some variation spots
- An "**average researcher**" might overlook a significant amount of the variation we observed due to the low number of speakers they usually consult with (the more possible values and the lower variable entropy, the higher the likelihood of overlooking variables)
- However, our experiment with 44 speakers also failed to show some of the variation we found in prior research on this dialect
- The observed variation should be explored in more detail using the collected sociolinguistic parameters (it looks like variation does not corelate with gender)
- The characteristics of the "**average researcher**" of Nakh-Daghestanian languages can be further eloborated using the parameters collected in the survey
- The observed variation remains a collection of isolated lexical, phonological and morphological facts…

Conclusions:

- Shannon information entropy helps to find some variation spots
- An "**average researcher**" might overlook a significant amount of the variation we observed due to the low number of speakers they usually consult with (the more possible values and the lower variable entropy, the higher the likelihood of overlooking variables)
- However, our experiment with 44 speakers also failed to show some of the variation we found in prior research on this dialect
- The observed variation should be explored in more detail using the collected sociolinguistic parameters (it looks like variation does not corelate with gender)
- The characteristics of the "**average researcher**" of Nakh-Daghestanian languages can be further eloborated using the parameters collected in the survey
- The observed variation remains a collection of isolated lexical, phonological and morphological facts…
    - Is it possible to study variation in syntax in this manner?
    - Could variational variables be interrelated?

Conclusions:

- Shannon information entropy helps to find some variation spots
- An "**average researcher**" might overlook a significant amount of the variation we observed due to the low number of speakers they usually consult with (the more possible values and the lower variable entropy, the higher the likelihood of overlooking variables)
- However, our experiment with 44 speakers also failed to show some of the variation we found in prior research on this dialect
- The observed variation should be explored in more detail using the collected sociolinguistic parameters (it looks like variation does not corelate with gender)
- The characteristics of the "**average researcher**" of Nakh-Daghestanian languages can be further eloborated using the parameters collected in the survey
- The observed variation remains a collection of isolated lexical, phonological and morphological facts…
    - Is it possible to study variation in syntax in this manner?
    - Could variational variables be interrelated?
- And what do all these results contribute to linguistic theory?

References

Dorian, N. C. (2010). *Investigating variation: The effects of social organization and social setting*. Oxford University Press.

Labov, W. (1963). The social motivation of a sound change. *Word*, 19(3):273–309.

Moroz, G. (2017). *lingtypology: easy mapping for Linguistic Typology*.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Trudgill, P. (1974). *The social differentiation of English in Norwich*. Cambridge University Press.

Wolfram, W. A. (1969). *A Sociolinguistic Description of Detroit Negro Speech., No. 5.*, volume 5 of *Urban language*. Center for Applied Linguistics.