Learning reduplicative templates as hidden structures: the case of reduplication-phonology interactions

Yang Wang
University of Utah
yangx.wang@utah.edu

1 Defining the problem

Models of morphophonological learning have focused primarily on concatenative processes, leaving the challenges of non-concatenative morphology largely unaddressed. Reduplication, the systematic copying operation (e.g., Ilokano pluralization [kal-kaldíŋ] 'goats'), is particularly revealing because successful learning requires the joint inference of prosodic templates that govern copying, underlying representations (URs) of stems and other affixes, as well as the phonological grammar.

Consider the schematic language in Table (1). The challenge of moving from the surface forms in Table (1a) to the analysis in Table (1b) lies in the mutual dependence between the grammar and the hidden structures. Solely based on the CAT paradigm, the learner cannot determine whether the reduplicative morpheme is a heavy syllable or a prosodic word, nor whether the stem UR is /bet/ with voicing assimilation or /bed/ with final devoicing.

Meanings	*Forms*
CAT_{STEM}	bet
CAT-PL	bed-a
CAT-DIM	bed-bet
DOG_{STEM}	panat
DOG-PL	panat-a
DOG-DIM	panat-nat

(a) Surface forms as input.

Morphemes	*Hidden units*
CAT_{STEM}	/bed/
DOG_{STEM}	/panat/
PL	/-a/
DIM	$\text{RED} = \sigma_{\mu\mu}$
Phonology: $\text{d} \rightarrow \text{t} \,/\, _\#$	

(b) The analysis.

Table 1: A schematic language with word-final devoicing and suffixing heavy-syllable reduplication.

The DOG paradigm provides the crucial disambiguating evidence. If the reduplicative template is the whole word, the predicted DOG-DIM form would be *[panat-panat], while a heavy-syllable template yields [panat-nat], which is what surfaces. The choice between /bed/ and /bet/ for CAT works analogously: positing /panat/ for DOG predicts [panat-a] for DOG-PL, which is observed; positing intervocalic voicing instead would wrongly predict *[panad-a]. Hence the CAT alternation must reflect devoicing of an underlying /bed/.

Within the typology of reduplication, Table (1) illustrates only one attested pattern of interaction between reduplication and segmental alternations, namely *normal application*, in which the process applies in all and only the expected environments. Other attested interactions include *overapplication* and *underapplication* (Wilbur, 1973); see McCarthy and Prince (1995) and Wang (2024) for plausible real language examples. Table 2 contrasts the three patterns: in *over*application (2a), devoicing applies to the first copy despite the absence of the conditioning environment. In *under*application (2b), devoicing fails to apply despite the presence of a word-final conditioning context.

Meanings	*Forms*
CAT_{STEM}	bet
CAT-PL	bed-a
CAT-DIM	bet-bet

(a) Lg2. *Overapplication*.

Meanings	*Forms*
CAT_{STEM}	bet
CAT-PL	bed-a
CAT-DIM	bed-bed

(b) Lg3. *Underapplication*.

Table 2: Typology of reduplication–phonology interactions. Only the CAT paradigm is shown; the DOG paradigm is identical to Table 1 in all three languages.

Prior computational work on reduplication has treated it as a morphology problem while setting aside its interactions with the rest of the phonological grammar (Vania and Lopez, 2017; Wilson, 2018; Dolatian and Heinz, 2018; Xu et al., 2020; Nelson et al., 2020; Beguš, 2021; Todd et al., 2022; Ellis et al., 2022). But the three attested patterns of reduplication–phonology interactions are precisely where copying and segmental phonology cannot be learned in isolation: each requires the learner to resolve hidden-structure ambiguities about templates

and URs jointly with the grammar that relates them. We present a learner that performs this joint inference, a combination not, to our knowledge, directly modeled in prior work. We show that it captures the attested typology.

Our goal is to model phonological acquisition: given a small paradigmatic input of the kind a child is likely to encounter, what hypotheses about underlying representations, templates, and grammar must the learner entertain in order to converge on the attested patterns? The EM-MaxEnt framework (O'Hara, 2017; Wang and Hayes, to appear) is well-suited to this question because (i) the constraints encode substantive linguistic hypotheses Base-Reduplicant Correspondence Theory (BRCT) (McCarthy and Prince, 1995) and Prosodic Morphology (McCarthy and Prince, 1986) whose learnability is itself the object of study; and (ii) the learned weights and posterior distributions over hidden structures are directly interpretable as theoretical claims, supporting analysis of what the learner does and why.

2 The model

2.1 The EM-MaxEnt framework

Wang and Hayes (to appear) propose a general framework for jointly learning URs and segmental phonology, in which URs are represented as latent categorical variables with multinomial distributions over candidates drawn from observed allomorph sets. This model, however, is limited to concatenative processes. To extend it to reduplication, we treat the reduplicative morpheme's prosodic template as a latent representation of the same formal type as a segmental UR: for each morpheme m the learner maintains a categorical prior θ_m over a finite set $\mathrm{UR}(m)$ of candidate representations. For a segmental morpheme, $\mathrm{UR}(m)$ is an allomorph set; for the reduplicative morpheme DIM, $\mathrm{UR}(\textsc{dim}) = \{\sigma_\mu, \sigma_{\mu\mu}, \mathrm{Ft}, \mathrm{PrWd}\}$. The same EM machinery learns both.

The probability of a surface form s given a meaning ω is the mixture

$$P(s \mid \omega; W, \theta) = \sum_{u \in \mathrm{UR}(\omega)} P(s \mid u; W) \prod_{m \in \omega} \theta_m(u_m), \tag{1}$$

where $u = (u_m)_{m \in \omega}$ ranges over joint UR assignments. The MaxEnt formula for each (UR, SR) pair is as in (2).

$$P(s \mid u; W) = \frac{exp\big(-\sum_k w_k\, C_k(u, s)\big)}{\sum_{s'} exp\big(-\sum_k w_k\, C_k(u, s')\big)} \tag{2}$$

Learning maximizes the regularized log conditional likelihood $\mathcal{L}(W, \theta) = \ln P(D \mid W, \theta) - \sum_i \frac{(w_i - \mu_i)^2}{2\sigma_i^2}$ via Expectation-Maximization (Dempster et al., 1977): the E-step computes $E(u, s, \omega) \propto P(s \mid u; W) \prod_m \theta_m(u_m)$; based on the computed expected counts, the M-step re-estimates W via L-BFGS-B and each θ_m by normalized expected counts. We use $2\sigma_i^2 = 10^5$ throughout, with non-negative weight constraints, and terminate when $\Delta\mathcal{L} < 10^{-3}$.

2.2 The reduplication learning component.

We augment the framework with three major components, the first of which is the *template space* as above.

Constraint set We use 12 standard constraints from BRCT (McCarthy and Prince, 1995) and related OT literature, supported by empirical evidence (Siah et al., under review); this includes input-base identity, identity between copies, anchoring, and contiguity constraints. We additionally propose RED=X, a single constraint in which X is bound at evaluation time by the template UR under consideration in the current (UR, SR) pair. It assigns a violation when the reduplicant in the surface form does not have the prosodic shape X specified by the input pair. For example, paired with the template UR $\sigma_{\mu\mu}$, a candidate reduplicant be violates RED=X, while bet does not. This diverges from standard BRCT, where each prosodic template corresponds to a separate constraint; the single weighted RED=X acts as a gatekeeper controlling the extent to which the posited template influences the reduplicant's shape.

GEN Candidate surface forms are constructed in three stages: ***reduplicant generation*** (each base segment may or may not surface, yielding 2^n candidates; e.g., for /bed/: {bed, be, ed, bd, b, e, d, ∅}), ***reduplicant placement*** (prefixing, suffixing, infixing, back-copying), and ***alternation substitution*** (each candidate is expanded with segmental alternants from the regular phonology; e.g. if [bed-bed] is in the candidate set, a [t]∼[d] alternation would also produce [bed-bet], [bet-bed] and [bet-bet]). Although each toy language contains only six

observed surface forms, the learner evaluates 1,992 hidden-structure/surface-form candidate pairs.

3 Evaluation and results

We evaluate the learner on the three schematic languages in Tables 1 and 2. The hypothesis space is identical across runs, same four candidate templates, same UR candidates from observed allomorphs, and the same 13 constraints. Only the CAT-DIM form differs, and this difference drives the learner to different grammars.

At initialization, all θ_m are uniform and constraint weights are zero. Across all three languages, the learner recovers the alternating-stem UR /bed/ and the template $\sigma_{\mu\mu}$ with posterior probability $> .999$. For Lg1 and Lg2, the learned grammar assigns probability $> .999$ to the observed outputs. For Lg3, the data are consistent only with a variable grammar, and the learner converges on a 0.50/0.50 distribution that frequency-matches the input. Table 3 presents the learned weights of key constraints across the three languages.

Constraint	Lg1 *normal*	Lg2 *over*	Lg3 *under*
*FINALVOICEDOBS	17.32	22.19	**8.56**
IDENT-IO[VOICE]	9.35	13.38	**8.56**
*VTV	1.34	6.99	1.16
*MARGINCLUSTER	6.28	3.53	6.22
MAX-IO	11.21	7.7	11.20
IDENT-BR[VOICE]	0.05	21.8	15.92
MAX-BR	6.66	0	6.70
CONTIG-BR	10.45	9.32	10.16
L-ANCHOR-BR	0.39	0.00	0.00
R-ANCHOR-BR	9.02	21.64	8.62
ALIGNREDL	0.00	0.00	0.00
ALIGNREDR	5.10	7.69	5.16
RED=X	21.44	7.34	21.18

Table 3: Learned constraint weights for the three languages. In Lg3, *FINALVOICEDOBS and IDENT-IO[VOICE] converge to equal weights, yielding free variation.

In Lg1 (*normal* application), the training datum [bed-bet] already exhibits different voicing across copies, so the learner has no pressure to push IDENT-BR[VOICE] up; it stays near zero ($w = 0.05$). Devoicing applies independently to each copy via *FINALVOICEDOBS outweighing IDENT-IO.

In Lg2 (*over*application), [bet-bet] forces the learner to explain why /bed/ surfaces as [bet] in the left copy. Devoicing in the left copy has no markedness motivation; the only available explanation in the hypothesis space is that BR identity pulls the base toward the markedness-driven [t] of the reduplicant. EM accordingly raises IDENT-BR to $w = 21.8$, with *FINALVOICEDOBS at $w = 22.19$ providing markedness pressure on the second copy.

In Lg3 (*under*application), the training data are mutually incompatible under any categorical grammar in the hypothesis space. The bare stem [bet] requires *FINALVOICEDOBS to outweigh IDENT-IO; the reduplicated form [bed-bed] requires the opposite. The learner settles to *frequency match*: *FINALVOICEDOBS and IDENT-IO converge to equal weights ($w = 8.56$ each), with IDENT-BR high ($w = 15.92$) to enforce identity between copies. The resulting grammar predicts 0.50/0.50 free variation between [bed]∼[bet] for CAT_{STEM} and [bed-bed]∼[bet-bet] for CAT-DIM, as in Table 4.

	*FINVCDOBS 8.56	ID-IO 8.56	ID-BR 15.92	H	P
Input: /bed/					
[bed]	1			8.56	.500
[bet]		1		8.56	.500
*Input: /bed-*RED/					
[bed-bet]			1	15.92	.000
[bed-bed]	1			8.56	.500
[bet-bet]		1		8.56	.500
[bet-bed]	1	1	1	33.04	.000

Table 4: Underapplication grammar: equal weights on *FINVCDOBS (abbreviated) and ID-IO yield free variation. Candidates with BR mismatches lose categorically.

This recapitulates the classical asymmetry of McCarthy and Prince (1995, p5) as an emergent property of the hypothesis space. Under BRCT, categorical underapplication requires an independently motivated constraint to break the tie. Without such a constraint, no grammar in the space prefers [bed-bed] categorically over [bet-bet], and the learner converges on the free-variation grammar. Although it remains unclear how humans acquire underapplication, some documented underapplication processes do seem to involve free variation (see Tagalog in Zuraw, 2002).

4 Discussion

As a first step, the model demonstrates that EM-MaxEnt inference over multiple interacting hidden structures, both prosodic templates and segmental URs, is feasible and effective for learning

reduplication-phonology interactions. While the underlying optimizer is the EM-MaxEnt framework of Wang and Hayes (to appear), the contribution here is in the *hypothesis space*: representing prosodic templates as URs of the reduplicative morpheme, introducing RED=X as the mechanism by which templates exert grammatical influence, and constructing GEN to generate reduplication candidates parametrically.

The asymmetry between overapplication and underapplication arises from this hypothesis-space structure rather than by stipulation. Overapplication is acquired categorically when the data leave only BR-identity as an available explanation; underapplication, lacking an independently motivated tie-breaker, drives the learner to free variation.

The 1,992 candidate pairs here arise from the combinatorial structure of GEN. This grows with morpheme length and the alternation set rather than with lexicon size; adding more stems sharing the same affix without introducing new alternations gives the learner stronger evidence about that affix's UR rather than enlarging the search. Future work will evaluate the model on natural-language reduplication systems and compare it to alternative frameworks for non-concatenative morphology such as Generalized Non-linear Affixation (Bermúdez-Otero, 2012). More broadly, the same EM-MaxEnt machinery extends from segmental URs to prosodic templates, suggesting a unified route to learning phonological hidden structures.

References

Gašper Beguš. 2021. Identity-based patterns in deep convolutional networks: Generative adversarial phonology and reduplication. *TACL*, 9:1180–1196.

Ricardo Bermúdez-Otero. 2012. The architecture of grammar and the division of labour in exponence. In Jochen Trommer, editor, *The Morphology and Phonology of Exponence*, pages 8–83. Oxford University Press, Oxford.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Hossep Dolatian and Jeffrey Heinz. 2018. Learning reduplication with 2-way finite-state transducers. In *Proceedings of the 14th International Conference on Grammatical Inference*, volume 93 of *Proceedings of Machine Learning Research*, pages 67–80. PMLR.

Kevin Ellis, Adam Albright, Armando Solar-Lezama, Joshua B Tenenbaum, and Timothy J O'Donnell. 2022. Synthesizing theories of human language with bayesian program induction. *Nature communications*, 13(1):5024.

John J. McCarthy and Alan S. Prince. 1986. Prosodic morphology. Ms., University of Massachusetts, Amherst, and Brandeis University, Waltham, Mass.

John J. McCarthy and Alan S. Prince. 1995. Faithfulness and reduplicative identity. In *Papers in Optimality Theory*, Amherst, MA. Graduate Linguistic Student Association, Dept. of Linguistics, University of Massachusetts.

Max Nelson, Hossep Dolatian, Jonathan Rawski, and Brandon Prickett. 2020. Probing RNN encoder-decoder generalization of subregular functions using reduplication. *Society for Computation in Linguistics*, 3(1).

Charlie O'Hara. 2017. How abstract is more abstract? Learning abstract underlying representations. *Phonology*, 34(2):325–345.

Jian Leat Siah, Sam Zukoff, and F. F. Hsieh. under review. Resolving reduplicative opacity in Malay nasal spreading: Argument for Base-Reduplicant correspondence theory. *Language*.

Simon Todd, Annie Huang, Jeremy Needle, Jennifer Hay, and Jeanette King. 2022. Unsupervised morphological segmentation in a language with reduplication. In *Proceedings of the 19th SIGMORPHON Workshop*, pages 12–22, Seattle, Washington. Association for Computational Linguistics.

Clara Vania and Adam Lopez. 2017. From characters to words to in between: Do we capture morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2016–2027, Vancouver, Canada. Association for Computational Linguistics.

Yang Wang. 2024. *Studies in morphophonological copying: Analysis, experimentation and modeling*. Ph.D. thesis, University of California, Los Angeles.

Yang Wang and Bruce Hayes. to appear. Learning phonological underlying representations: the role of abstractness. *Linguistic Inquiry*.

Ronnie Bring Wilbur. 1973. *The phonology of reduplication*. University of Illinois at Urbana-Champaign.

Colin Wilson. 2018. Modeling morphological affixation with interpretable recurrent networks: sequential rebinding controlled by hierarchical attention. In *CogSci*.

Hongzhi Xu, Jordan Kodner, Mitch Marcus, and Charles Yang. 2020. Modeling morphological typology for unsupervised learning of language morphology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6672–6681.

Kie Zuraw. 2002. Aggressive reduplication. *Phonology*, 19(3):395–439.