The Morphological Handover: Prefixal Attrition and Compounding Growth in Khmer

Phenomenon. A central question in morphological theory concerns how productive derivational systems transition into lexicalized and fossilized forms (Alves, 2015; Gehrmann, 2018; Haiman, 2013). Recent quantitative approaches recast this transition not as a categorical loss of rules, but as a gradient decline in cue reliability, where morphological productivity reflects the extent to which a formal element supports novel formations (Chuang and Baayen, 2021; Stupak and Baayen, 2022). Within this view, morphological units function as weighted cues in a discriminative network rather than discrete rules. This perspective is particularly relevant for Mon-Khmer languages, which are typologically characterized by largely analytic and isolating morphological systems (Donegan and Stampe, 1983). Across the family, remnants of earlier productive morphology persist primarily as fossilized prefixes, often realized as unstressed minor syllables preceding stressed lexical bases. As a result, grammatical and derivational functions are increasingly expressed through lexical strategies and word order rather than productive affixation. In this study, we focus on Modern Khmer to examine this transitional morphological state through quantitative analyses of prefix productivity and semantic transparency.

Methodology. We analyzed a large-scale corpus of Modern Khmer to compare the morphological vitality of prefixal structures spanning different degrees of phonological reduction, including CrV-, CVN-, and monoconsonantal C-prefixes extracted from CCVC forms, alongside modern compounding strategies. Prefixes were manually annotated based on the phonotactic shape of the minor syllable, with etymological verification from *The Mon-Khmer Etymological Dictionary* (Shorto et al., 2006) to confirm their derivation from historical prefix–base constructions. Following Stupak and Baayen (2022), we computed potential productivity ($P = V(N, 1)/N$) to estimate the likelihood of novel formations within each category. Semantic transparency was quantified independently using fastText word embeddings (Bojanowski et al., 2017) as the cosine similarity between the semantic vectors of prefixed words and their lexical bases (the stressed main syllables).

Results. The results reveal a clear gradient in prefixal vitality across prefix shapes. Reduced CrV prefixes exhibit comparatively higher potential productivity at smaller vocabulary sizes and relatively higher semantic transparency (Figure 1), whereas CVN prefixes (e.g., *kaN*, *saN*) show near-zero potential productivity despite large token frequencies, indicating strong lexical entrenchment and fossilization (Baayen, 2009). When viewed jointly in productivity–type space (Figure 1, left), CVN and CrV forms pattern along a shared gradient rather than forming categorically distinct clusters, consistent with a continuum of declining prefixal productivity. Notably, some CVN prefixes (e.g., *taN*) retain slightly higher potential productivity, suggesting residual capacity for novel formations. Monoconsonantal C-prefixes extracted from CCVC compounds represent an extreme endpoint of this continuum. Although they occur with substantial token frequencies, these prefixes are associated with very small type inventories (*m-*: 2; *p-*: 19; *s-*: 9; *t-*: 6) and zero or near-zero hapax counts, yielding effectively zero potential productivity. This pattern indicates that C-prefixes persist only as lexically entrenched residues and no longer function as productive morphological cues, a conclusion supported by their uniformly low cosine similarity to lexical bases. Figure 2 situates these patterns within the broader space of compounding by reporting semantic transparency as cosine similarity between compounds and their first components. Compounds with complex initial shapes (CCVC, CrVN, CrVVN) show consistently high transparency, indicating stable and compositional contributions of the initial element. Many CCVC forms are historically derived from C–CVC structures but function synchronically as unified lexical units in compound formation.

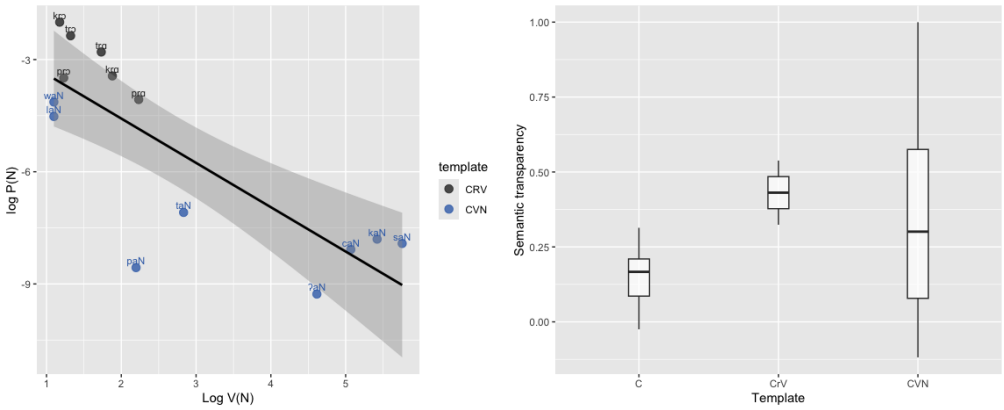

Figure 1: Potential productivity (left); semantic transparency between base (major syllable) and the whole word based on prefixes (right).

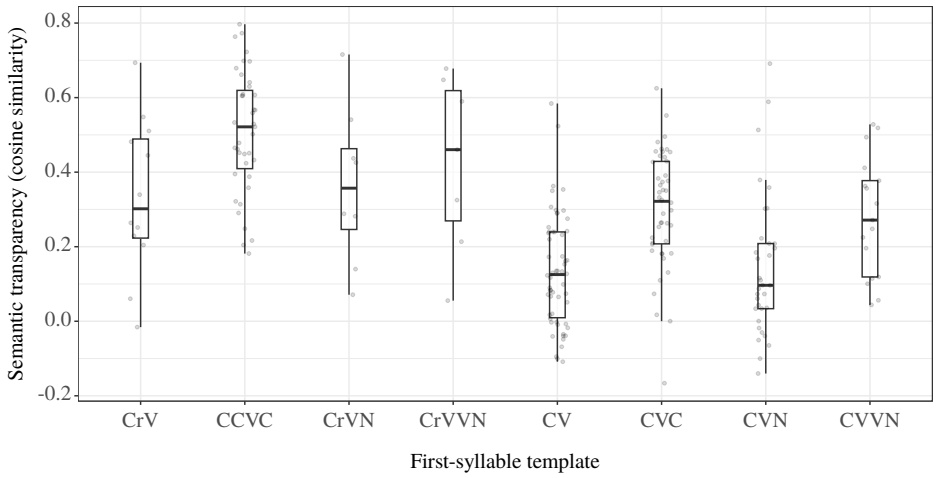

Figure 2: Semantic transparency scores by initial syllable template in compounds.

M. J. Alves. Morphological functions among Mon-Khmer languages: Beyond the basics. *Mainland Southeast Asian Languages: The State of the Art. Berlin, Mouton de Gruyter*, 531:557, 2015.

R. H. Baayen. Corpus linguistics in morphology: Morphological productivity. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics: An International Handbook*, volume 2, pages 899–919. Mouton de Gruyter, Berlin, 2009. 10.1515/9783110213881.2.899.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. ISSN 2307-387X.

Y.-Y. Chuang and R. H. Baayen. Discriminative learning and the lexicon: NDL and LDL. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press, 2021.

P. J. Donegan and D. Stampe. Rhythm and the holistic organization of language structure. In J. F. Richardson, M. Marks, and A. Chukerman, editors, *Papers from the Parasession on the Interplay of Phonology, Morphology, and Syntax*, pages 337–353. Chicago Linguistic Society, Chicago, 1983.

R. Gehrmann. Katuic presyllables and derivational morphology in diachronic perspective. In *JSEALS Special Publication No. 3, Papers from the Seventh International Conference on Austroasiatic Linguistics*, pages 132–56, 2018.

J. Haiman. Decorative morphology in Khmer. *The aesthetics of grammar: Sound and meaning in the languages of mainland Southeast Asia*, pages 61–82, 2013.

H. L. Shorto, P. Sidwell, D. L. Cooper, C. Bauer, et al. *A Mon-Khmer comparative dictionary*. The Australian National University, 2006.

I. V. Stupak and R. H. Baayen. How much semantics does grammatical number carry in Ukrainian: A distributional semantic approach. OSF Preprints, 2022.