

Reading group: (Barth et al. 2021)

**G. Moroz**

Linguistic Convergence Laboratory, NRU HSE

13 September 2022

# Outline of the talk

Context of the paper

Goal of the paper

Modeling

Domains

## Family Problems picture task, (San Roque et al. 2012)

This task was developed for a research of the grammaticized expression of social cognition, that combines

- “social facts” such as kinship relationships;
- “psychological facts” (such as feelings, desires, or attentional states);
- and something else.

## Family Problems picture task, (San Roque et al. 2012)

This task was developed for a research of the grammaticized expression of social cognition, that combines

- “social facts” such as kinship relationships;
- “psychological facts” (such as feelings, desires, or attentional states);
- and something else.

Authors wanted to create a task that stimulates production

- of different speech types (descriptions, dialogic interactions, narratives);
- of linguistic expressions relevant to the domain of social cognition;
- of some known interesting grammatical categories (e.g. evidentiality);
- of parallel but spontaneous texts (in the broad sense) across different languages.

## Family Problems picture task, (San Roque et al. 2012)

“Two participants (A and B) sit together, either side-by-side or oriented slightly toward each other, with an open flat space (e.g., the ground, a table) in front of them. The activity involves three components:

- i.* the participants are given the pictures one at a time and describe the content of each;
- ii.* the participants are asked to arrange the pictures as an ordered narrative, working together;
- iii.* participants are asked to tell the story from start to finish, from both third- person and first-person perspectives, typically to an audience that had not seen the pictures or heard the story before.” (San Roque et al. 2012: 144–145)

# Canonical order defined by (San Roque et al. 2012)



# Presentation order defined by (San Roque et al. 2012)



# The Social Cognition Parallax<sup>1</sup> Interview Corpus (SCOPIC)

There are just seven languages in (San Roque et al. 2012). The main drivers of this project (D. Barth and N. Evans) got funding, and as a result in (Barth and Evans 2017) they claimed to have 24 (and over 30 languages now).

- [corpus website](#);
- recordings and some annotations are available in [PARADISEC catalog](#);
- 31 publications including (Barth et al. 2021).

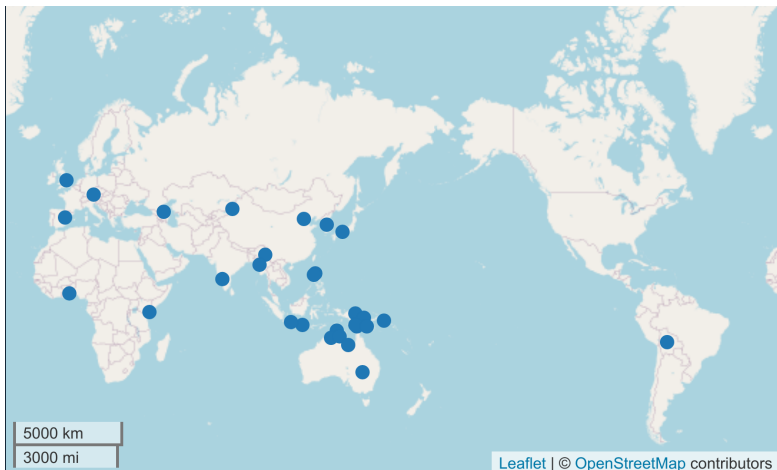
---

<sup>1</sup>By ‘parallax corpus’ we mean ‘broadly comparable formulations resulting from a comparable task’, to avoid the implications of ‘parallel corpus’ that there will be exact semantic equivalence across languages. (Barth and Evans 2017: 1)



# The Social Cognition Parallax Interview Corpus (SCOPIC)

30 languages of the database (present in PARADISEC catalog):



created with lingtypology

# The Social Cognition Parallax Interview Corpus (SCOPIC)

13 languages from (Barth et al. 2021):



created with `lingtypology`

# Outline of the talk

Context of the paper

Goal of the paper

Modeling

Domains

## Goal of the paper

The authors make it clear on the second page of the paper:

- “Can we measure the relative contributions of language structures and individual variation to how people formulate statements about the world?”
- Do accounts of typological differences need to take individual variation into account, and is such variation more prevalent in some kinds of linguistic domains than others?” (Barth et al. 2021: 181)

## Goal of the paper

The authors make it clear on the second page of the paper:

- “Can we measure the relative contributions of language structures and individual variation to how people formulate statements about the world?”
- Do accounts of typological differences need to take individual variation into account, and is such variation more prevalent in some kinds of linguistic domains than others?” (Barth et al. 2021: 181)

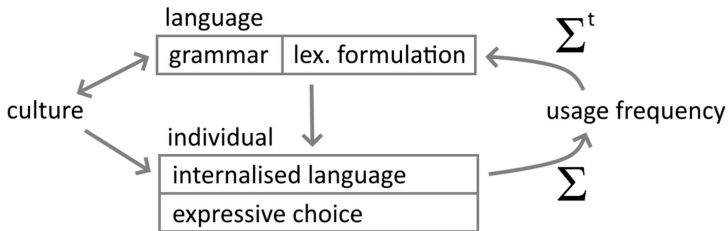


Figure 1: Language and individuals affect usage frequency and each other. 

# Outline of the talk

Context of the paper

Goal of the paper

Modeling

Domains

## Variation

Imagine some variation pattern (e.g. [pə'teɪ.tou] vs [pə'taɪ.tou]).

- some speakers (Type 1) could use just one variant

## Variation

Imagine some variation pattern (e.g. [pə'teɪ.tou] vs [pə'taɪ.tou]).

- some speakers (Type 1) could use just one variant
- some speakers (Type 2) can use both variants with some probability... 0.7 of [eɪ]



# Variation

Imagine some variation pattern (e.g. [pə'teɪ.tou] vs [pə'taɪ.tou]).

- some speakers (Type 1) could use just one variant
- some speakers (Type 2) can use both variants with some probability... 0.7 of [eɪ]
- there should be some proportion of speakers of Type 1 (0.3 - [eɪ], 0.1 - [aɪ]) and Type 2 (0.6)

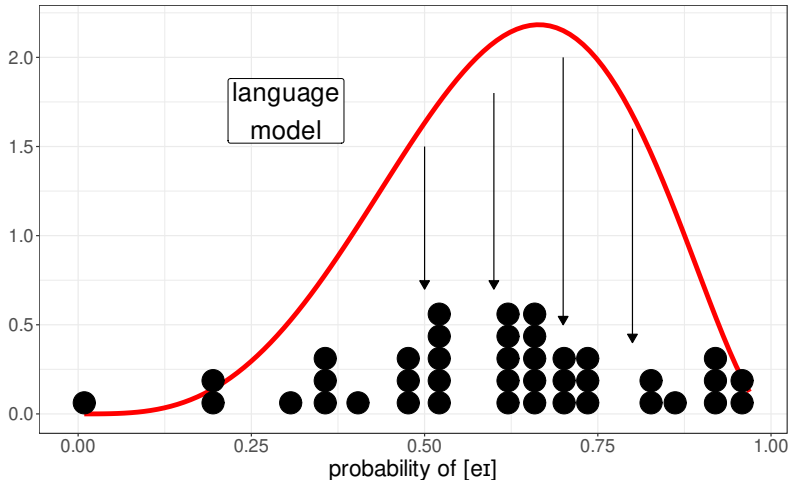
# Variation

Imagine some variation pattern (e.g. [pə'teɪ.tou] vs [pə'taɪ.tou]).

- some speakers (Type 1) could use just one variant
- some speakers (Type 2) can use both variants with some probability... 0.7 of [eɪ]
- there should be some proportion of speakers of Type 1 (0.3 - [eɪ], 0.1 - [aɪ]) and Type 2 (0.6)
- But we don't know anything like this just observe real usage!

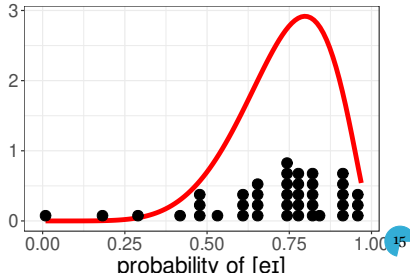
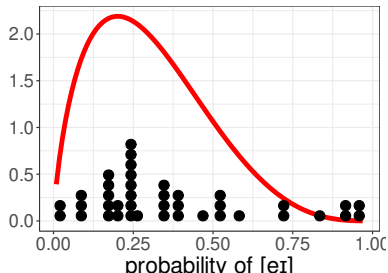
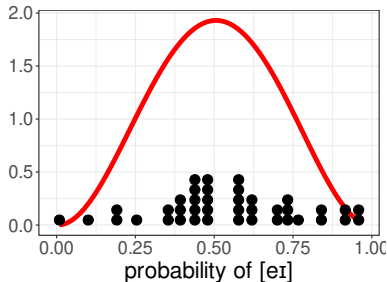
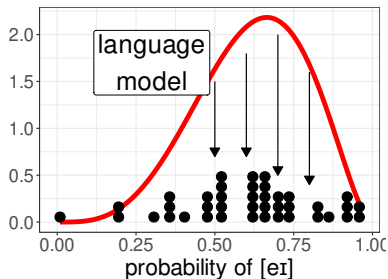
## Variation

So we can run some tests, obtain some results and create a language model:



# Variation

We can expect slightly different models for different English dialects:



## Modeling in (Barth et al. 2021)

- In (Barth et al. 2021) they use logistic mixed effect regressions:

name	PREDICT	fixed ef.	random ef.	random ef.
<i>Mod-L</i>	variation	language		stimuli card
<i>Mod-LI</i>	variation	language	individual	stimuli card
<i>Mod-I</i>	variation		individual	stimuli card

- Compare models

## Modeling in (Barth et al. 2021)

- In (Barth et al. 2021) they use logistic mixed effect regressions:

name	PREDICT	fixed ef.	random ef.	random ef.
<i>Mod-L</i>	variation	language		stimuli card
<i>Mod-LI</i>	variation	language	individual	stimuli card
<i>Mod-I</i>	variation		individual	stimuli card

- Compare models
- Show model estimates with confidence intervals for the contribution of language

# Outline of the talk

Context of the paper

Goal of the paper

Modeling

Domains

## Domains examined

In the paper they examined four domains:

- Propositional framing (grammatical), 6 languages: usage of subordination vs paratactic structures.
- Reported speech, thought, and action constructions (grammatical), 12 languages: usage of direct vs indirect speech.
- Reported speech, thought, and action predicates (lexical/semantic), 12 languages: usage of cognitive vs. quotative predicates.
- Human reference lexical choice (semantic/lexical), 11 languages: usage of kinship terms vs. others



- *Mod-LI* is the best across all four domains. They make a conclusion that researchers need to take into account individual variation in typological studies.
- All analyzed variables are scales and languages/individuals take different position on those scales.
- Some languages seem to have more inter-speaker variability than others.
- Genre and the picture card stimuli influence linguistic behaviour for semantic domains, but not for grammatical ones.

- D. Barth and N. Evans. Scopic: design and overview. *Language Documentation and Conservation*, (12), 2017.
- D. Barth, N. Evans, I. W. Arka, H. Bergqvist, D. Forker, S. Gipper, G. Hodge, E. Kashima, Y. Kasuga, C. Kawakami, Y. Kimoto, D. Knuchel, N. Kogura, K. Kurabe, J. Mansfield, H. Narrog, D. P. E. Pratiwi, S. van Putten, C. Senge, and O. Tykhostup. Language vs. individuals in cross-linguistic corpus typology. In G. Haig, S. Schnell, and F. Seifart, editors, *Doing corpus-based typology with spoken language data : State of the art*, pages 179–232. University of Hawai'i Press, Honolulu, 2021. URL <http://hdl.handle.net/10125/74661>.

- L. San Roque, L. Gawne, D. Hoenigman, J. Miller, S. Spronck, A. Rumsey, A. Carroll, and N. Evans. Getting the story straight: Language fieldwork using a narrative problem-solving task. *Language Documentation and Conservation*, 6:135–174, 2012. URL <http://hdl.handle.net/10125/4504>.