Quantifying performance constraints on case-alignment typology with information theory

Steven **Foley**
✉ **Contact:** sfoley@princeton.edu
✱ **Info:** https://stevenrfoley.github.io/

Acknowledgements ❀ Thanks to Brian Dillon, Matt Wagers, and audiences at Princeton & UMass Amherst for helpful input.

References ❧ **[1]** Ackerman, F. & R. Malouf. 2013. *Language*, 89(3). **[2]** Cotterell, R. et al. 2019. *Transactions of the Association for Computational Linguistics*, vol 7. **[3]** Harris, A. 1985. *Diachronic Syntax: The Kartvelian case*. **[4]** Gippert, J. & M. Tandashvili. 2015. In *Historical Corpora*, eds. J. Gippert & R. Gehrke. **[5]** Dell, G. & F. Chang. 2014. *Philosophical Transactions of the Royal Society*, vol 369. **[6]** Futrell, R. et al. 2021. *Cognitive Science*, vol 44. **[7]** Levshina, N. 2021. *Frontiers in Psychology*, vol 12.

PRINCETON UNIVERSITY

Conditional entropy quantifies burdens case systems place on **producers and comprehenders**. Entropy values for a sample of **naturalistic and simulated** grammars are interpolated from Georgian **corpus data**. Attested case patterns are **simpler** than expected; **production constrains** more than comprehension.

1. Information theory & grammatical complexity

- Conditional entropy (i) measures the **unpredictability** of one variable given a known value of another variable.
  - Consider a pro-drop, scrambling SOV language with case. How hard is it to **inflect NP$_1$ for case** when speaking (ii)? How hard is it to **parse its syntactic role** while listening (iii)?

(i) $H(Y|X) = \Sigma_{x\in X}\Sigma_{y\in Y} p(x,y)\, I(y|x)$
$I(y|x) = -\log_2 p(y|x)$
$p(y|x) = p(x,y) / p(x)$

(ii) $H(Case|Role) \equiv H(k|r) \equiv$ **Production Burden**

(iii) $H(Role|Case) \equiv H(r|k) \equiv$ **Comprehension Burden**

- The **Low Conditional Entropy Conjecture** [1,2] posits a complexity ceiling on acquirable grammars.
  - Inflectional **paradigms can only be so irregular** before the most entropic forms are regularized during language transmission.
- Research questions
  - Do **syntagmatic** patterns like case alignment exhibit similarly constrained conditional entropy?
  - How might $H(k|r)$ and $H(r|k)$ shape morphosyntactic **typology**?

2. Case-alignment typology & Georgian split ergativity

- **Alignment** refers to the way **morphological** categories are associated with different **syntactic** positions.
  - Classic typology: Does case marking (α, β, γ) of intransitive subjects (**S**) pattern with that of transitive subjects (**A**) or direct objects (**P**)?

(iv)

	A	S	P
Nom-Acc	α		β
Erg-Abs	α	β	
Tripartite	α	β	γ

- Some languages like **Georgian** [3] have very complex case alignment systems, not neatly categorizable above (iv).
  - Seven syntactic roles: transitive (**A**), unergative (**Z**), unaccusative (**S**), and experiencer (**E**) subjects; patient (**P**) and theme (**T**) direct objects; indirect objects (**G**)
  - Three-way **split**, conditioned in different environments (tenses).

(v)

Macro-role	SU				DO		IO
Micro-role	A	Z	S	E	P	T	G
Env 1	α			β		α	β
Env 2	γ		α	β	α		β
Env 3	β		α	β	α		(β)

3. Georgian case complexity calculated with corpus data

- The relative frequencies of all case–role combinations were estimated from **Georgian National Corpus** data [4].

(vi)

Macro-role	SU				DO		IO
Micro-role	A	Z	S	E	P	T	G
Env 1	37,474	15,235	38,467	14,639	43,143	26,516	26,148
Env 2	61,928	6,572	48,717	3,596	75,825	5,356	36,980
Env 3	9,607	574	7,947	786	12,101	1,181	3,013

(vii)

Macro-role	SU				DO		IO
Micro-role	A	Z	S	E	P	T	G
Case *α*	0.079	0.032	0.200	0	0.185	0.07	0
Case *β*	0.020	0.001	0	0.040	0.091	0	0.129
Case *γ*	0.130	0.014	0	0	0	0	0

- Now calculate production & comprehension burdens of Georgian case alignment, over macro- & micro-roles.

(viii) $H(r_M|k) = 0.999$ $H(k|r_M) = 0.968$ $H(r_\mu|k) = 1.732$ $H(k|r_\mu) = 0.599$

4. Simulating a typology of case alignments

- Holding frequencies constant, corpus data were remapped to 17 attested alignments (ix) and >10k simulated ones (x).

(ix)

	A	Z	S	E	P	T	G
Env 1	α			β	γ	α	β
Env 2	α			β	γ	α	β
Env 3	α			β	γ	α	β

Naturalistic sample: Case alignment attested in Basque, Batsbi, Cebuano, Chamorro, Chechen, Hindi, Icelandic, Inuktitut, Laz, Lezgian, Megrelian, Nez Perce, Russian, Sakha, Shipibo, Svan, Tabasaran

(x)

	A	Z	S	E	P	T	G
Env 1	α	β	γ	α	β	γ	α
Env 2	γ	α	β	γ	α	β	γ
Env 3	β	γ	α	β	γ	α	β

Simulated sample: All 301 logically possible 3-case/7-role non-split alignments, and 10k randomly generated 2- or 3-way split alignments.

- Except for non-split $H(k|r)$s, all attested samples are less entropic than simulated ones ($p < 0.05$, via Welch's ind. samples t-tests).

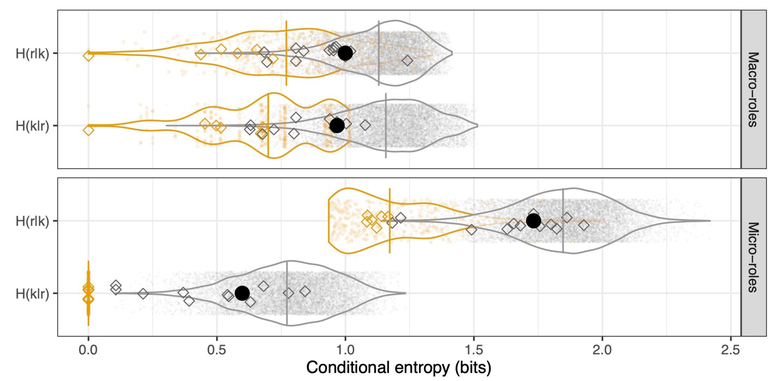
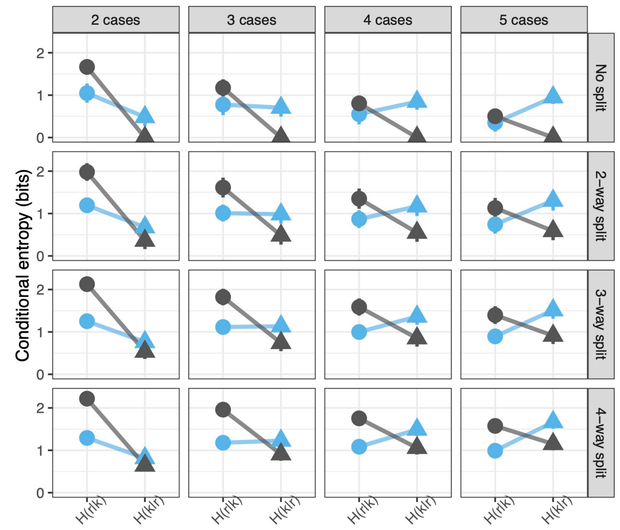


Figure 1: Conditional entropy estimates for Georgian (black dots), naturalistic sample (diamonds), and simulated sample (violins/clouds). Values for split alignments are grey/black; values for non-split alignments are gold.

5. Effects of case-inventory size and number of splits

- Question for **'theoretical typology'**: how do grammatical factors contribute to complexity, independent of alignment?
  - Languages tend to have ~3 core case categories, and no more than a two-way alignment split. Is this a coincidence?
- To explore this question: 16 new simulated typologies of 5k alignments, varying in numbers of cases and splits



6. Implications for theories of case alignment

- Standard case typology is descriptive, taxonomical. Information theory provides an **explanatory foothold**.
  - Sentence-processing theories like the P-Chain [5] might explain a complexity ceiling on case: more entropic grammars are harder to use, making acquisition channels noisier [cf. 6].
- Tentative conclusions
  - Attested case-alignment patterns are **less entropic** than they could be (unsurprising — how would very complex ones arise?)
  - Case is **more streamlined for production** than comprehension, especially when calculated over syntactic micro-roles.
  - Systems with **more cases** benefit comprehension but impede production; **more splits** make everything harder.
  - But **2-case** languages are the hardest to parse — and rarer?
- Limitations of this approach: highly **abstract**, doesn't account for **other cues** like word order or animacy [cf. 7].
  - Tantalizing observation: the most entropic languages in the sample are **verb-initial**!