Areal, family, and sociolinguistic effects on conditional constructions

Laura Becker*, Matías Guzmán Naranjo**, Samira Ochs*

*University of Freiburg, **University of Tübingen

Family, areal, and sociolinguistic effects in typology

Family and areal effects in typology

Controlling for family effects

- sampling:
  Bell (e.g. 1978), Bickel (2008), Bickel (2011), Dahl (2008), Nichols (1992), and Rijkhoff et al. (1993)

- statistical control (group-level effects):
  Bentz and Winter (e.g. 2013), Blasi et al. (2019), and Levshina (2019)

Family and areal effects in typology

Controlling for family effects

- sampling:
  Bell (e.g. 1978), Bickel (2008), Bickel (2011), Dahl (2008), Nichols (1992), and Rijkhoff et al. (1993)
- statistical control (group-level effects):
  Bentz and Winter (e.g. 2013), Blasi et al. (2019), and Levshina (2019)

Controlling for areal effects

- macro areas:
  (e.g. Dryer 1989; Dryer 1992; Hammarström and Donohue 2014; Donohue and Whiting 2011; Hammarström and Güldemann 2014)
- relative measure of space between languages:
  Dryer (2018), Jaeger et al. (2011), and Cysouw, Dediu, and Moran (2012)

Sociolinguistic effects in typology

- there is ample evidence for socio-linguistic factors having an impact on the crosslinguistic distribution of grammatical features
(Trudgill 2011b; Trudgill 2011a; Trudgill 2008; Perkins 1992; De Busser and LaPolla 2015; Karlsson, Miestamo, and Sinnemäki 2008; Wray and Grace 2007; Lupyan and Dale 2016; Ladd, Roberts, and Dediu 2015; Biber 2009; Chafe 1982; Chafe and Tannen 1987; Mithun 1984; Nettle 2012; Sinnemäki and Di Garbo 2018; Sinnemäki 2020; Bentz and Winter 2013; Biber 2009)

Sociolinguistic effects in typology

- there is ample evidence for socio-linguistic factors having an impact on the crosslinguistic distribution of grammatical features
  (Trudgill 2011b; Trudgill 2011a; Trudgill 2008; Perkins 1992; De Busser and LaPolla 2015; Karlsson, Miestamo, and Sinnemäki 2008; Wray and Grace 2007; Lupyan and Dale 2016; Ladd, Roberts, and Dediu 2015; Biber 2009; Chafe 1982; Chafe and Tannen 1987; Mithun 1984; Nettle 2012; Sinnemäki and Di Garbo 2018; Sinnemäki 2020; Bentz and Winter 2013; Biber 2009)

- socio-linguistic factors
    - population size, proportion of L2 speakers
    - language contact (multilingual settings)
    - domains of use
    - modality of use (written vs. spoken), literary tradition
    - register (formal vs. informal) ...

Sociolinguistic effects in typology

- there is ample evidence for socio-linguistic factors having an impact on the crosslinguistic distribution of grammatical features
  (Trudgill 2011b; Trudgill 2011a; Trudgill 2008; Perkins 1992; De Busser and LaPolla 2015; Karlsson, Miestamo, and Sinnemäki 2008; Wray and Grace 2007; Lupyan and Dale 2016; Ladd, Roberts, and Dediu 2015; Biber 2009; Chafe 1982; Chafe and Tannen 1987; Mithun 1984; Nettle 2012; Sinnemäki and Di Garbo 2018; Sinnemäki 2020; Bentz and Winter 2013; Biber 2009)

- socio-linguistic factors
  - population size, proportion of L2 speakers
  - language contact (multilingual settings)
  - domains of use
  - modality of use (written vs. spoken), literary tradition
  - register (formal vs. informal) ...

- associated linguistic features
  - pronominal, demonstrative systems
  - spatial orientation
  - inflectional morphology
  - word order flexibility
  - subordination (syntactic complexity)
  - nominal case ...

Family, areal, and sociolinguistic effects

? How do all family, areal, and sociolinguistic factors interact?

? Are associations with current population sizes due to areal biases in samples (cf. Nichols 2009)?

Family, areal, and sociolinguistic effects

? How do all family, areal, and sociolinguistic factors interact?

? Are associations with current population sizes due to areal biases in samples (cf. Nichols 2009)?

- such questions have received even less attention in linguistic typology

- notable exceptions
  - Sinnemäki and Di Garbo (2018) and Sinnemäki (2020)
  - GramAdapt project
  - WS 14 at SLE 2021!

Family, areal, and sociolinguistic effects

? How do all family, areal, and sociolinguistic factors interact?

? Are associations with current population sizes due to areal biases in samples (cf. Nichols 2009)?

- such questions have received even less attention in linguistic typology

- notable exceptions
    - Sinnemäki and Di Garbo (2018) and Sinnemäki (2020)
    - GramAdapt project
    - WS 14 at SLE 2021!

? To what extent is the **number of dedicated conditional constructions** associated with sociolinguistic predictors once areal and family effects are taken into account?

Family, areal, and sociolinguistic effects

? How do all family, areal, and sociolinguistic factors interact?

? Are associations with current population sizes due to areal biases in samples (cf. Nichols 2009)?

- such questions have received even less attention in linguistic typology

- notable exceptions
  - Sinnemäki and Di Garbo (2018) and Sinnemäki (2020)
  - GramAdapt project
  - WS 14 at SLE 2021!

? To what extent is the **number of dedicated conditional constructions** associated with sociolinguistic predictors once areal and family effects are taken into account?

☞ Martowicz (2011) finds an association between the grammaticalization, lexicalization and explicitness of conditional markers and various sociolinguistic factors

[...] the evidence gathered for anteriority and conditionality suggest that encoding of these two relations is very prone to the influence of socio-cultural factors. (Martowicz 2011: 310)

A note on conditional constructions

Types of conditionals

We can (roughly) distinguish between 3 types of conditionals (Kortmann 1997: 85, Thompson, Longacre, and Hwang 2007: 255-256)

❶ **real conditionals**
present, past, future, or general situations for which it is unresolved whether or not they are, were, or will be true

(1)   a.   If it's raining on my way home, I am getting wet.
        b.   If you do not get enough sleep, you will be tired all day.

❷ **hypothetical conditional**
imagined situations which might happen

(2)   If I went for a run, I'd feel less stressed afterwards.

❸ **counterfactual conditional**

(3)   If you had been at the meeting, you would have seen the new CEO.

Types of conditionals

We can (roughly) distinguish between 3 types of conditionals (Kortmann 1997: 85, Thompson, Longacre, and Hwang 2007: 255-256)

❶ **real conditionals**
present, past, future, or general situations for which it is unresolved whether or not they are, were, or will be true
(1) a. If it's raining on my way home, I am getting wet.
  b. If you do not get enough sleep, you will be tired all day.

❷ **hypothetical conditional**
imagined situations which might happen
(2) If I went for a run, I'd feel less stressed afterwards.

❸ **counterfactual conditional**
(3) If you had been at the meeting, you would have seen the new CEO.

Examples of (non-)dedicated conditional constructions

Conditional marker (often in the protasis)

(4) [**Là** góe=p'ét]   [t'òng góe=múút].
 COND 2SG.M.S=exit.SG IRR 2SG.M.S=die.SG
 'If you go out, you will die.'   Goemai (Hellwig 2011: 457)

Verbal marker

(5) [wà-**á**-gám-yà]   [e-èké-gúnówó]
 S:2SG-**COND**-greet-O:3SG S:3SG-NEG.FUT-answer
 'If you greet X, X will not respond.'   Oko (Atoyebi 2010: 94)

No overt marker (juxtapposition)

(6) [Dimilyi-rni   maja-nga-yi], [wukurni ngilma-nga-yi].
 lancewood.bark-FOC get-1SG-FUT humpy   make-1SG-FUT
 If I get some lancewood bark, I can make a humpy.   Jingulu (Pensalfini 1997: 211)

Relative construction

(7) [Nyila=ma=rna=nga warlagu=ma ba-rru guliyan=ma] [**nyamu**=yi=nga
 that=TOP=1MIN.S=DUB dog=TOP   hit-POT dangerous=TOP **REL**=1MIN.O=DUB
 baya-wu].
 bite=POT
 'I'll hit the agressive dog, if it bites me.' Bilinarra (Meakins and Nordlinger 2013: 307)

Examples of (non-)dedicated conditional constructions

Conditional marker (often in the protasis)

(4) [**Là** góe=p'ét]   [t'òng góe=múút].
     COND 2SG.M.S=exit.SG IRR   2SG.M.S=die.SG
     'If you go out, you will die.'   Goemai (Hellwig 2011: 457)

Verbal marker

(5) [wà-**á**-gám-yà]   [e-èké-gúnówó]
     S:2SG-**COND**-greet-O:3SG S:3SG-NEG.FUT-answer
     'If you greet X, X will not respond.'   Oko (Atoyebi 2010: 94)

No overt marker (juxtapposition)

(6) [Dimilyi-rni   maja-nga-yi], [wukurni ngilma-nga-yi].
     lancewood.bark-FOC get-1SG-FUT   humpy   make-1SG-FUT
     If I get some lancewood bark, I can make a humpy.   Jingulu (Pensalfini 1997: 211)

Relative construction

(7) [Nyila=ma=rna=nga warlagu=ma ba-rru guliyan=ma] [**nyamu**=yi=nga
     that=TOP=1MIN.S=DUB dog=TOP   hit-POT dangerous=TOP REL=1MIN.O=DUB
     baya-wu].
     bite=POT
     'I'll hit the agressive dog, if it bites me.' Bilinarra (Meakins and Nordlinger 2013: 307)

Data & methods

Sample

our sample includes 300 languages
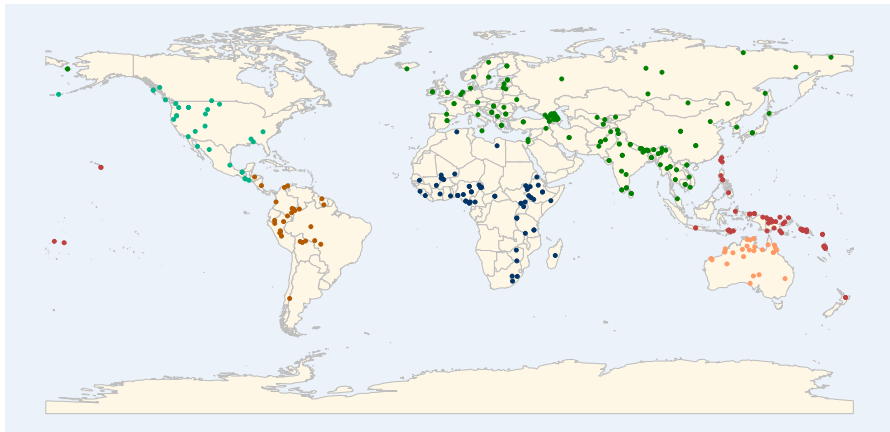
Sociolinguistic annotation

- the relevant information is not available for many of the languages
- ☞ established measures could be applied as such

Speaker information

- N speakers
- proportion of L2 speakers (45 languages)
- multilingual speakers: no < some < many < most < all (213 languages)

Language information

- literature tradition: yes, no
- used in writing: no < little < yes
- used in education: no < language classes < little < yes

Hypotheses

- **N speakers**
 - more speakers $\rightarrow$ more variation and more conditional constructions

Hypotheses

- **N speakers**
- more speakers → more variation and more conditional constructions
- **modality & register**
- formal, written use: need to be more explicit → more dedicated constructions
- languages with a literary tradition tend to have more adverbial subordinators (Kortmann 1997: 256)

- **N speakers**
- more speakers $\rightarrow$ more variation and more conditional constructions

- **modality & register**
- formal, written use: need to be more explicit $\rightarrow$ more dedicated constructions
- languages with a literary tradition tend to have more adverbial subordinators (Kortmann 1997: 256)

- **degree of multilingualism**
- the borrowing of conditional markers is not uncommon (cf. Bakker and Hekking 2012; Matras 2008; Martowicz 2011)
- multilingual settings favor contact-induced borrowing

Hypotheses

- **N speakers**
- more speakers → more variation and more conditional constructions
- **modality & register**
- formal, written use: need to be more explicit → more dedicated constructions
- languages with a literary tradition tend to have more adverbial subordinators (Kortmann 1997: 256)
- **degree of multilingualism**
- the borrowing of conditional markers is not uncommon (cf. Bakker and Hekking 2012; Matras 2008; Martowicz 2011)
- multilingual settings favor contact-induced borrowing
- **areal effects**
- low degree of explicitness of conditional markers in Australia and Papunesia (Martowicz 2011: 276)

Results

General distributions

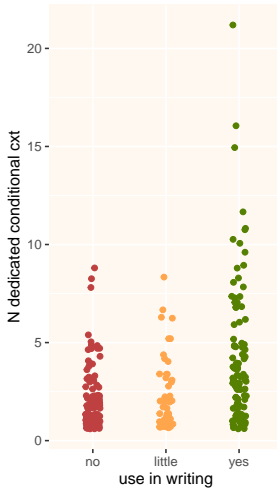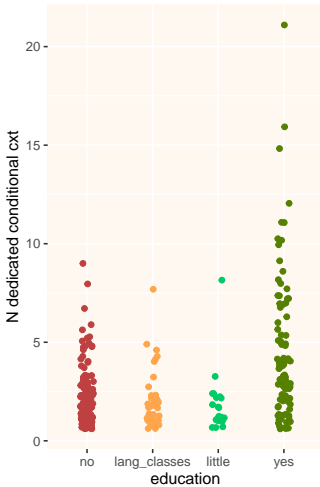N dedicated cxt ● 0 ● 1–2 ● 3–5 ● > 5
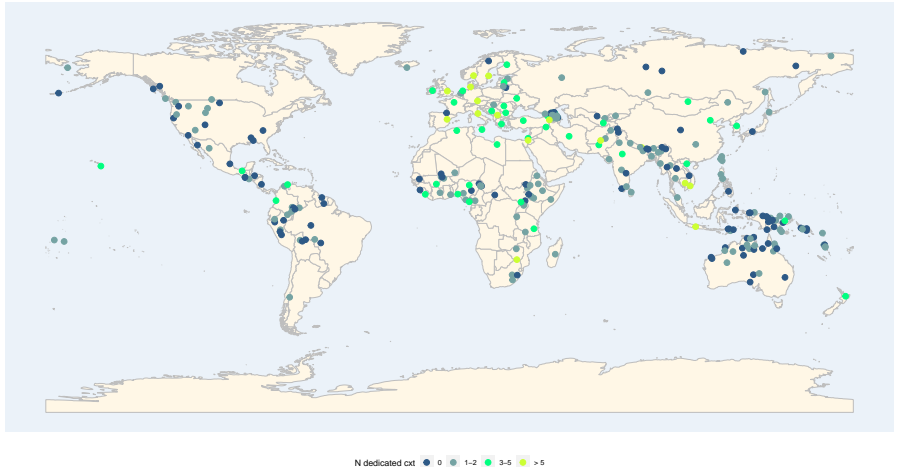
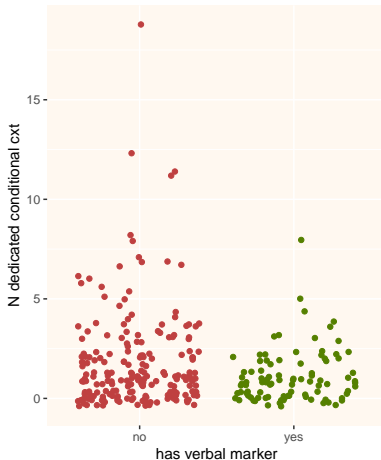# Number of speakers

- more speakers → more constructions
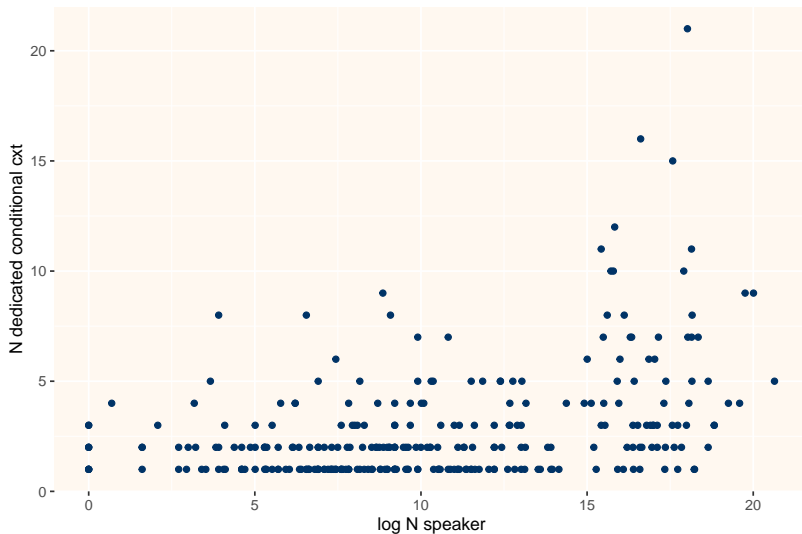
Other sociolinguistic factors

- higher degree of written/formal use → more explicit conditional expressions
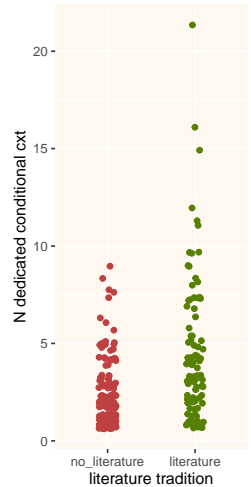
Availability of a verbal conditional marker

- languages with a conditional marker that is part of the verbal (inflection) paradigm tend to have fewer dedicated conditional markers overall

Modelling

Choosing the predictors

- we fitted a number of Bayesian poisson regression models to estimate the number of conditional markers per language

Choosing the predictors

- we fitted a number of Bayesian poisson regression models to estimate the number of conditional markers per language

models	elpd diff	se diff
`N_cxt ~ log_speaker + edu + verbal`	0.0	0.0
`N_cxt ~ log_speaker + edu`	-5.5	4.6
`N_cxt ~ log_speaker + edu + lit + writing + verbal`	-7.0	4.9
`N_cxt ~ log_speaker`	-13.6	7.4
`N_cxt ~ edu`	-17.2	8.8
`N_cxt ~ lit`	-30.0	11.2
`N_cxt ~ writing`	-35.3	11.8

Choosing the predictors

- we fitted a number of Bayesian poisson regression models to estimate the number of conditional markers per language

| models	elpd diff	se diff
`N_cxt ~ log_speaker + edu + verbal`	0.0	0.0
`N_cxt ~ log_speaker + edu`	-5.5	4.6
`N_cxt ~ log_speaker + edu + lit + writing + verbal`	-7.0	4.9
`N_cxt ~ log_speaker`	-13.6	7.4
`N_cxt ~ edu`	-17.2	8.8
`N_cxt ~ lit`	-30.0	11.2
`N_cxt ~ writing`	-35.3	11.8

Choosing the predictors

- we fitted a number of Bayesian poisson regression models to estimate the number of conditional markers per language

models	elpd diff	se diff
`N_cxt ~ log_speaker + edu + verbal`	0.0	0.0
`N_cxt ~ log_speaker + edu`	-5.5	4.6
`N_cxt ~ log_speaker + edu + lit + writing + verbal`	-7.0	4.9
`N_cxt ~ log_speaker`	-13.6	7.4
`N_cxt ~ edu`	-17.2	8.8
`N_cxt ~ lit`	-30.0	11.2
`N_cxt ~ writing`	-35.3	11.8

Choosing the predictors

- we fitted a number of Bayesian poisson regression models to estimate the number of conditional markers per language

models	elpd diff	se diff
`N_cxt ~ log_speaker + edu + verbal`	0.0	0.0
`N_cxt ~ log_speaker + edu`	-5.5	4.6
`N_cxt ~ log_speaker + edu + lit + writing + verbal`	-7.0	4.9
`N_cxt ~ log_speaker`	-13.6	7.4
`N_cxt ~ edu`	-17.2	8.8
`N_cxt ~ lit`	-30.0	11.2
`N_cxt ~ writing`	-35.3	11.8

- we used approximate LOO-CV to calculate the expected log predictive density (ELPD) (Vehtari, Gelman, and Gabry 2017)

No-controls model: conditional effects

`N_cxt ~log_speaker + education + verbal_marker`
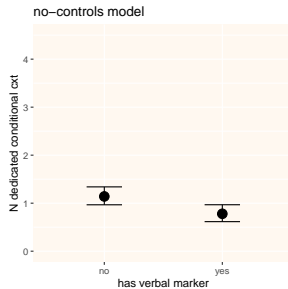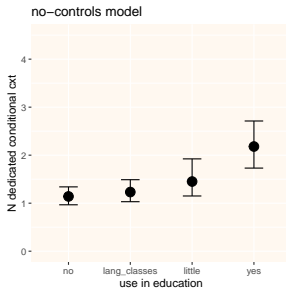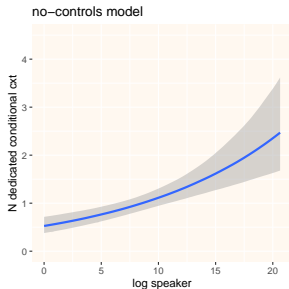
- with no additional linguistic controls, both the number of speakers and the use of the language in education have a clear effect on the number of conditional constructions
- the presence of a verbal conditional marker also has a robust weak effect

Modelling family and areal effects

Including family and areal controls

- we then fitted a series of poisson models to examine the robustness of the predictors from
  the **no-controls model**
  `N_cxt ~ 1 + log_speaker + edu + verbal`

Including family and areal controls

- we then fitted a series of poisson models to examine the robustness of the predictors from
  the **no-controls model**
  ```
  N_cxt ~ 1 + log_speaker + edu + verbal
  ```

- we first added "traditional" random intercepts for family and macro area:
  the **hierarchical model**
  ```
  N_cxt ~ 1 + log_speaker + edu + verbal + (1|family +
  (1|macroarea)
  ```

Including family and areal controls

- we then fitted a series of poisson models to examine the robustness of the predictors from
  the **no-controls model**
  ```
  N_cxt ~ 1 + log_speaker + edu + verbal
  ```

- we first added "traditional" random intercepts for family and macro area:
  the **hierarchical model**
  ```
  N_cxt ~ 1 + log_speaker + edu + verbal + (1|family +
  (1|macroarea)
  ```

- we controlled for family and areal/contact effects using a phylogenetic term and a Gaussian process, respectively (Guzmán Naranjo and Becker accepted):
  the **full model**
  ```
  N_cxt ~ 1 + log_speaker + edu + verbal + PT + GP
  ```

Including family and areal controls

- we then fitted a series of poisson models to examine the robustness of the predictors from
  the **no-controls model**
  ```
  N_cxt ~ 1 + log_speaker + edu + verbal
  ```

- we first added "traditional" random intercepts for family and macro area:
  the **hierarchical model**
  ```
  N_cxt ~ 1 + log_speaker + edu + verbal + (1|family +
  (1|macroarea)
  ```

- we controlled for family and areal/contact effects using a phylogenetic term and a Gaussian process, respectively (Guzmán Naranjo and Becker accepted):
  the **full model**
  ```
  N_cxt ~ 1 + log_speaker + edu + verbal + PT + GP
  ```

- as a sanity check, we used those controls without the fixed effects:
  ```
  N_cxt ~ 1 + PT + GP
  ```

Family bias: Phylogenetic regression

- the phylogenetic term (PT) can account for the fact that language families are trees

Family bias: Phylogenetic regression

- the phylogenetic term (PT) can account for the fact that language families are trees
- ☞ we do not need to choose any cut-off point between phylogenetic levels
  (family, genus, etc.)

Family bias: Phylogenetic regression

- the phylogenetic term (PT) can account for the fact that language families are trees
- ☞ we do not need to choose any cut-off point between phylogenetic levels (family, genus, etc.)

- a PT includes information about all relations between the languages in the sample:
- e.g. **Spanish** is more closely related to **Catalan** than to **Italian**, but these three are closer to each other than to **German**

Family bias: Phylogenetic regression

- the phylogenetic term (PT) can account for the fact that language families are trees
- ☞ we do not need to choose any cut-off point between phylogenetic levels (family, genus, etc.)

- a PT includes information about all relations between the languages in the sample:
- e.g. **Spanish** is more closely related to **Catalan** than to **Italian**, but these three are closer to each other than to **German**
- ☞ the model estimates effects for languages which must respect the phylogenetic distances

  (the more closely-related two languages are, the less their effects can vary)

Geographic bias: Gaussian process

- we use latitude and longitude information to represent the location of languages (Hammarström, Forkel, et al. 2020)
- with this, we include a **Gaussian process (GP)** in our model (a surface of 2 dimensions)
- the model estimates whether there are regions in the map that are strongly associated with the response variable
- ☞ we do not need to assume that the effect of distances is constant
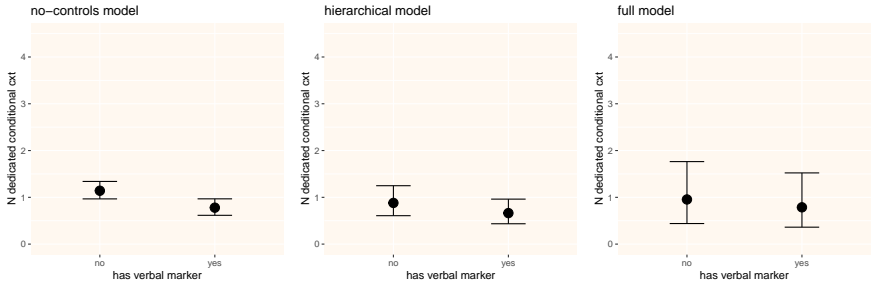
Model comparisons

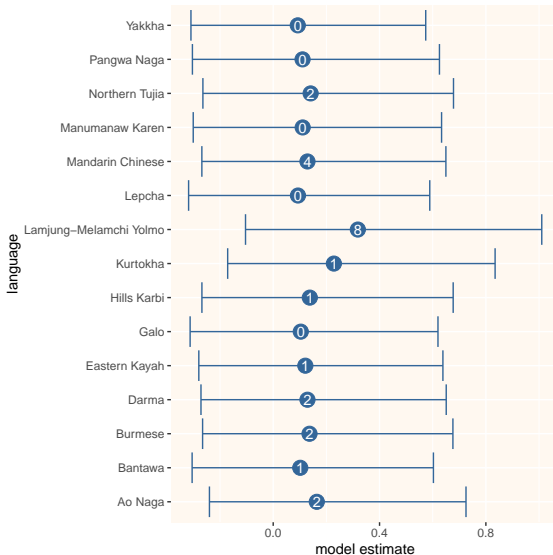# Comparing the conditional effects for N speakers
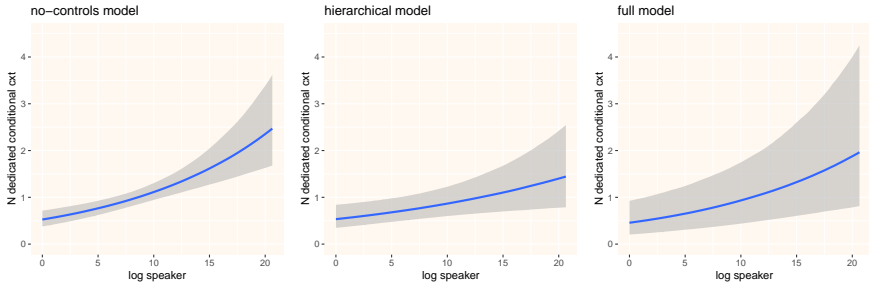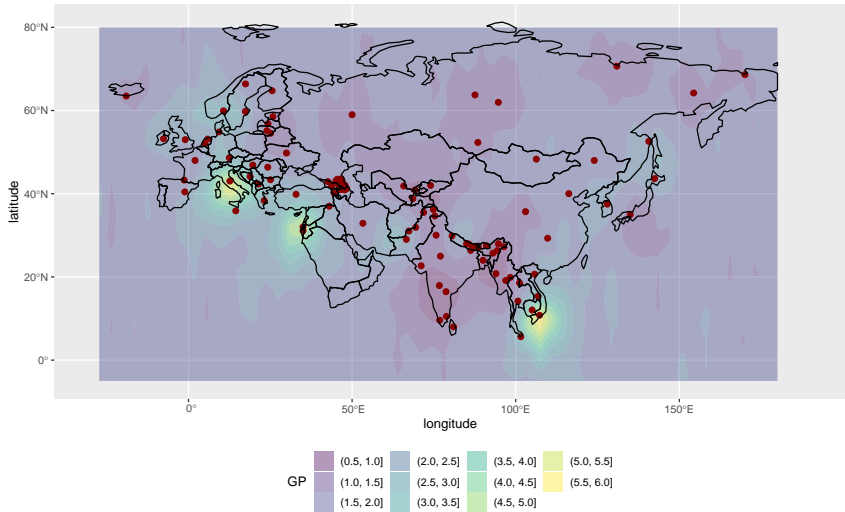
Comparing the conditional effects for use in education
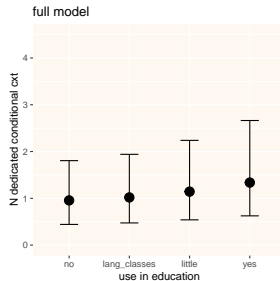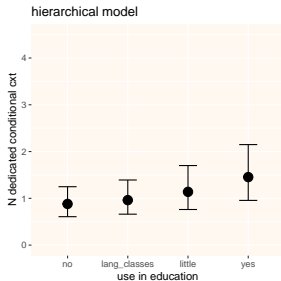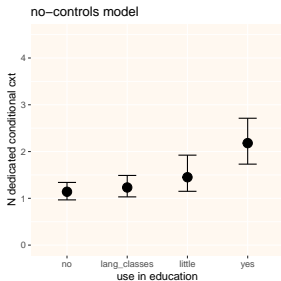
Comparing the conditional effects for verbal marker

no–controls model

hierarchical model

full model

Phylogenetic effects

Areal effects: Eurasia

Comparing the predictive power

- we performed approximate LOO-CV to compare the predictive power of the different models

predictors	elpd diff	se diff
`log_speaker + edu + verbal + `PT` + `GP	0.0	0.0
PT` + `GP	-10.3	9.1
`log_speaker + edu + verbal + (1|`family`) + (1|`area`)`	-24.4	11.2
`log_speaker + edu + verbal`	-34.7	15.1

☞ precise family and areal controls are crucial

Conclusion

Little evidence for sociolinguistic effects

- when family and areal effects are properly controlled for, the effects of the sociolinguistic predictors (used here) are much less certain

We need more sociolinguistic data

Controlling for a potential bibliographic bias

Conclusion

Little evidence for sociolinguistic effects

- when family and areal effects are properly controlled for, the effects of the sociolinguistic predictors (used here) are much less certain

We need more sociolinguistic data

Controlling for a potential bibliographic bias

Conclusion

Little evidence for sociolinguistic effects

- when family and areal effects are properly controlled for, the effects of the sociolinguistic predictors (used here) are much less certain

We need more sociolinguistic data

- of course, we cannot exclude that other sociolinguistic predictors / more information would show a more robust effect

Controlling for a potential bibliographic bias

Conclusion

Little evidence for sociolinguistic effects

- when family and areal effects are properly controlled for, the effects of the sociolinguistic predictors (used here) are much less certain

We need more sociolinguistic data

- of course, we cannot exclude that other sociolinguistic predictors / more information would show a more robust effect

Controlling for a potential bibliographic bias

- include information on the length of the descriptions to control for a potential bibliographical bias

Thank you!

References I

Atoyebi, Joseph Dele (2010). *A Reference Grammar of Oko.* Köln: Köppe.

Bakker, Dik and Ewald Hekking (2012). "Clause Combining in Otomi before and after Contact with Spanish." In: *Linguistic Discovery* 10.1, pp. 42–61.

Bell, Alan (1978). "Language Samples." In: *Universals of Human Language. Volume 1: Method and Theory.* Ed. by Joseph H. Greenberg and Charles Albert Ferguson. Stanford, CA: Stanford University Press, pp. 123–156.

Bentz, Christian and Bodo Winter (2013). "Languages with More Second Language Learners Tend to Lose Nominal Case." In: *Language Dynamics and Change* 3, pp. 1–27.

Biber, Douglas (2006). *University Language: A Corpus-Based Study of Spoken and Written Registers.* Amsterdam: Benjamins.

Biber, Douglas (2009). "Are There Linguistic Consequences of Literacy? Comparing the Potentials of Language Use in Speech and Writing." In: *The Cambridge Handbook of Literacy.* Cambridge: Cambridge University Press.

Bickel, Balthasar (2008). "A Refined Sampling Procedure for Genealogical Control." In: *Sprachtypologie und Universalienforschung* 61.3, pp. 221–233.

Bickel, Balthasar (2011). "Statistical Modeling of Language Universals." In: *Linguistic Typology* 15.2, pp. 401–413.

Blasi, Damián et al. (2019). "Human Sound Systems Are Shaped by Post-Neolithic Changes in Bite Configuration." In: *Science* 363.6432.

Chafe, Wallace (1982). "Integration and Involvement in Speaking, Writing, and Oral Literature." In: *Spoken and Written Language: Exploring Orality and Literacy.* Ed. by Deborah Tannen. Norwood, NJ: Ablex, pp. 35–53.

Chafe, Wallace and Deborah Tannen (1987). "The Relation between Written and Spoken Language." In: *Annual Review of Anthropology* 16, pp. 383–407.

Cysouw, Michael, Dan Dediu, and Steven Moran (Feb. 2012). "Supporting Online Material for: Comment on "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa"."

Dahl, Östen (2008). "An Exercise in a Posteriori Language Sampling." In: *Language Typology and Universals* 61.3, pp. 208–220.

De Busser, Rik and Randy LaPolla, eds. (2015). *Language Structure and Environment: Social, Cultural, and Natural Factors.* Amsterdam: Benjamins.

References II

Donohue, Mark and Bronwen Whiting (2011). "Quantifying Areality: A Study of Prenasalisation in Southeast Asia and New Guinea." In: *Linguistic Typology* 15.1, pp. 101–121.

Dryer, Matthew S. (1989). "Large Linguistic Areas and Language Sampling." In: *Studies in Language* 13.2, pp. 257–292.

Dryer, Matthew S. (1992). "The Greenbergian Word Order Correlations." In: *Language* 68.1, pp. 81–138.

Dryer, Matthew S. (2018). "On the Order of Demonstrative, Numeral, Adjective and Noun." In: *Language* 94.4, pp. 798–833.

Guzmán Naranjo, Matías and Laura Becker (accepted). "Statistical Bias Control in Typology." In: *Linguistic Typology*.

Hammarström, Harald and Mark Donohue (2014). "Some Principles on the Use of Macro-Areas in Typological Comparison." In: *Language Dynamics and Change* 4.1, pp. 167–187.

Hammarström, Harald, Robert Forkel, et al. (2020). *Glottolog 4.3.* Jena: Max Planck Institute for the Science of Human History.

Hammarström, Harald and Tom Güldemann (2014). "Quantifying Geographical Determinants of Large-Scale Distributions of Linguistic Features." In: *Quantifying Language Dynamics*, pp. 67–95.

Hellwig, Birgit (2011). *A Grammar of Goemai.* Berlin: De Gruyter.

Jaeger, T. Florian et al. (Nov. 2011). "Mixed Effect Models for Genetic and Areal Dependencies in Linguistic Typology." In: *Linguistic Typology* 15.2, pp. 281–319.

Karlsson, Fred, Matti Miestamo, and Kaius Sinnemäki (2008). *Linguistic Complexity. Typology, Contact, Change.* Amsterdam: Benjamins.

Koppen, Kim, Mirjam Ernestus, and Margot van Mulken (2019). "The Influence of Social Distance on Speech Behavior: Formality Variation in Casual Speech." In: *Corpus Linguistics and Linguistic Theory* 15.1, pp. 139–165.

Kortmann, Bernd, ed. (1997). *Adverbial Subordination: A Typology and History of Adverbial Subordinators Based on European Languages.* Berlin: De Gruyter.

Ladd, D Robert, Seán G Roberts, and Dan Dediu (2015). "Correlational Studies in Typological and Historical Linguistics." In: *Annual Review of Linguistics* 1, pp. 221–241.

Levshina, Natalia (2019). "Token-Based Typology and Word Order Entropy: A Study Based on Universal Dependencies." In: *Linguistic Typology* 23.3, pp. 533–572.

References III

Lupyan, Gary and Rick Dale (2010). "Language Structure Is Partly Determined by Social Structure." In: *PLoS ONE* 5.1, e8559.

Lupyan, Gary and Rick Dale (2016). "Why Are There Different Languages? The Role of Adaptation in Linguistic Diversity." In: *Trends in Cognitive Sciences* 20.9, pp. 649–660.

Martowicz, Anna (2011). "The Origin and Functioning of Circumstantial Clause Linkers: A Cross-Linguistic Study." PhD thesis. Edinburgh: University of Edinburgh.

Matras, Yaron (2008). "The Borrowability of Structural Categories." In: *Grammatical Borrowing in Cross-Linguistic Perspective*. Ed. by Yaron Matras and Jeanette Sakel. Berlin: De Gruyter, pp. 31–74.

Meakins, Felicity and Rachel Nordlinger (2013). *A Grammar of Bilinarra, an Australian Aboriginal Language of the Northern Territory*. Berlin: De Gruyter.

Mithun, Marianne (1984). "How to Avoid Subordination." In: *Annual Meeting of the Berkeley Linguistics Society* 10.0, pp. 493–509.

Nettle, Daniel (July 2012). "Social Scale and Structural Complexity in Human Languages." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1597, pp. 1829–1836.

Nichols, Johanna (1992). *Linguistic Diversity in Space and Time*. Chicago: The University of Chicago Press.

Nichols, Johanna (2009). "Linguistic Complexity: A Comprehensive Definition and Survey." In: *Language Complexity as an Evolving Variable*. Ed. by Geoffrey Sampson, David Gil, and Peter Trudgill. Oxford: Oxford University Press.

Pensalfini, Robert J. (1997). "Jingulu Grammar, Dictionary, and Texts." PhD thesis. Massachusetts, MA: MIT.

Perkins, Revere (1992). *Deixis, Grammar and Culture*. Amsterdam: Benjamins.

Rijkhoff, Jan et al. (1993). "A Method of Language Sampling." In: *Studies in Language* 17.1, pp. 169–203.

Sinnemäki, Kaius (Oct. 2020). "Linguistic System and Sociolinguistic Environment as Competing Factors in Linguistic Variation: A Typological Approach." In: *Journal of Historical Sociolinguistics* 6.2.

Sinnemäki, Kaius and Francesca Di Garbo (2018). "Language Structures May Adapt to the Sociolinguistic Environment, but It Matters What and How You Count: A Typological Study of Verbal and Nominal Complexity." In: *Frontiers in Psychology* 9.

Thompson, Sandra A., Robert E. Longacre, and Shin Ja J. Hwang (2007). "Adverbial Clauses." In: *Language Typology and Syntactic Description. Volume 2*. Ed. by Timothy Shopen. Cambridge: Cambridge University Press, pp. 237–300.

Trudgill, Peter (2008). "Linguistic and Social Typology." In: *The Handbook of Language Variation and Change*. London: Wiley, pp. 707–728.

Trudgill, Peter (Nov. 2011a). "Social Structure and Phoneme Inventories." In: *Linguistic Typology* 15.2, pp. 155–160.

Trudgill, Peter (2011b). *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry (2017). "Practical Bayesian Model Evaluation Using Leave-One-out Cross-Validation and WAIC." In: *Statistics and Computing* 27.5, pp. 1413–1432.

Wray, Alison and George W. Grace (Mar. 2007). "The Consequences of Talking to Strangers: Evolutionary Corollaries of Socio-Cultural Influences on Linguistic Form." In: *Lingua* 117.3, pp. 543–578.

Linguistic annotation

- type: real/hypothetical (N=455), general (N=364) , counterfactual (N=136)
- N other uses of the marker
- marker that is part of the verbal inflection paradigm
- form & position of the marker
- balanced / deranked construction
- N constructions per language
- N markers per language
- N dedicated constructions per language
- ...

Multilingual speakers

- for 213 languages, we have information on multilingual speakers
- more multilingual speakers → more borrowings → more conditional constructions
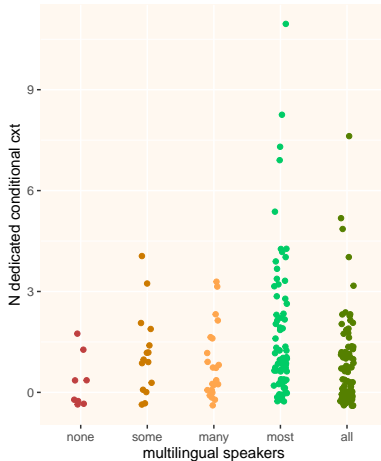
☞ however, in comparison to the other predictors (log_speaker, education), this is not a relevant predictor

Association between sociolinguistic factors

- the different sociolinguistic factors are associated with each other
- ☞ potential collinearity!
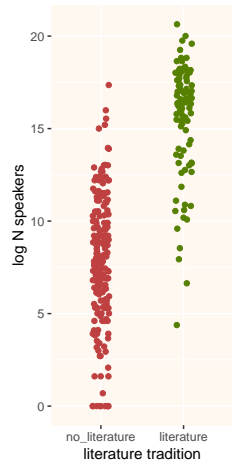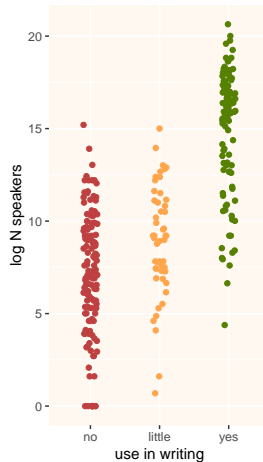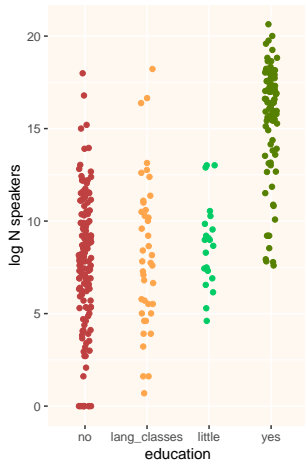
Phylogenetic term

	Hindi	German	Dutch	Spanish	French	Italian	Seri	Chukchi
Hindi	1.00	0.67	0.67	0.67	0.67	0.67	0	0
German	0.67	1.00	0.83	0.67	0.67	0.67	0	0
Dutch	0.67	**0.83**	1.00	0.67	0.67	0.67	0	0
Spanish	0.67	0.67	0.67	1.00	0.91	0.90	0	0
French	0.67	**0.67**	0.67	**0.91**	1.00	0.90	0	0
Italian	0.67	0.67	0.67	0.90	0.90	1.00	0	0
Seri	0	0	0	0	0	0	1.00	0
Chukchi	0	0	0	0	0	0	0	1.00

Gaussian process: A toy example