What representations do RNNs learn and use from morpho-phonological processes?: an exploration of PCA and PC neutralizations on Turkish vowel harmony

Jane Li, Kyle Rawlins, Paul Smolensky. Johns Hopkins University.
{sli213, kgr, paul.smolensky}@jhu.edu

Phonologists and cognitive scientists have long debated about the types of generalizations that humans infer from the observations of the applications of morpho-phonological processes [1-3]. For instance, upon observing the following alternations, one can form two possible hypotheses.

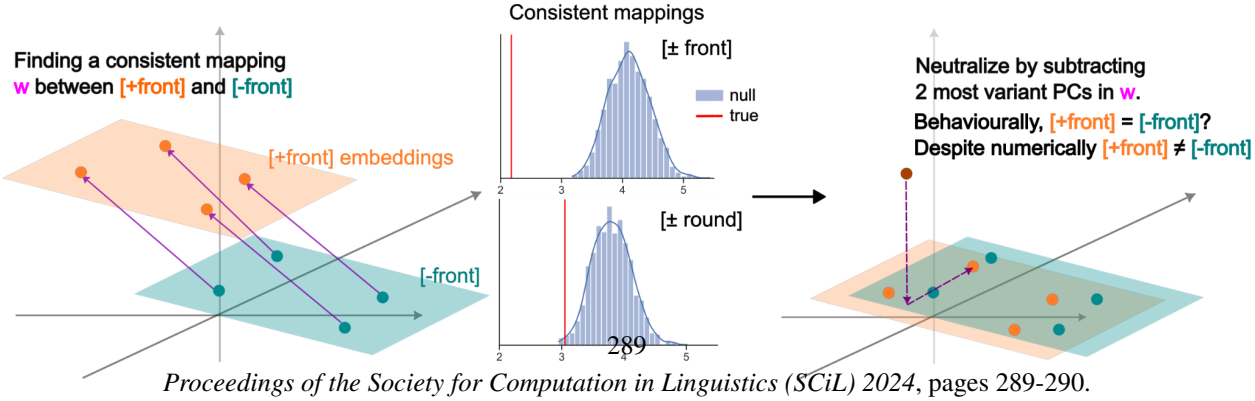
Observe (data from [1])		**Segmental generalization**		**Sub-segmental generalization**
piki-mi	tudu-mu	V → i / iC_	V → u / uC_	V → [α back] / [α back] C_
gæmæ-mi	dapa-mu	V → i / æC_	V → u / aC_	
tipæ-mi	napu-mu			

These hypotheses about *human* knowledge have been explored extensively via artificial language learning experiments [1-2] and infant learning experiments [3], but little is known about what is inferred by *machines* that observe morpho-phonology. Previous work has probed at the generalization abilities of recurrent neural networks (RNNs) by observing how they inflect novel stems or comparing production probabilities to human ratings of such inflections [4-5]. Success on these measures demonstrate that RNNs can learn to appropriately inflect, but do not differentiate between the types of generalizations. Symbolic models of morpho-phonological learning that build in sub-segmental information and generalization algorithms have also shown success in predicting human behavior on inflection tasks [6-7]. **Do RNNs implicitly learn and utilize the sub-segmental information necessary to succeed in an inflection task?**

In this work, we explore the hypothesis that the contrasting types of generalizations can be distinguished through analyses of the learned phoneme embeddings. This was tested on RNNs performing a noun inflection task in Turkish, which involved the learning of Turkish complex vowel harmony. Descriptively, vowels harmonize in [± front] and [± round] and neutralize to the [α front, α round, +closed] vowel. After observing the full range of vowel harmony alternations present in Turkish noun inflection, two contrasting generalizations can be formed:

Segmental generalization		**Sub-segmental generalization**
V → i / {i, e}C_	V → ɯ / {ɯ, a}C_	[+closed] → [α front, α round] /
V → y / {y, œ}C_	V → u / {u, o}C_	[α front, α round] C_

We conjecture that, if a sub-segmental feature is represented in an RNN, we expect a consistent mapping between principal component (PC) projections of minimal pairs that differ by that feature. In this case study, we expect consistent mappings of [± front] minimal pairs {i/ɯ, e/a, y/u, œ/o} and [± round] minimal pairs {i/y, e/œ, ɯ/u, a/o}. To test whether these distinctions are *used* in the system, we neutralize by subtracting the strongly contributing PCs from the [+front] (or [+round]) phoneme. If the feature is used in the RNN, then we expect that test stems with final [+front] vowels have harmony behavior consistent with their [-front] counterparts.



Proceedings of the Society for Computation in Linguistics (SCiL) 2024, pages 289-290.
Irvine, California, June 27-29, 2024

Model. 100,461 inflected nouns that obey the Turkish harmony system were selected from the UniMorph-Turkish dataset [8]. The inputs to the RNN are strings that start with three syntactic tokens and a stem, e.g., `<PL> <PSS1P> <NOM> g r u p`. The RNN learns to output the inflected form of the stem combined with its syntactic specifications, e.g., `g r u p l a r ɯ m ɯ z`. Overall performance was at 97.0%, with minimal differences between syntactic categories (range: [96.7%, 97.2%], 52 categories, 11,000 test items). Phoneme embeddings (d = 100) were shared across encoder and decoder. All following reported results are an average of five models.

All phoneme embeddings were submitted to a principal components analysis with all PCs (n = 56) extracted, such that we have a covariance matrix M and a projection $\mathbf{v}$ for each phoneme embedding $\mathbf{u}$ ($\mathbf{u} = M\mathbf{v}$). First, we inspect whether there exists a consistent mapping between [± front] minimal pairs (and likewise, for [± round]), restricting our search of mappings to the addition of a constant vector, i.e., for any [± front] minimal pair embeddings $\mathbf{v}_{[+f]}$ and $\mathbf{v}_{[-f]}$, $\mathbf{v}_{[+f]} \approx \mathbf{w} + \mathbf{v}_{[-f]}$. This is akin to the visual inspection of analogical relationships in word embeddings in PC space, e.g., *king* − *man* ≈ *queen* − *woman* [9]. Statistically, we consider a mapping to be consistent if the average pairwise Euclidean distance between pairs (e.g., 4 minimal pairs → 6 pairwise distances) is rejected from a null, randomly-permuted, distribution. For both [± front] and [± round], we found a consistent mapping between their minimal pairs (pictured above).

We then ask whether numerically neutralizing some feature-representing PCs can behaviorally render the same effect as replacing the vowel with its minimal counterpart. A PC is considered to be "feature-representing" if the difference vector $\mathbf{w}$ has accounted for a significant amount of variance within the vowel embeddings on that PC dimension, and we "neutralize" a PC p by subtracting w_p from the [+front] or [+round] phoneme projection (illustrated above as the mapping of the brown point to the orange). We neutralize the top two feature-representing PCs, generating a new embedding set each for [± front] and [± round]. If these the distinctions encoded in these feature-representing PCs are subsequently utilized in the RNN, we should expect that a model with the respective neutralized embeddings to be functionally equivalent to a model tested with altered inputs, where the last stem vowel is replaced with its [-front] or [-round] counterpart. For example, the numerical neutralization of the front vowel [y] to its back counterpart [u] should, by hypothesis, inflect the stem [dyʃ] 'dream' as [dyʃ**u**] instead of [dyʃ**y**].

We tested the model with the neutralized embeddings on the original test set and compared the outputs against (i) the original model with the same test set and (ii) the original model with an altered test set (last vowel neutralized). Strikingly, we found that the neutralized model had **no difference** in output to (i) – the neutralized embeddings across five runs rendered the same outputs. This is unexpected because we anticipated at least some perturbation of the outputs. At first blush, this suggests that whatever distinction is encoded in latent embedding space ends up being overlooked or outweighed by other information in the stem. This is partially supported by the fact that in the outputs of (ii), we anecdotally notice (and plan to quantify) that altered items tend to follow the vowel harmony pattern that was consistent with their original stems, suggesting the importance of other information present in the stem (e.g., tendency for harmonic stems). However, these results do not necessarily indicate that there is no symbolic-like manipulation within RNNs – it could be that symbolic manipulations occur post-embedding, e.g., after the encoding of the stem, in which case our embedding neutralizations techniques do not end up taking effect after multiple non-linear transformations.

References. **[1]** Finley & Badecker (2009). *JML*. **[2]** Peperkamp & Dupoux (2007). *LabPhon*. **[3]** Zamuner, Kerkhoff, & Fikkert (2012). *Appl. Psycholing.* **[4]** Kirov & Cotterell (2018). *TACL*. **[5]** Corkery et al. (2019). *ACL*. **[6]** Albright & Hayes (2002). *ACL*. **[7]** Hayes & Wilson (2008). *Linguistic Inquiry.* **[8]** McCarthy et al. (2019). *ACL*. **[9]** Mikolov et al. (2013). *NAACL-HLT.*