Fossilized Prefixes, Living Compounds?

Productivity and Semantic Shift in Khmer Word Formation

Tianyi Ni (ni.386@buckeyemail.osu.edu)

Department of Linguistics
The Ohio State University

SEALS 35 @ Nanyang Technological University

Khmer word formation combines residual prefix-like material, minor-syllable structure, and several types of compounding.

- Khmer is often described as relatively analytic, but it preserves older Austroasiatic derivational morphology (Alves, 2015a,b, 2019; Jenny and Sidwell, 2014).
- Minor syllables and presyllables make recurrent left-edge material analytically separable from stems, but initial material is not automatically an affix (Pittayaporn, 2015; Butler, 2014, 2015).
- Khmer compounding roughly includes subordinate, coordinate, decorative, and reduplicative patterns (Ourn and Haiman, 2000; Haiman, 2013; Ourn and Haiman, 2003).

Received view: historical prefixation is reduced or fossilized; compounding remains synchronically active.

Question: is this received view really true, and how is each category organized?

Main Claim

Some prefix families look regular, but they do not become fully compound-like once dispersion is size-matched.

Roadmap

Three studies separate size, productivity, and semantic organization.

---

1. **Productivity vs. family size**
   Do prefix families and compound families have different productivity geometry?
2. **Productivity vs. semantic shift coherence**
   Is productivity tied to semantic organization?
3. **Semantic-shift entropy**
   Do larger families spread through semantic space in the same way?

Method Summary

Corpus frequency and distributional semantics are measured on the same families.

1. Extract Khmer word types and token frequencies from Khmer Wikipedia.
2. Use those counts to estimate family size V, family token frequency N, and hapax productivity P, following corpus-based productivity measures (Baayen, 1992, 2009).
3. Use pretrained fastText word vectors for complex forms, stems, and compound components; items without vector coverage are excluded from semantic tests (Bojanowski et al., 2017; Grave et al., 2018).
4. Compute semantic transparency and semantic-shift vectors in distributional space (Bonami and Paperno, 2018; Shen and Baayen, 2022).

Family-level shift coherence and entropy are then computed from those stem-to-word or other-component-to-word shift vectors.

Khmer Wikipedia supplies word types and frequencies; word vectors supply semantic measures.

Class	Word types	Family units
Prefix families	195 forms; 242 pairs	21 labels; 9 broader families
Subordinate compounds	995	562 first; 692 second
Coordinate compounds	264	377 pooled members
Decorative compounds	238	139 first; 223 second
Reduplicative compounds	13	small comparison class

- "Family" is operational: it groups comparable word-formation items without assuming a fully productive morpheme.
- Compounds are coded into four descriptive categories, following Khmer descriptive work (Ourn and Haiman, 2000; Haiman, 2013; Ourn and Haiman, 2003).
  - **Subordinate**: asymmetric component relation; head + dependent, where one component narrows or elaborates the other.
  - **Coordinate**: symmetrical relation; both components have parallel semantic status.
  - **Decorative**: recurrent sound/form patterning with weaker compositionality.
  - **Reduplicative**: repeated or near-repeated material; small comparison class.

Study 1 Question

Do productivity statistics separate prefix families from compound families?

- We expect larger family size V to be associated with lower hapax productivity P, because extent of use and current low-frequency productivity can diverge (Baayen, 1992, 2009; Shen and Baayen, 2022).

- Plot each family in the $(\log V, \log P)$ plane.
- If well-accepted productivity measures are diagnostic, prefix families and compound families should diverge.

Extent of use and low-frequency productivity are measured separately.

Family size

V = number of attested types

- how large the family is
- realized extent of use
- English *un-* family: *unhappy*, *unfair*, *unclear*

Hapax productivity

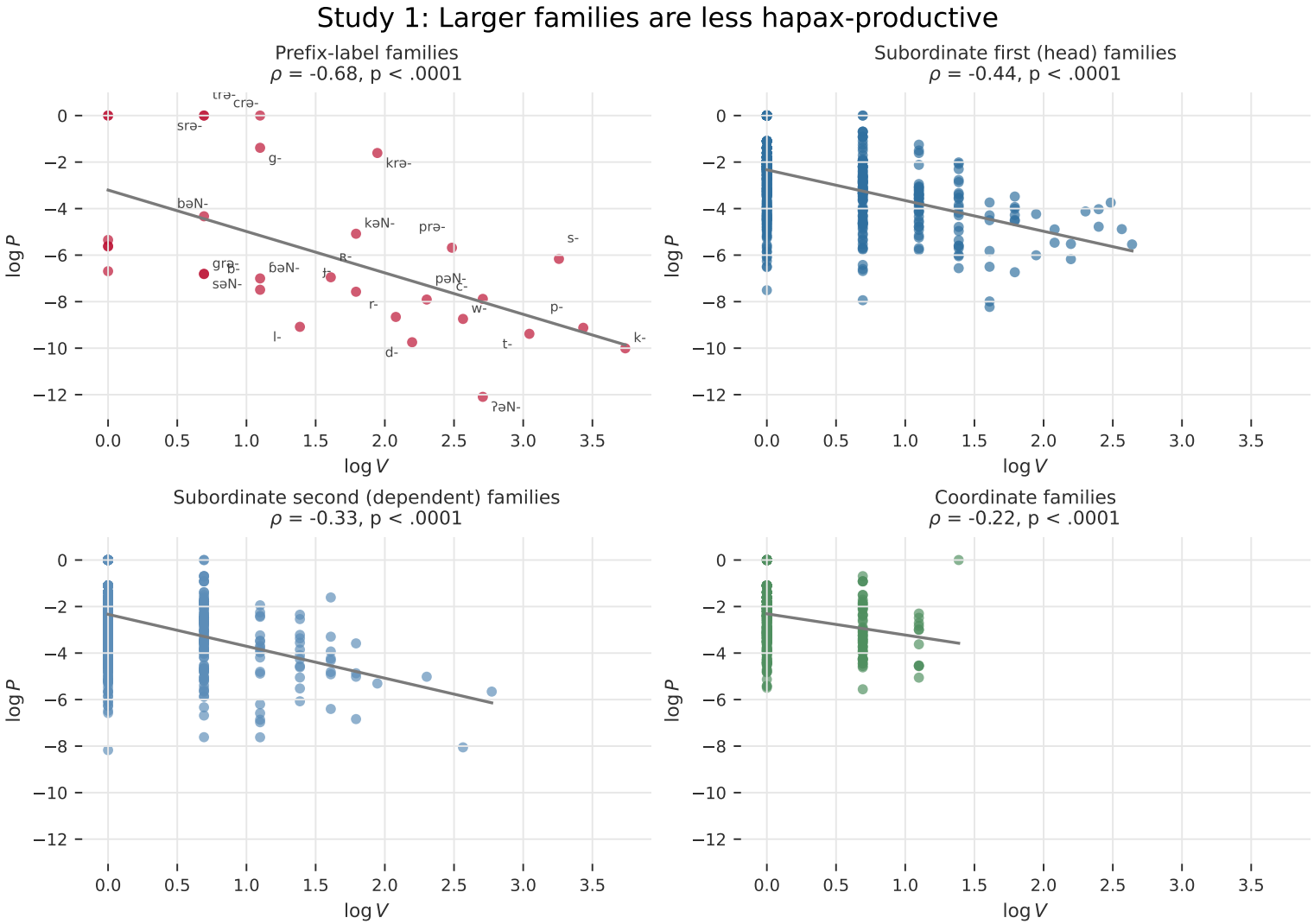
$$P = \frac{V_1 + 1}{N + 1}$$

- V_1: hapax types
- N: family token frequency

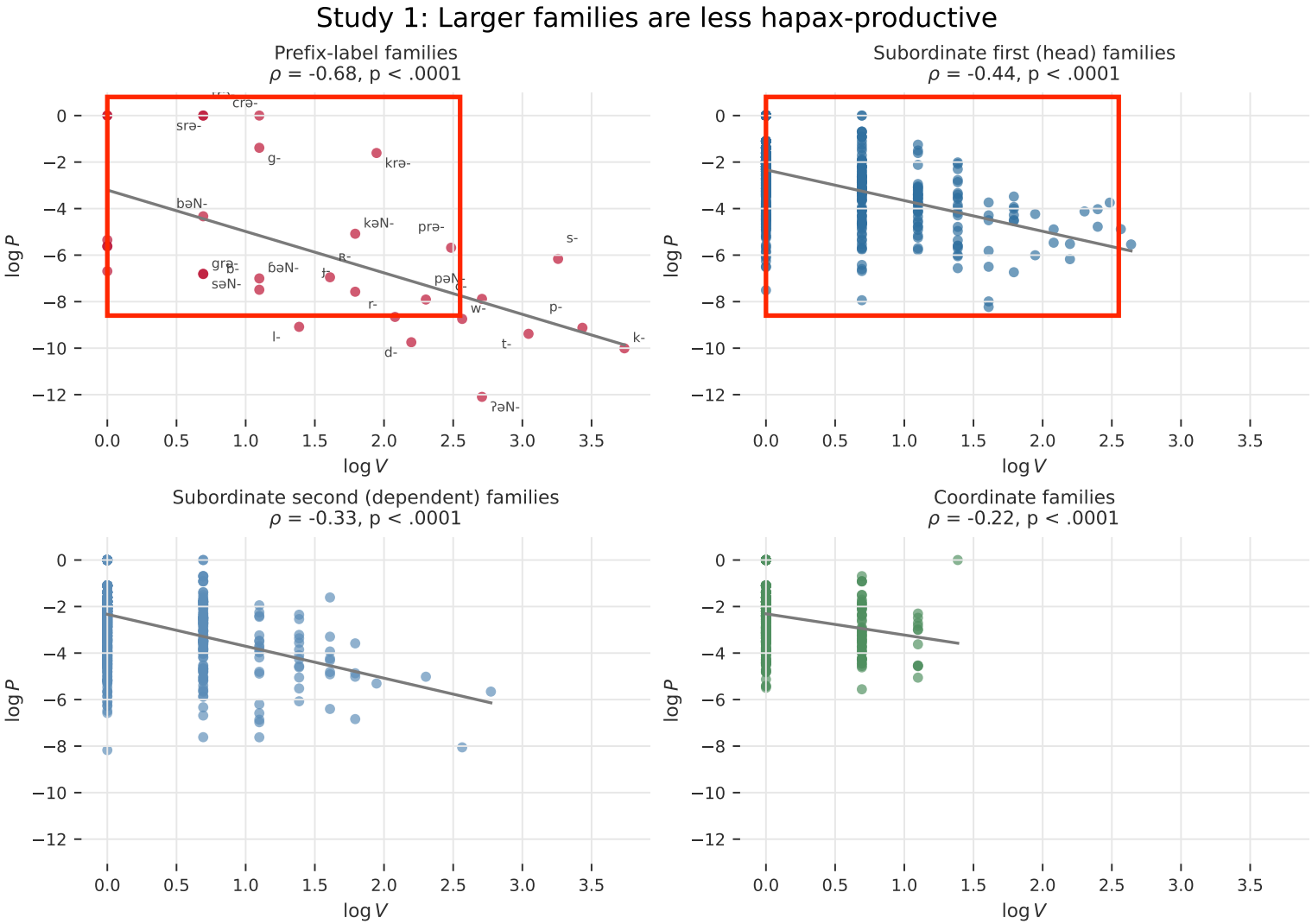
Each word type counts once; repeated tokens do not increase V.

The logic follows Baayen's corpus-based productivity measures (Baayen, 1992, 2009).

Study 1 Result



Study 1 Result



Productivity measures do not separate prefix families from compound families.

- Prefix families and subordinate first (head) families both show the same V–P geometry.
  - In both cases, larger families tend to have lower hapax productivity.
  - This separation between extent of use and current productivity is also central in corpus studies of English derivation (Baayen, 1992, 2009), German word formation (Stupak and Baayen, 2022), and Mandarin Chinese compounds (Shen and Baayen, 2022).
- Some prefix families therefore look like subordinate first (head) compound families under the productivity measures.
  - Relatively higher-P prefix families such as *bəN-*, *kəN-*, and *prə-* overlap with subordinate first (head) families in the productivity plane.
  - Large saturated prefix families such as *k-*, *p-*, and *t-*.

Study 2 Question

If productivity statistics are not enough, does meaning separate the systems?

- A general expectation in productivity research is that more productive patterns are more semantically **transparent** or **regular** (Baayen and Lieber, 1991; Baayen, 1992; Bauer, 2001); distributional work makes this testable in vector space (Shen and Baayen, 2022).

Semantic transparency

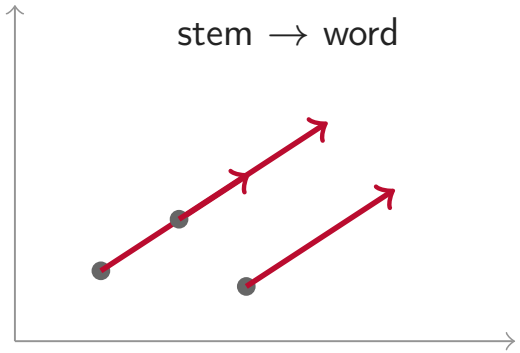
- Is a prefixed form close to its stem?
  - × no word vectors for prefixes themselves
- Is a compound close to its combined components?

Shift coherence

- Do family members move in a similar semantic direction?
- This asks a more morphological question.

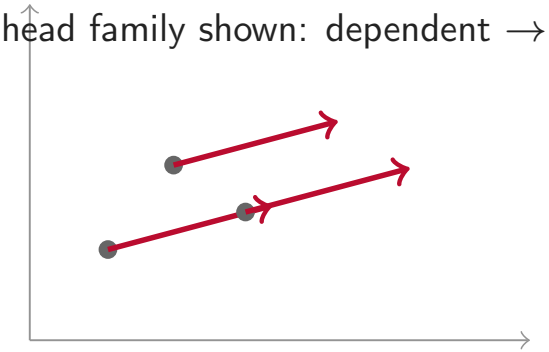
Shift coherence asks whether family members move in the same semantic direction.

Prefix-label families



Group: same prefix-like label ℓ
Shift: $\Delta_i = \overrightarrow{\text{word}}_i - \overrightarrow{\text{stem}}_i$

Subordinate component families

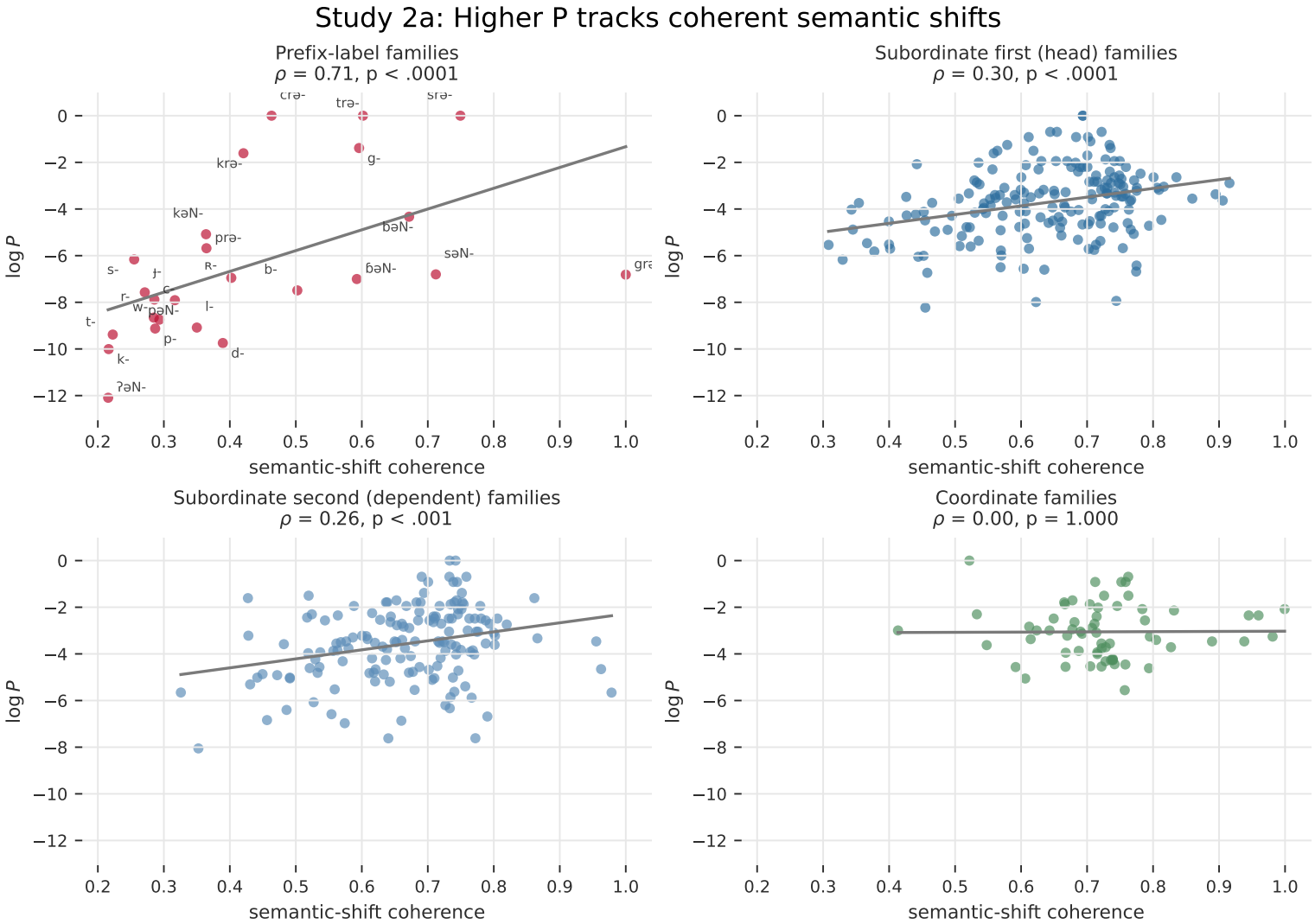


Group: same head/pivot h shown
Shift: $\Delta_i = \overrightarrow{\text{word}}_i - \overrightarrow{\text{dep}}_i$

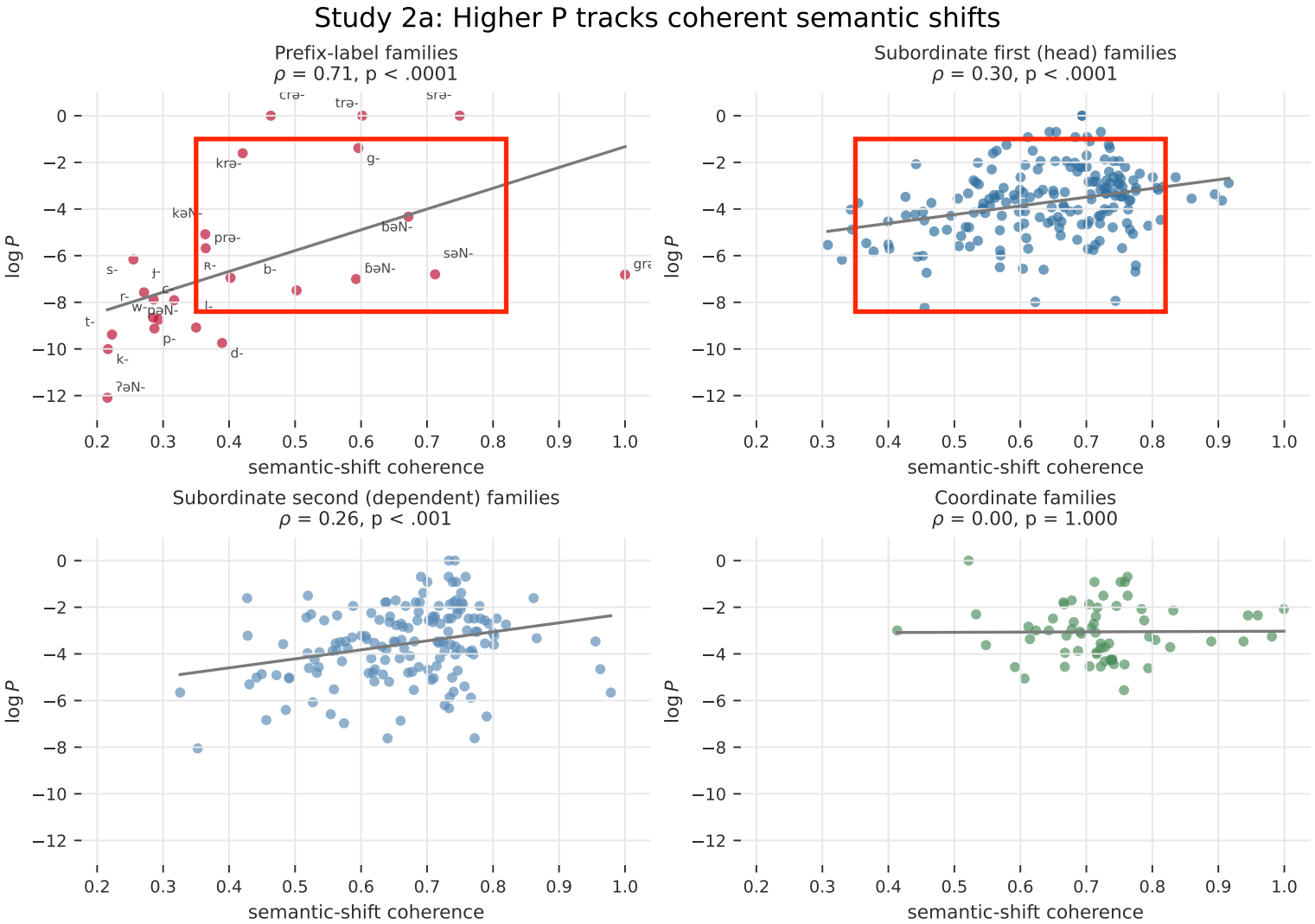
$$\bar{\Delta}_F = \frac{1}{m}\sum_{i=1}^{m}\Delta_i \qquad C(F) = \frac{1}{m}\sum_{i=1}^{m}\cos(\Delta_i, \bar{\Delta}_F)$$

Prefix side: grouped by shared prefix-like label, measured by stem–word shifts. Compound side: grouped by shared component, measured by other-component–word shifts.

Study 2a Result: Shift Coherence



Study 2a Result: Shift Coherence



Higher-P families tend to have coherent semantic shifts.

- Prefix-label families show the strongest coherence–productivity relation.
  - This means higher-P prefix labels contain stem–derived-form pairs that shift in a consistent semantic direction.
  - They are therefore semantically coherent, not simply random lexical residues.
- Subordinate first (head) and second (dependent) compound families show weaker but reliable parallels.
  - This means a shared compound component can organize a productive family when the other-component-to-word shifts are consistent.

× Coordinate families do not show the same relation.

Examples in the overlapping region include rhotacized labels such as *srə-*/*trə-* and nasal labels such as *səN-*/*bəN-*, which preserve organized stem-to-derived-form relations without implying that all prefix-like material is synchronically productive.

Study 3 Question

Since prefix and subordinate families both show semantic shifts, how do those shifts spread as families get larger?

- **Regular semantic shift** → shift vectors concentrate in fewer regions → lower entropy.
    - Large family size can also raise entropy: more members have more chances to enter more semantic-shift regions.
- **Less regular semantic shift** → shift vectors spread across more regions → higher entropy (Shannon, 1948; Bonami and Paperno, 2018; Shen and Baayen, 2022).

- **Raw entropy** measures how broadly a family occupies semantic-shift regions.
- The **size-matched comparison** asks whether a family is more or less dispersed than random same-size type samples from the same domain.

Entropy measures how widely a family's semantic shifts spread across regions.

Higher H means the family's semantic shifts are less systematic and less consistent.

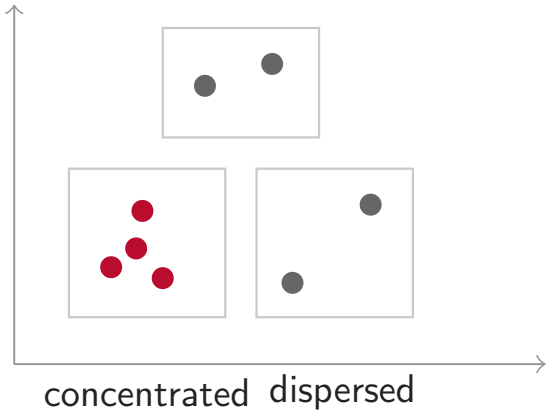
From shift vectors to regions

1. For each item, calculate a shift vector:

$$\Delta_i = \overrightarrow{\text{word}}_i - \overrightarrow{\text{stem}}_i$$
$$\Delta_i = \overrightarrow{\text{word}}_i - \overrightarrow{\text{other}}_i$$

2. Normalize the shift vectors and cluster them into semantic-shift regions.
3. For each family, count how many members fall in each region.

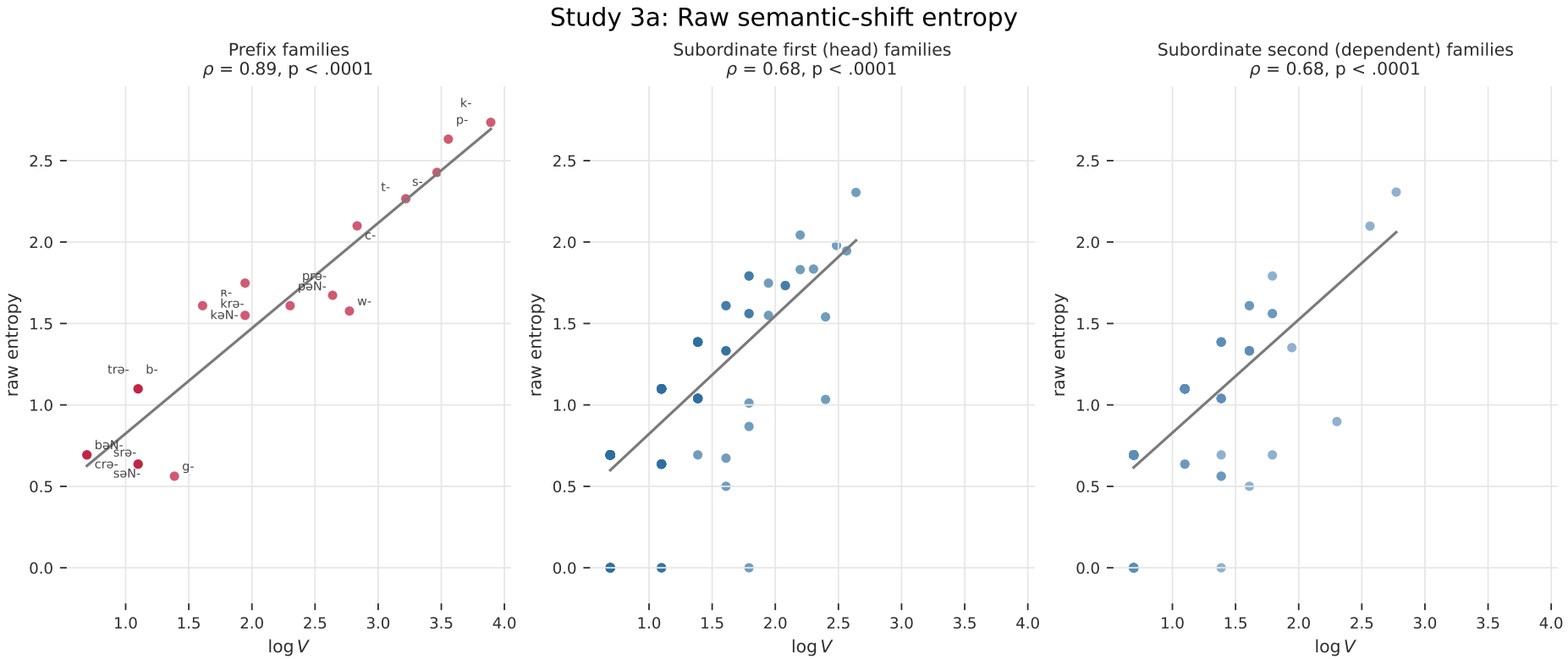
Family entropy (Shannon, 1948)



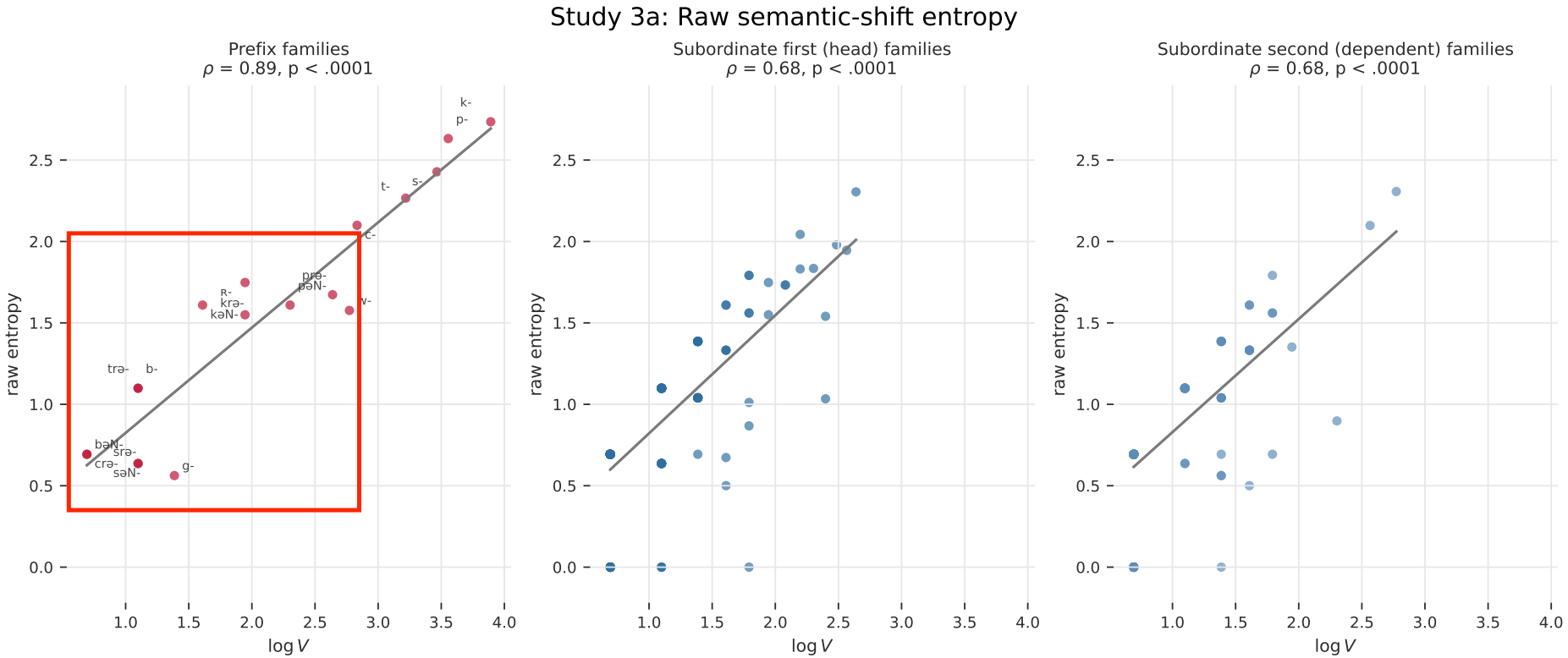
$$p_r = \frac{n_r}{V} \qquad H = -\sum_r p_r \log p_r$$

Each type contributes one count; for head families, $\overrightarrow{\text{other}}$ is the dependent component. Higher H means more occupied semantic-shift regions.

Study 3a Result: Raw Entropy



Study 3a Result: Raw Entropy



Raw entropy is size-sensitive: as prefix and subordinate component families get larger, they spread into more semantic-shift regions.

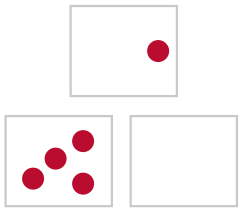
☞ Low raw entropy may simply reflect small family size, just as many head/dependent families sit below $\log V < 3.0$.

How We Calculate Size-Matched Entropy

A larger family has more chances to enter more semantic regions, so size must be controlled.

---

1. Observed

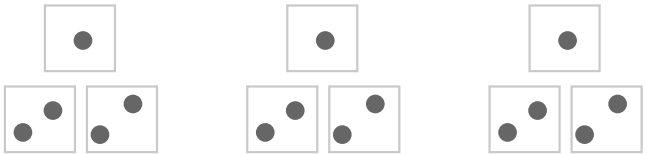


one real family
$V = 5$

Measure the real family's dispersion:

$$H_{\text{obs}}$$

2. Random baseline



random samples, same V

Draw random type samples with the same V from the same constructional domain.

This asks: how dispersed would this many types be by chance?

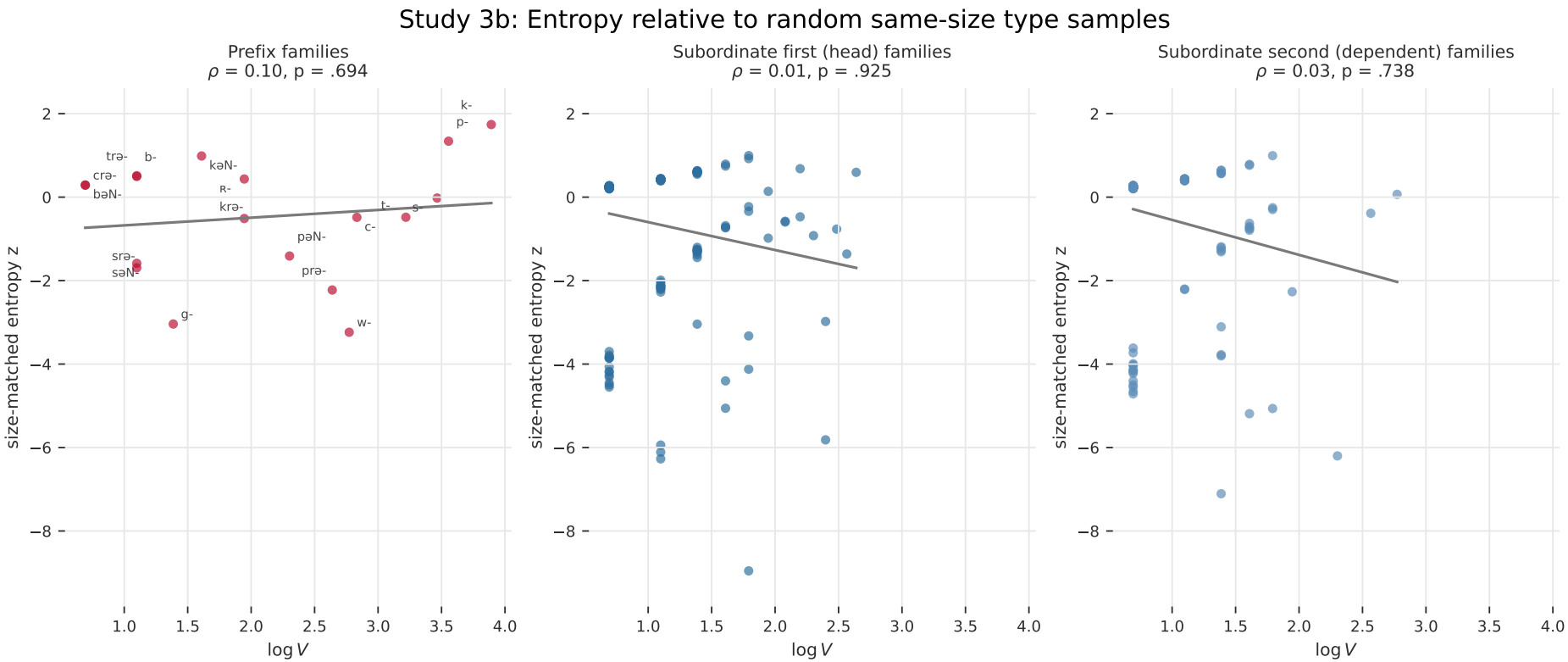
3. Read the result

- Lower: constrained for its size.
- Similar/higher: no size-controlled constraint.

Now large V is not enough: the family must beat its same-size baseline.

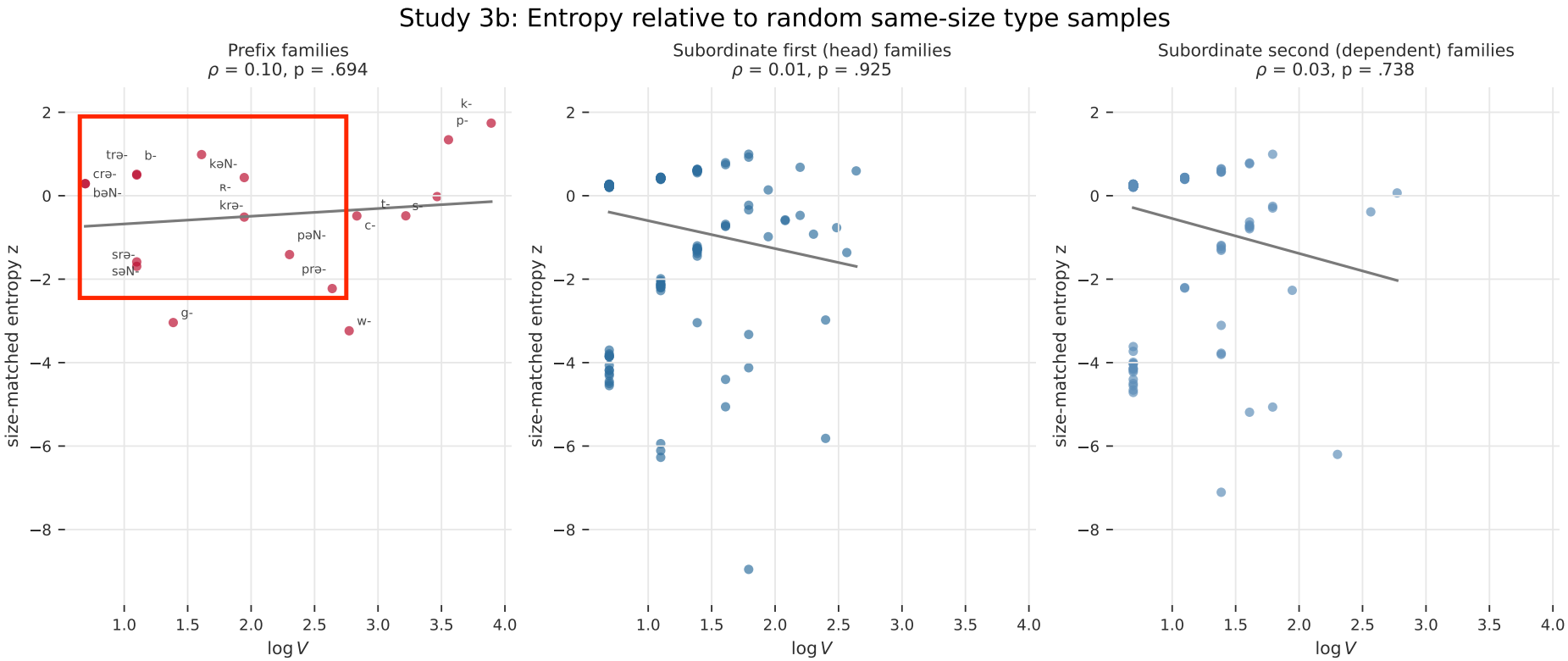
Each observed point gets its own same-V random baseline.

Study 3b Result: Size-Matched Entropy



Titles show rank correlations; gray lines are tendency from linear regression. $z < 0$: less dispersed than random same-size type samples; $z \approx 0$: no size-matched constraint.

Study 3b Result: Size-Matched Entropy



Titles show rank correlations; gray lines are tendency from linear regression. $z < 0$: less dispersed than random same-size type samples; $z \approx 0$: no size-matched constraint.

Size-Matched Summary

Size-matched entropy separates local prefix regularity from compound-like constraint.

- Some nasal and rhotacized prefix labels look regular in productivity and shift coherence.
- But in size-matched entropy, these prefix labels are not consistently low-dispersion for their size.
- Subordinate head and dependent families show a weak downward linear tendency after size matching; prefix families do not.

Interpretation: productive-looking prefix labels show residual organization, but they do not behave like fully compound-like active families.

Conclusion

Prefix-like and subordinate component families overlap in productivity and coherence, but differ in size-matched dispersion.

Study	Diagnostic	Prefix-like families	Subordinate component families
Study I	Hapax productivity	Larger families have lower P	Larger families have lower P
Study II	Coherent shifts	**Yes**: higher-P families show coherent stem-to-word shifts	**Yes**: higher-P families show coherent other-component-to-word shifts
Study III	Raw entropy vs. size	**Strong positive size effect**: larger families spread across more regions	**Strong positive size effect**: larger families spread across more regions
Study III	Size-matched entropy	**Not consistently low**: productive-looking labels can remain dispersed for their size	**Weak downward tendency**: larger families are more constrained

Take-home message: some nasal and rhotacized prefix labels preserve regularity, but size-matched dispersion keeps them distinct from compound-family productivity.

References

Alves, M. J. (2015a). Mon-khmer. In Lieber, R. and Štekauer, P., editors, *The Oxford Handbook of Derivational Morphology*, pages 520–544. Oxford University Press, Oxford.

Alves, M. J. (2015b). Morphological functions among Mon-Khmer languages. In Enfield, N. J. and Comrie, B., editors, *Languages of Mainland Southeast Asia: The State of the Art*, pages 524–550. De Gruyter Mouton, Berlin.

Alves, M. J. (2019). Morphology in Austroasiatic languages. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Baayen, R. H. (1992). Quantitative aspects of morphological productivity. *Yearbook of Morphology*, 1991:109–149.

Baayen, R. H. (2009). Corpus linguistics in morphology: Morphological productivity. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics: An International Handbook*, pages 900–919. Mouton de Gruyter, Berlin.

Baayen, R. H. and Lieber, R. (1991). Productivity and English derivation: A corpus-based study. *Linguistics*, 29(5):801–843.

Bauer, L. (2001). *Morphological Productivity*. Cambridge University Press, Cambridge.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bonami, O. and Paperno, D. (2018). Inflection vs. derivation in a distributional vector space. *Lingue e Linguaggio*, 17(2):173–195.

Butler, B. (2014). *Deconstructing the Southeast Asian sesquisyllable: A gestural account*. PhD thesis, Cornell University.

Butler, B. (2015). Approaching a phonological understanding of the sesquisyllable with phonetic evidence from Khmer and Bunong. *Studies in Asian Linguistics*, 8:217–242.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan. European Language Resources Association.

References

Haiman, J. (2013). Decorative morphology in Khmer. In Williams, J. P., editor, *The Aesthetics of Grammar: Sound and Meaning in the Languages of Mainland Southeast Asia*, pages 61–82. Cambridge University Press, Cambridge.

Jenny, M. and Sidwell, P., editors (2014). *The Handbook of Austroasiatic Languages*. Brill, Leiden.

Ourn, N. and Haiman, J. (2000). Symmetrical compounds in Khmer. *Studies in Language*, 24(3):483–514.

Ourn, N. and Haiman, J. (2003). Cambodian. In Thurgood, G. and LaPolla, R. J., editors, *The Sino-Tibetan Languages*, pages 395–411. Routledge, London.

Pittayaporn, P. (2015). Typologizing sesquisyllabicity: The role of structural analysis in the study of linguistic diversity in mainland southeast asia. In Enfield, N. J. and Comrie, B., editors, *Languages of Mainland Southeast Asia: The State of the Art*, pages 500–528. De Gruyter Mouton, Berlin.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Shen, T. and Baayen, R. H. (2022). Productivity and semantic transparency: An exploration of word formation in Mandarin Chinese. *The Mental Lexicon*, 17(3):458–479.

Stupak, I. V. and Baayen, R. H. (2022). An inquiry into the semantic transparency and productivity of German particle verbs and derivational affixation. *The Mental Lexicon*, 17(3):422–457.

Thank you

សូមអរគុណ

Questions and suggestions are very welcome.