Probing Subphonemes in Morphology Models

Gal Astrach **Yuval Pinter**
Department of Computer Science and Data Science Research Center
Ben-Gurion University of the Negev
Beer Sheva, Israel
{galastra@post,uvp@cs}.bgu.ac.il

Abstract

Transformers have achieved state-of-the-art performance in morphological inflection tasks, yet their ability to generalize across languages and morphological rules remains limited. One possible explanation for this behavior can be the degree to which these models are able to capture implicit phenomena at the phonological and subphonemic levels. We introduce a language-agnostic probing method to investigate phonological feature encoding in transformers trained directly on phonemes, and perform it across seven morphologically diverse languages. We show that phonological features which are local, such as final-obstruent devoicing in Turkish, are captured well in phoneme embeddings, whereas long-distance dependencies like vowel harmony are better represented in the transformer's encoder. Finally, we discuss how these findings inform empirical strategies for training morphological models, particularly regarding the role of subphonemic feature acquisition.

1 Introduction

The transformer architecture has revolutionized natural language processing and computational linguistics since its introduction by Vaswani et al. (2017). While it has achieved state-of-the-art results across various tasks, much remains to be understood about its inner workings and representations. Investigating how transformers acquire and use linguistic knowledge is crucial for assessing their ability to generalize beyond shallow pattern recognition. One such aspect is morphological knowledge, such as that examined in tasks like *morphological inflection* (Cotterell et al., 2017), where a model predicts a word's inflected form given a lemma and morphosyntactic attributes. For example, for the English lemma "hug" and the morphosyntactic attributes `VERB;PAST`, the model should output "hugged". In many languages, morphology interacts in meaningful ways with *phonological* attributes, for example through the phenomenon of *harmony*, raising the question of how this correspondence is manifested in model representations. This question is hard to pursue due to the general scarcity of multilingual morphological data, which hinders models' ability to generalize to new lemmas and morphosyntactic attributes (Goldman et al., 2022; Kodner et al., 2023b) and to adapt to the diversity of morphological processes (Kodner et al., 2022). In this work, we present a language-agnostic method for testing phonological features and long-context feature agreement in models trained on a morphological task.[1] We do so by training designated probing classifiers that predict linguistic properties from a model's internal representations (Belinkov, 2022). We show how a morphological transformer implicitly acquires phonological knowledge, complementing previous findings regarding the representations found in neural phoneme embeddings (Rodd, 1997; Silfverberg et al., 2021; Muradoglu and Hulden, 2023; Mirea and Bicknell, 2019; Silfverberg et al., 2018; Kolachina and Magyar, 2019; Steuer et al., 2023) and the information conveyed by morphological models (Muradoglu and Hulden, 2023; Kodner et al., 2023a; Gorman et al., 2019). Unlike previous work, we demonstrate via explainability methods that model representations explicitly encode phonological features, and quantify how well they are encoded across languages and features. We test the hypothesis that when trained on reliable phonological representations, models acquire subphonemic features such as `VOICE` or `ROUND` (Chomsky and Halle, 1968) that play a role in morphology, and that this ability depends on a language's reliance on such features in encoding inflectional properties. We find that local phenomena, such as final consonant devoicing, are captured in the character embeddings, while long-distance phenomena are better represented in contextualized embeddings

[1] https://github.com/MeLeLBGU/probing-subphonemes

Findings of the Association for Computational Linguistics: ACL 2025, pages 12954–12961

from the transformer encoder. Finally, we argue for current practices in language transfer of morphological models based on our results.

2 Probing Phonological Features

Our experiments consist of three stages: training a phonemic transformer model on a morphological task for a language; probing the embeddings for phonological features; and analyzing the probe using minimum description length (MDL).

2.1 Phoneme-based Transformer

We use a character-based encoder-decoder transformer which achieves state-of-the-art results on morphological inflection (Wu et al., 2021). This architecture is relatively small and employs a feature-invariant positional encoding for morphosyntactic tags, making their order irrelevant to the model. We modify the architecture by weight tying, using the same embedding table for both the encoder and decoder during training and evaluation (Press and Wolf, 2017). We train the transformer on the SIGMORPHON 2017 shared task dataset (Cotterell et al., 2017) which covers multiple languages with diverse typologies.[2] To directly analyze the representation of phonological features, we transcribe the lemmas from standard orthographic form to International Phonetic Alphabet (IPA) using Epitran (Mortensen et al., 2018),[3] a rule-based grapheme-to-phoneme tool.[4] We refer to IPA characters as *phonemes* interchangeably.

We train two versions of the transformer: (i) an **inflection model**, trained on the phonemic transcriptions of the morphological inflection task; and (ii) a **lemma copying model**, where we replace each morphosyntactic attribute with `COPY` and set the inflected form as identical to the lemma. We then probe the phoneme embeddings and the encoder using a set of probe tasks.

2.2 Probe Tasks

We design two types of probes to evaluate how well phonological features are embedded by a model. The **phoneme probe** assesses the phoneme embeddings, while the **harmony probe** evaluates the encoder's output vectors. Separate probes are trained for each (phonological feature, language) pair.

[2]For Hebrew (vocalized), we use the dataset from Kodner et al. (2022).

[3]Version 1.25.1.

[4]For Hebrew (voc), we use a dedicated API (Cohen, 2019).

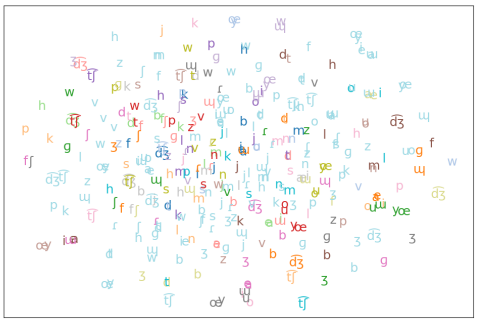

Figure 1: t-SNE projection of phoneme embeddings after training on Turkish morphological inflection. Characters from each seed are presented in a distinct color.

Extracting phonological features. We use PanPhon (Mortensen et al., 2016)[5] to map each phoneme to its corresponding phonological features, each represented as a ternary value: $+$, $-$, or 0 (meaning "irrelevant"), demonstrated in Table 1.

Phoneme probe. For each phonological feature, we train a probe using the phoneme embeddings as input and the feature values as labels. However, the limited number of phonemes per language makes it insufficient for training a probe and may leave some feature values out of distribution. To mitigate this, we augment the dataset by training the transformer with multiple random seeds, generating a diverse set of phoneme embeddings. Due to computational constraints, we additionally apply oversampling by a factor of three. To see whether embeddings from different seeds exhibit inherent structure, we project them into a 2D plane (Figure 1) using t-SNE (Van der Maaten and Hinton, 2008). The lack of clustering among identical phonemes suggests that this data augmentation strategy effectively diversifies the embeddings.

Harmony probe. To investigate how well the transformer encodes long-distance phonological dependencies, we design a probe that mimics vowel and consonant harmony rules. We generate nonce words using the method and code from Muradoglu and Hulden (2023, §3). The probe's inputs are the contextualized phoneme vectors produced from the encoder for these nonce words, taken from the last layer of the encoder when passed on nonce words, with beginning-of-sequence and end-of-sequence tokens attached. The probe classifies the harmony

[5]Version 0.21.1.

	syl	son	cons	cont	delrel	lat	nas	strid	voi	sg	cg	ant	cor	distr	lab	hi	lo	back	round	tense	long
k (phoneme)	−	−	+	−	−	−	−	0	−	−	−	−	−	0	−	+	−	+	−	0	−
köpek (v. harmony)	+	+	−	+	−	−	−	0	+	−	−	0	−	0	−	−	−	−	0	0	−
köpek (c. harmony)	−	−	+	−	−	−	−	0	−	−	−	0	−	0	0	0	−	0	−	0	−

Table 1: Phonology features extracted via Panphon for the probes: a single phoneme and (vowel / consonant) harmony type for the word köpek (dog in Turkish).

type of each word for both vowel and consonant harmony: + if all phonemes are + or 0, − if all are − or 0, and 0 if the word is disharmonic, containing both + and − values, demonstrated in Table 1. We train separate probes for vowel and consonant harmony for each phonological feature, but only if at least two phonemes in the language exhibit + and − values each for that feature.

2.3 MDL Probes

Traditional probing methods use metrics like accuracy or F1 score to estimate how well embeddings encode linguistic properties. However, this approach has several limitations. A probe may perform well even with randomly assigned labels or when applied to randomly-initialized representations. To address this, we adopt an information-theoretic approach (Voita and Titov, 2020) and report a metric based on the probe's minimum description length (MDL) instead. This method accounts for the probe's complexity, making it more robust and comparable across different models and linguistic properties. For each phonological feature, we compute MDL using the online coding approach: We segment the probe dataset at sequential indices $t_0 < t_1 < \cdots < t_S = n$, where n is the size of the probe's dataset. A probe θ_i is trained on each prefix of the data while measuring cross-entropy loss on the next segment.[6] Summing these losses yields the total description length of the feature:

$$L = t_0 \log_2 K - \sum_{i=0}^{S-1} \log_2 p_{\theta_i}(y_{t_i+1:t_{i+1}} | x_{t_i+1:t_{i+1}}), \quad (1)$$

where K is the number of classes (in our case, 3).

To normalize across datasets of different sizes, we compute a *compression score* by dividing the uniform coding length by the MDL, providing a comparable measure of how efficiently phonological information is encoded across different languages with varying phoneme inventories:

$$\mathcal{C} = \frac{n \log_2 K}{L}. \quad (2)$$

Analyze a probe using compression score characterizes the strength of regularity in the embeddings with respect to the labels. While it is a strictly relative metric, higher values indicate stronger regularity and therefore that the labels are better encoded in the embeddings.

3 Results

We apply our method to a set of seven languages selected to provide reasonable diversity with respect to morphophonological phenomena.[7] These languages represent different morphological typologies: agglutinative / fusional, prefixing / suffixing / non-concatenative; and exhibit diverse phonological rules, such as palatalization and vowel harmony. We focus on languages with low *orthographic depth*, which allows us to probe phoneme embeddings in a way comparable to probing character embeddings in standard orthography.

We first compare the results to a control probing task where the phonological features (labels) are randomly shuffled (Hewitt and Liang, 2019). The results, presented in Figure 2, validate that the compression score is a good indicator of phonological feature representation in the embeddings, as all scores remain below 1.0. We next discuss specific morphophonological phenomena and how they manifest in the data.

Turkish final-obstruent devoicing. In Turkish, a word-final [−CONTINUANT] consonant is devoiced. Moreover, in accusative case the consonant becomes voiced, and a [+HIGH, −ROUNDED] vowel is suffixed, with the BACK value subject to vowel harmony. For example, *kebap* (the kebab) becomes *kebabı* (ACC the kebab). Our probe shows that both VOICE and CONTINUANT features have a relatively good compression score in both probes, with more prominence compared to other features in the inflection task.

[6] To address class imbalance, we weigh the loss by the inverse frequency of each feature. The probe's architecture follows Voita and Titov (2020) and is implemented as a multi-layer perceptron with two hidden layers of 100 neurons each.

[7] Feature inclusion score (Ploeger et al., 2024) of 0.63.

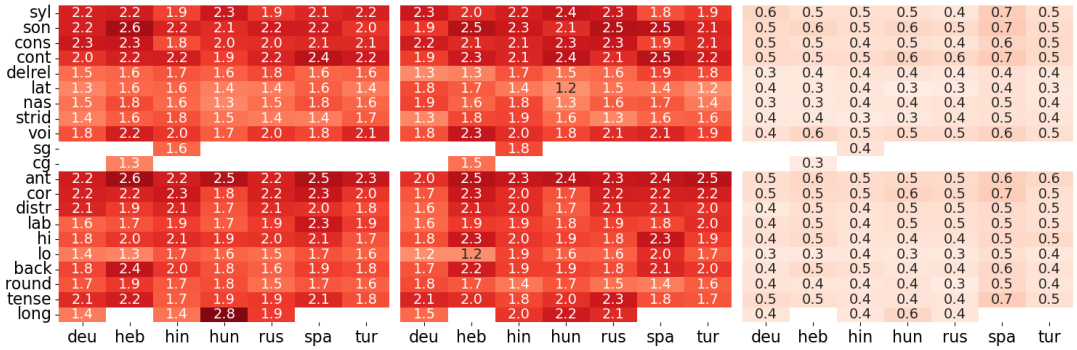

Figure 2: Compression scores ($\mathcal{C}$) of phoneme embeddings: phonological features are plotted on the y-axis (abbreviated), and languages are on the x-axis (represented by their ISO 639-3 code). From left to right: inflection model, lemma copying, and control task.

Hungarian gemination. Nearly every phoneme in Hungarian has a corresponding [+LONG] variant of it in the phonetic inventory. There are two processes that can alter the value of the LONG feature: gemination, where consonants at the end of a verb become [+LONG] before a suffix, and degemination, where [+LONG] consonants become [−LONG] when preceded or followed by another consonant. Among all languages and features in the phoneme probe, the compression score for LONG in Hungarian is the highest in the inflection model. We hypothesize that this is due to two factors: (i) morphological alternations affecting the gemination process, and (ii) the high entropy of LONG, which effectively separates Hungarian's phonetic inventory, allowing the probe to achieve a relatively high compression score.

Long-context feature agreement. Vowel harmony is a rule requiring all vowels in a word to share a specific phonological feature. For example, Turkish and Hungarian exhibit vowel harmony for ROUND and BACK. Since this rule influences both phonotactics and morphology, we expect these features to have high compression scores. While this is not observed in the phoneme probe (Figure 2), the harmony probe results (Figure 3) show high compression scores for context-dependent embeddings in the inflection model. Results for probing consonant harmony are provided in Appendix A.

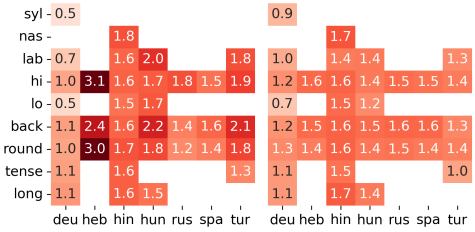

Figure 3: Compression scores ($\mathcal{C}$) for probing vowel harmony. Inflection model on the left, lemma copying on the right.

4 Discussion

We showed that a morphological transformer can effectively acquire phonological features. The quality of their representation, as reflected by the compression score ($\mathcal{C}$), varies across features and languages, and is influenced by how informative they are per language. Features prominent in phonotactics or in short-context environments are represented better in the phoneme embeddings, while those more present across long contexts are represented better through the encoder. Higher scores are generally observed for features that are more central to morphology and phonology, though these results may also be influenced by the quality of the datasets or grapheme-to-phoneme tools. Surprisingly, in the phoneme probe, the lemma copying model achieves on par with or even better than the inflection model. We believe this might be due to dataset noise, as explored in prior work (Wiemerslage et al., 2023). Future work could investigate the variance across languages and models.

Our findings complement work that showed that adding subphonemic features hardly improves model performance, suggesting these are already present in their representations (Wiemerslage et al., 2018; Guriel et al., 2023). Our lemma copying findings reinforce the common practice of pretraining models for this task before turning to inflection (Yang et al., 2022; Liu and Hulden, 2022; Anastasopoulos and Neubig, 2019), which has been argued to succeed due to inducing "copy bias"

and to coaxing attention modules towards monotonicity (Aharoni and Goldberg, 2017). Finally, our results imply that the demonstrated success of transfer learning in morphological inflection, even between typologically unrelated languages (McCarthy et al., 2019; Elsner, 2021), might stem from the model's ability to acquire subphonemic features, which are approximately universal (Mielke, 2008) and therefore transferable.

5 Conclusion

In this paper we analyze phonological transformers trained on a type-level morphological task, finding that these models acquire subphonemic features. We show that the degree to which these features are embedded in the transformer's representation depends on the feature's importance for the morphology and phonology of the language it is trained for; and on the locality of the feature's importance: in the encoder, long-context features are more salient.

We use these results to explain empirical training methods used in the morphology inflection domain. We hope this analysis will add an analytical tool in explaining morphological models using phonology acquisition.

Limitations

Our suggested probing method that outputs a compression score, although language-agnostic, might have underlying biases in its components: the transliteration tools, the character-based transformer and the morphology inflection datasets' quality. While discussing common strategies in morphology inflection, we omitted a popular data augmentation called *data hallucination* (Anastasopoulos and Neubig, 2019), where new training examples are synthesized from existing training examples by identifying a (possibly discontinuous) word stem and replacing this with a random character sequence. Since this augmentation might indulge phonologically invalid words, we decided not to incorporate it to our method and results.

Acknowledgments

We thank Evyatar Cohen for his valuable guidance regarding Hebrew grapheme-to-phoneme conversion, and the anonymous reviewers for their helpful feedback. This research was supported in part by the Israel Science Foundation (grant No. 1166/23).

References

Roee Aharoni and Yoav Goldberg. 2017. Morphological inflection generation with hard monotonic attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2004–2015, Vancouver, Canada. Association for Computational Linguistics.

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 984–996, Hong Kong, China. Association for Computational Linguistics.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

N. Chomsky and M. Halle. 1968. *The Sound Pattern of English*. Studies in English. Harper & Row.

E. Cohen. 2019. Hebrew to ipa transcriber. Zemereshet. https://www.zemereshet.co.il/m/hebrewToIpa.asp, updated 2024.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.

Micha Elsner. 2021. What transfers in morphological inflection? experiments with analogical models. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 154–166, Online. Association for Computational Linguistics.

Omer Goldman, David Guriel, and Reut Tsarfaty. 2022. (un)solving morphological inflection: Lemma overlap artificially inflates models' performance. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 864–870, Dublin, Ireland. Association for Computational Linguistics.

Kyle Gorman, Arya D. McCarthy, Ryan Cotterell, Ekaterina Vylomova, Miikka Silfverberg, and Magdalena Markowska. 2019. Weird inflects but OK: Making sense of morphological generation errors. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 140–151, Hong Kong, China. Association for Computational Linguistics.

David Guriel, Omer Goldman, and Reut Tsarfaty. 2023. Morphological inflection with phonological features.

In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 613–622, Toronto, Canada. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Jordan Kodner, Salam Khalifa, Khuyagbaatar Batsuren, Hossep Dolatian, Ryan Cotterell, Faruk Akkus, Antonios Anastasopoulos, Taras Andrushko, Aryaman Arora, Nona Atanalov, Gábor Bella, Elena Budianskaya, Yustinus Ghanggo Ate, Omer Goldman, David Guriel, Simon Guriel, Silvia Guriel-Agiashvili, Witold Kieraś, Andrew Krizhanovsky, Natalia Krizhanovsky, Igor Marchenko, Magdalena Markowska, Polina Mashkovtseva, Maria Nepomniashchaya, Daria Rodionova, Karina Scheifer, Alexandra Sorova, Anastasia Yemelina, Jeremiah Young, and Ekaterina Vylomova. 2022. SIGMORPHON–UniMorph 2022 shared task 0: Generalization and typologically diverse morphological inflection. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 176–203, Seattle, Washington. Association for Computational Linguistics.

Jordan Kodner, Salam Khalifa, and Sarah Ruth Brogden Payne. 2023a. Exploring linguistic probes for morphological generalization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8933–8941, Singapore. Association for Computational Linguistics.

Jordan Kodner, Sarah Payne, Salam Khalifa, and Zoey Liu. 2023b. Morphological inflection: A reality check. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6082–6101, Toronto, Canada. Association for Computational Linguistics.

Sudheer Kolachina and Lilla Magyar. 2019. What do phone embeddings learn about phonology? In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 160–169, Florence, Italy. Association for Computational Linguistics.

Ling Liu and Mans Hulden. 2022. Can a transformer pass the wug test? tuning copying bias in neural morphological inflection models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 739–749, Dublin, Ireland. Association for Computational Linguistics.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

Jeff Mielke. 2008. *The emergence of distinctive features*. Oxford University Press, Oxford.

Nicole Mirea and Klinton Bicknell. 2019. Using LSTMs to assess the obligatoriness of phonological distinctive features for phonotactic learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1595–1605, Florence, Italy. Association for Computational Linguistics.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.

Saliha Muradoglu and Mans Hulden. 2023. Do transformer models do phonology like a linguist? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8529–8537, Toronto, Canada. Association for Computational Linguistics.

Esther Ploeger, Wessel Poelman, Miryam de Lhoneux, and Johannes Bjerva. 2024. What is "typological diversity" in NLP? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5681–5700, Miami, Florida, USA. Association for Computational Linguistics.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain. Association for Computational Linguistics.

Jennifer Rodd. 1997. Recurrent neural-network learning of phonological regularities in Turkish. In *CoNLL97: Computational Natural Language Learning*.

Miikka Silfverberg, Francis Tyers, Garrett Nicolai, and Mans Hulden. 2021. Do RNN states encode abstract phonological alternations? In *Proceedings of the 2021 Conference of the North American Chapter of*

the Association for Computational Linguistics: Human Language Technologies, pages 5501–5513, Online. Association for Computational Linguistics.

Miikka P. Silfverberg, Lingshuang Mao, and Mans Hulden. 2018. Sound analogies with phoneme embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pages 136–144.

Julius Steuer, Johann-Mattis List, Badr M. Abdullah, and Dietrich Klakow. 2023. Information-theoretic characterization of vowel harmony: A cross-linguistic study on word lists. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 96–109, Dubrovnik, Croatia. Association for Computational Linguistics.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Adam Wiemerslage, Miikka Silfverberg, and Mans Hulden. 2018. Phonological features for morphological inflection. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 161–166, Brussels, Belgium. Association for Computational Linguistics.

Adam Wiemerslage, Changbing Yang, Garrett Nicolai, Miikka Silfverberg, and Katharina Kann. 2023. An investigation of noise in morphological inflection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3351–3365, Toronto, Canada. Association for Computational Linguistics.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the transformer to character-level transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online. Association for Computational Linguistics.

Changbing Yang, Ruixin (Ray) Yang, Garrett Nicolai, and Miikka Silfverberg. 2022. Generalizing morphological inflection systems to unseen lemmas. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 226–235, Seattle, Washington. Association for Computational Linguistics.

A Consonant Harmony Results

Figure 4 displays the results of the consonant harmony probe.

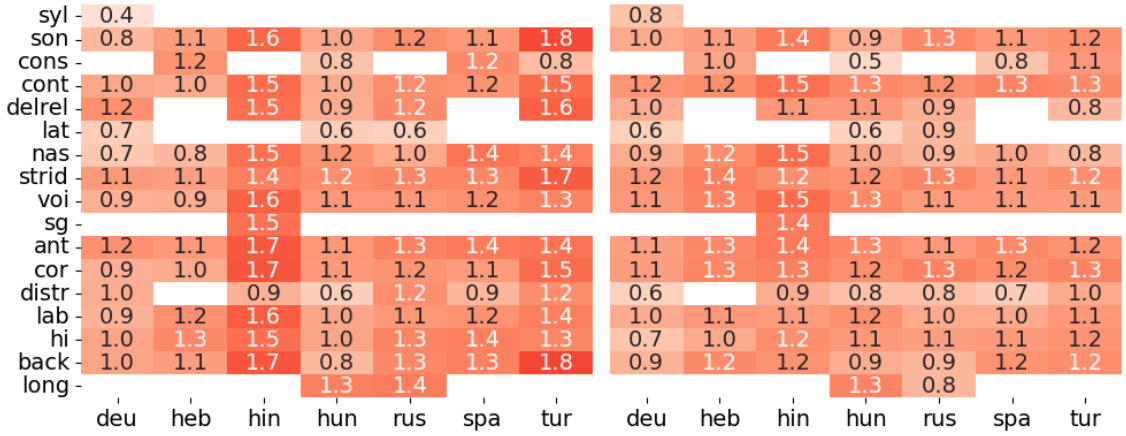

Figure 4: Compression scores for probing consonant harmony. Inflection model on the left, lemma copying on the right.