Acquiring inflectional gaps with indirect negative evidence: Evidence from Russian

Kyle Gorman & Daniel Yakubov (CUNY Graduate Center)

Inflectional gaps

Children must learn not only what they can say, but also what they cannot.

- In Russian there are roughly 40 nouns which lack a well-formed gen.pl. (Pertsova 2005, Zaliznyak 1977) and roughly 70 verbs which lack a well-formed 1sg. non-past (Halle 1973, Pertsova 2016).

(1)

nom.sg.	gen.pl.	
дно	*дн, *ден	'bottom (of s.t.)'
кочерга	*кочерг	'fire poker'
мечта	*мечт	'dream; fantasy'

(2)

infinitive	1sg. non-past	
дерзить	*держу	'be insolent'
победить	*побежу, *побежду	'win; overcome'
пылесосить	*пылесошу	'vacuum'

- There is no consensus about how such inflectional gaps are represented grammatically (e.g., Rice & Blaho 2009, Baerman et al. 2010, Gorman & Yang 2019, Sims 2005); even less is known about how they are acquired.
- There is no evidence that inflectional gaps are learned via *direct* negative evidence (e.g., they are not part of Russian prescriptive norms) but *indirect* negative evidence (INE) has been suggested as a possible mechanism (e.g., Orgun & Sprouse 1999, Pertsova 2005, Daland et al. 2007).
- Reviewer #4: *I think there is a large group of scientists who have the intuition "Obviously, statistics and indirect negative evidence can easily account for the acquisition of inflectional gaps!"* […] *On the other hand, I think there are also many people who have the intuition, "Clearly, statistics alone can't differentiate inflectional gaps from forms that are grammatical but accidentally absent."*

Our contribution

- We perform a realistic simulation of INE learning using morphologically annotated Russian text.
  - Child-directed speech would be ideal, but available corpora are not nearly large enough.
  - Our simulation is provided with orders of magnitude more than children encounter during the critical period, and thus we are giving the INE hypothesis the **maximal benefit of the doubt**.

Indirect negative evidence

We propose three naïve models of how a child might infer the presence of an inflectional gap using INE.

The models

- Let *w* denote a (possible or attested) inflected word, with a corresponding lemma *l* (representing the lexeme) and morphosyntactic feature bundle *f* (representing a cell in the lexeme's inflectional paradigm).
- Let p_w, p_l, and p_f denote the probabilities of a word, lemma, and feature bundle, respectively; these probabilities can be estimated using maximum likelihood estimation from a morphologically annotated corpus.
- Under a naïve independence assumption, $p_w = p_l \times p_f$.
- Alternatively, if p_w is substantially less than $p_l \times p_f$ we posit that the corresponding cell in the paradigm is a gap.
- We consider three ways to quantify "substantially less":

(3) *absolute divergence*:

$$\delta_w = p_w - p_l \times p_f$$

(4) *standard score*:

$$z_w = \delta_w / \sqrt{[p_f^2 \times s_l^2 + p_l^2 \times s_f^2 + s_l^2 \times s_f^2]}$$

where s^2 is sample variance, and

(5) *log-odds ratio*:

$$L_w = \log (p_w / p_l \times p_f) = \log p_w - \log p_l - \log p_f.$$

The data

- We train state-of-the-art neural morphological analyzer UDTube (UDTube; Yakubov 2024) on the 1.5m-sentence SynTagRus corpus (Droganova et al. 2018) with 25 rounds of Bayesian hyperparameter tuning.
- We use the best model to morphologically analyze 232m sentences (3b tokens) of Russian web text filtered from the CC-100 corpus (Wenzek et al. 2020).
- We use the analyzed corpus to compute metrics (3-5) for all gen.pl. nouns and 1sg. non-past verbs, and then fit a *stump classifier* to find any possible threshold for distinguishing defective and non-defective forms.

Results

Morphological analysis

- On held-out Russian data, the UDTube morphological analyzer achieves 99% POS accuracy, 98% lemmatization accuracy, and 95% morphological feature accuracy.

Attestation

- Out of the 36 gen.pl. nouns labeled *затрудн.* 'difficult' or *нет* 'does not exist' in Zaliznyak's (1977) dictionary:
  - 34 of these lemmas occur in at least one inflected form.
  - We find 107 tokens of **мечт* and few others (e.g., 13 tokens of **мулл* < *мулла* 'mullah').
- Out of the 65 1sg. non-past verbs listed as defective in Pertsova's (2016) appendix:
  - 50 of these lemmas occur in at least one inflected form.
  - We find 11 tokens of **чту* (< *чтить* 'to honor').

Acquisition

- For both cases:
  - Point-biserial correlation between defectivity and metrics (3-5) are all in [–.001, +.001].
  - There is no operating point where (3-5) effectively distinguishes between defective and non-defective forms.

Conclusions

- **Inflectional gaps cannot be discriminated on the basis of indirect negative evidence** using metrics (3-5).
- To acquire inflectional gaps, children must make use of additional information:
  - Albright (2003) and Gorman & Yang (2019), a.o., propose that inflectional gaps arise primarily in the presence of competing grammatical generalizations.
  - In Russian nouns, there are competing allomorphs of the gen.pl. and stem allomorphy introduced by *yer* deletion and stress shift (and concomitant reduction).
  - In the 1sg. non-past there are competing forms of stem-final mutation, and some English loanwords do not undergo them at all.

We will make available our software and database under appropriate open-source licenses to encourage others to development and evaluate alternative INE models.