Information-Theoretic Characterization of Vowel Harmony: A Cross-Linguistic Study on Word Lists

Julius Steuer[ʊ] **Badr M. Abdullah**[ʊ] **Johann-Mattis List**[ʏ] **Dietrich Klakow**[ʊ]
[ʊ]Language Science and Technology (LST), Saarland University
[ʏ]MPI-EVA, Univ. of Passau
{ jsteuer, babdullah, dietrich }@lsv.uni-saarland.de, mattis.list@uni-passau.de

Abstract

We present a cross-linguistic study that aims to quantify vowel harmony using data-driven computational modeling. Concretely, we define an information-theoretic measure of harmonicity based on the predictability of vowels in a natural language lexicon, which we estimate using phoneme-level language models (PLMs). Prior quantitative studies have relied heavily on inflected word-forms in the analysis of vowel harmony. We instead train our models using cross-linguistically comparable lemma forms with little or no inflection, which enables us to cover more under-studied languages. Training data for our PLMs consists of word lists with a maximum of 1000 entries per language. Despite the fact that the data we employ are substantially smaller than previously used corpora, our experiments demonstrate the neural PLMs capture vowel harmony patterns in a set of languages that exhibit this phenomenon. Our work also demonstrates that word lists are a valuable resource for typological research, and offers new possibilities for future studies on low-resource, under-studied languages.

1 Introduction

1.1 Vowel Harmony

Many of the world's languages exhibit vowel harmony – a phonological co-occurrence constraint whereby vowels in polysyllabic words have to be members of the same natural class (Ohala, 1994). Natural classes of vowels are defined with respect to polar phonological features such as vowel backness (±BACK) and roundedness (±ROUND). In a prototypical language with backness, or ±BACK harmony, all vowels within a word tend to share the ±BACK feature, i.e. they are either all front (−BACK) or back (+BACK). Table 1 illustrates vowel harmony in Turkish, one of the languages best known to have this feature. In Table 1, the nominative plural and genitive plural are examples of −BACK harmony, while the genitive singular column of +BACK harmony. In the case of Turkish, vowel harmony can be defined as a constraint applying to almost all words and the entire inflectional system. In other languages vowel harmony may be restricted to the inflectional system, or even only a subset of inflectional suffixes. For example, In Estonian there are vestiges of vowel harmony in lexical items and it is absent from the inflectional system, while in Bislama it only occurs in a single suffix marking transivity (Crowley, 2014). Between these extremes of Turkish and Bislama lie languages such as Finnish and Hungarian, with intermediate vowel harmony systems where not all vowels participate in vowel harmony to the same extent. Both languages have ±BACK harmony, but a subset of the −BACK vowels allow +BACK harmony to spread: In a word like [lɑtikːo] 'box' (not [lɑtikːø]), +BACK harmony is not violated, whereas a word containing only neutral vowels triggers −BACK harmony, as in [merkitys] 'meaning' where the +BACK disharmonic form [merkitus] is not possible.

The rather broad application of the term has made it increasingly difficult to define it as a phonological process (cf. Anderson 1980). If vowel harmony is used as a typological feature to group languages into phylogenetic families, this broad application becomes perilous to the researcher since they have to be aware of the the degree of vowel harmonicity in the individual languages. Instead of searching for a necessarily complex definition of vowel harmony, research has consequentially concentrated on a quantitative description.

1.2 Prior Work and Scope

Prior approaches to a quantitative description of vowel harmony have mostly focused on strictly local harmony processes. Mayer et al. (2010) used vowel succession counts derived from corpora of inflected word-forms to quantify vowel harmony in a large number of languages in terms of χ^2-values,

Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2023), pages 96–109
May 6, 2023 ©2023 Association for Computational Linguistics

	Nom. Sg.	Gen. Sg.	Nom. Pl.	Gen. Pl.	Gloss
−BACK/−ROUND	[ip]	[ip-in]	[ip-lɛr]	[ip-lɛr-in]	’string’
+BACK/−ROUND	[kɯz]	[kɯz-ɯn]	[kɯz-lar]	[kɯz-lar-ɯn]	’girl’
−BACK/+ROUND	[jyz]	[jyz-yn]	[jyz-lɛr]	[jyz-lɛr-in]	’face’
+BACK/+ROUND	[pul]	[pul-un]	[pul-lar]	[pul-lar-ɯn]	’stamp’

Table 1: Illustration of the Turkish vowel harmony system following Polgárdi (1999). The first vowel of a word form determines the harmony type. If the first vowel is +BACK, the vowels of the following suffixes must agree w. r. t. the +BACK feature. ±ROUND harmony applies only in suffixes that have separate forms for this feature: The genitive suffix takes both ±BACK and ±ROUND forms, while the plural suffix varies only for ±BACK.

while Ozburn (2019) used count data to estimate succession probabilities and calculate the relative risk of encountering an harmonic vowel in a word form. These two approaches treated all positions in a word form identically. Goldsmith and Riggle (2012) argued that vowel harmony involves at least one type of non-local dependency, since it operates over consonants intervening between adjacent vowels. They employed a simple *n*-gram language model to learn the phonology of Finnish and calculated pointwise mutual information of vowel-vowel and consonant-vowel pairs based on the phoneme probabilities predicted by the language model, finding evidence for consonant-vowel harmony besides the expected ±BACK harmony, with a small bias towards +BACK harmony. However, *n*-gram language models are limited by their predefined context size. A language model with a left-hand context of $n = 3$ cannot capture the effect of vowel harmony if it operates over a neutral vowel intervening between two harmonic vowels. While this effect could be mitigated by allowing by allowing for a larger or flexible n, estimating probabilities from corpora becomes increasingly difficult with higher values of n. In this study we aim to improve over these methods by quantifying vowel harmony with a information-theoretic measure based on *surprisal*, capturing the relative strength of vowel harmony in language in terms of the likelihood of a vowel in a word to share a specific feature with preceding vowels. To do so, we employ neural recurrent language models with variable-length preceding phoneme context that are trained on cross-linguistically comparable lexical data. While some previous work on modeling vowel harmony with language models has been carried out (Rodd, 1997), finding evidence for Turkish vowel harmony in the hidden activations of a simple neural language model, it seems that this topic has not been further explored since then. In the following section, we first introduce feature surprisal as an information-theoretic measure of vowel harmony (§2). We then present our computational experiments with the introduced measure of vowel harmony and discuss the results of their application to a large collection of cross-linguistic lexical data (§3, §4). We conclude by discussing the implications of our study for future studies on vowel harmony in classical and computational studies (§5).

2 Quantifying Vowel Harmony

2.1 Phoneme-Level Language Models

Preliminaries and Notations. To quantify vowel harmony in our study, we make use of phoneme-level language models (PLMs). Consider a natural language with a lexicon $\mathcal{L}$ and a phoneme inventory $\mathbf{\Phi}$ (using IPA symbols). Using a cross-linguistic word list, we obtain K samples from the lexicon $\mathcal{D} = \{\boldsymbol{w}^k\}_{k=1}^{K} \sim \mathcal{L}$ where each sample is a word-form that is transcribed as a phoneme sequence $\boldsymbol{w} = (\varphi_1, \cdots, \varphi_{|\boldsymbol{w}|}) \in \mathbf{\Phi}^*$. Given this sample of word-forms as training data, a PLM can be trained to estimate a probability distribution over $\mathbf{\Phi}$ by maximizing the term

$$\begin{aligned} J(\theta, \mathcal{D}) &= \sum_{\boldsymbol{w} \in \mathcal{D}} p(\boldsymbol{w}; \theta) \\ &= \sum_{\boldsymbol{w} \in \mathcal{D}} \prod_{t \in \{1, \cdots, |\boldsymbol{w}|\}} p(\varphi_t | \boldsymbol{\varphi}_{<t}; \theta) \end{aligned} \tag{1}$$

Here, θ are the parameters of the model that are learned by maximizing the objective function above. Once a PLM has been trained, it can be used to compute the probability of unseen, held-out word-forms (i.e, word-forms that were not observed in the training data). Ideally, a PLM should assign a higher probability mass to plausible word-forms given the phonotactic rules of the language of the train data, and lower probability to implausible word-forms.

Recurrent PLMs. Although different architectures can be used to build a PLM, we choose to employ a recurrent architecture based on unidirectional long short-term memory (LSTM) cell (Hochreiter and Schmidhuber, 1997). Given a word-form as a sequence of phonemes $\boldsymbol{w} = (\varphi_1, \cdots, \varphi_{|\boldsymbol{w}|})$, each phoneme is first projected into a continuous-vector phoneme representation using an embedding matrix as $\mathbf{E}(\varphi_t) = \mathbf{x}_t \in \mathbb{R}^d$. Then, the LSTM takes as input the sequence at each position t within the word-form to compute the hidden state representation

$$\mathbf{h}_t = \mathcal{F}_{\text{LSTM}}(\mathbf{x}_t, \mathbf{h}_{t-1}) \in \mathbb{R}^h \tag{2}$$

To obtain a probability distribution over the phoneme inventory, a linear transformation is applied on the hidden state vector followed by a softmax function to obtain a probability vector as

$$p(\varphi_t|\boldsymbol{\varphi}_{<t}) = \text{SOFTMAX}(\mathbf{W}\mathbf{h}_t + \mathbf{b}) \tag{3}$$

Here, $\mathbf{W} \in \mathbb{R}^{|\boldsymbol{\Phi}| \times h}$ is a projection matrix at the network output and $\mathbf{b} \in \mathbb{R}^{|\boldsymbol{\Phi}|}$ is a bias term.

Nevertheless, we make a few (trivial) design modifications to the vanilla LSTM-based PLMs to make them more suitable for our study. First, since our main interest is to model the predictability of the vowels, we confine the output probability distribution to be over the set of vocalic segments, which is a subset of the phoneme inventory $\mathcal{V} \subset \boldsymbol{\Phi}$. Second, we train and evaluate our PLMs to predict the next vowel only in the intra-word positions where we know that the next phoneme is indeed a vowel, given a preceding phoneme context that contains at least one vowel. While the output in this modified PLM is over the set $\mathcal{V}$, the word-forms remain sequences in $\boldsymbol{\Phi}^*$. That is, both consonants and vowels could appear in the preceding context.

Note that we do not employ fixed-length context n-gram PLMs in our study since we aim to account for non-local phoneme dependencies within a word-form. Given that word-forms within a lexicon have arbitrary lengths, restricting the preceding context to a fixed number of phonemes does not enable us to model vowel harmony across variable-length contexts beyond phoneme n-grams. On the other hand, we do not employ more powerful architectures such as a transformer (Vaswani et al., 2017) or a bidirectional LSTM (Graves and Schmidhuber, 2005) on grounds of suitability for the task: (1) the dependencies between vowels are relatively short (the domain of vowel harmony is the phonological word), (2) vowel harmony is a progressive phenomenon (i.e., operates from left to right–unlike its regressive counterpart *umlaut*), and (3) the training sets of the individual languages in our study are likely too small to train a large transformer model. Moreover, several prior studies within the information-theoretic approaches to investigate phonological structure have also employed LSTM-based PLMs (e.g., Pimentel et al., 2020, 2021a).

2.2 Harmony as Surprisal

Given that our phoneme-level language model that was trained on a set of word-forms sampled from a natural language lexicon, we can quantify the vowel harmony phenomenon using Shannon's information content, or **surprisal**. Given a non-initial vocalic position t after a phoneme context $\varphi_{<t}$, vowel surprisal is

$$\eta(v, t) = -\log_2 p(v \,|\, t, \boldsymbol{\varphi}_{<t}) \tag{4}$$

which is measured in bits. Note that surprisal is maximal when the preceding context tells us nothing about which vowels are more likely to occur. That is, if the vowels are sampled from a uniform distribution over the vowel inventory $\mathcal{V}$, then $\eta(v, t) = \log_2 |\mathcal{V}|$ (bits). Therefore, surprisal in our case is mainly a metric of how "predictable" a vowel is in a given context. Now consider a set of vowels $\mathcal{H} \in \mathcal{V}$ that share a phonological feature. For a given vowel $v \in \mathcal{H}$, we refer to the set $\mathcal{H}$ as a harmonic group, while its disharmonic counterpart $\neg\mathcal{H} \in \mathcal{V} \setminus \mathcal{H}$ as a disharmonic group with respect to the vowel v. For example, consider the front vowel [i] in Turkish that has the feature $-$BACK. With respect to [i], the front vowels in the Turkish vowel inventory $\mathcal{H} = \{[\text{i}], [\text{e}], [\text{y}], [\text{œ}]\}$ make a harmonic group since they all share the feature $-$BACK, while the rest of the vowels make a disharmonic group $\neg\mathcal{H} = \{[\text{ɯ}], [\text{a}], [\text{u}], [\text{o}]\}$ since they all lack the feature $-$BACK. Given a phoneme context that contains at least one vowel v such that $v \in \mathcal{H}$, we compute the surprisal of a harmonic group at position t in a word-form by summing over the vowels in $\mathcal{H}$, i.e.

$$\eta(\mathcal{H}, t) = -\log_2 \sum_{\pi \in \mathcal{H}} p(\pi \,|\, t, \boldsymbol{\varphi}_{<t}) \tag{5}$$

We refer to the quantity $\eta(\mathcal{H}, t)$ as **feature surprisal**, since all members of the harmonic group

Language	Harmonic Groups		
Finnish	−BACK {y, ø, æ}	+BACK {u, o, ɑ}	BACK neutral {e, i}
Hungarian	−BACK {y, ø}	+BACK {u, o, ɒ}	BACK neutral {e, i}
Manchu	−BACK {e/ɤ}	+BACK {ɑ, ɔ}	BACK neutral {i, u}

Khalkha Mongolian	−ATR {e, u, ɔ}
	−ROUND {e, a, i}
		+ATR {a, ɒ, o}
		+ROUND {o}
			ATR neutral {i}
			ROUND neutral {u, ʊ}

Turkish	−BACK {i, e, y, œ}
	−ROUND {i, e, u, o}
		+BACK {ɯ, a, u, o}
		+ROUND {ɯ, a, y, œ}	

Arabic, Ainu, Armenian, Basque, Estonian†	–	–	–

Table 2: Languages from NorthEuraLex used in our sample along with their harmonic groups. Khalkha Mongolian has a special type of vowel harmony involving the placement of the tongue root: +ATR codes an advanced position of the tongue root in the vocal tract, while −ATR encodes an retracted or further back position. Languages in our sample that do not exhibit vowel harmony are marked with the symbol (†).

$\mathcal{H}$ share one phonological feature. Likewise, we compute the surprisal of a disharmonic group by summing over the vowels in $\neg\mathcal{H}$ as

$$\eta(\neg\mathcal{H}, t) = -\log_2 \sum_{\pi \in \neg\mathcal{H}} p(\pi \mid t, \boldsymbol{\varphi}_{<t}) \quad (6)$$

Assuming that a PLM has learned the vowel harmony constraints of a language from the training word-forms, we expect the model to predict that vowels in $\mathcal{H}$ are more likely to co-occur in a single word-form. By implication, we expect the model to "disfavour" the occurrence of a vowel in $\neg\mathcal{H}$ when observing members of $\mathcal{H}$ in the context. That is, in a language that exhibits this linguistic phenomenon, word-forms that conform to vowel harmony should be assigned a higher probability than word-forms that do not. For example, the Finnish word form [s i l m æ s ae] is expected to be assigned a high probability by our model since the sequence of vowels [i], [ae], [ae] is −BACK harmonic, and its disharmonic counterpart [s i l m æ s o] is expected to be assigned a lower probability.

Note in equations (5) and (6) we compute the surprisal at a single vocalic position in a given word-form. To quantify harmonic group surprisal across a set of held-out word-forms $\mathcal{W}$, we compute the quantity

$$\overline{\eta}(\mathcal{H}) = -\frac{1}{|\mathcal{W}|} \sum_{\boldsymbol{w} \in \mathcal{W}} \sum_{t \in \{\tau, \cdots, T\}} \eta(\mathcal{H}, t) \quad (7)$$

which is the average feature surprisal. Here, the outer sum $\sum_{\boldsymbol{w} \in \mathcal{W}}$ iterates over all word-forms in $\mathcal{W}$, while the inner sum $\sum_{t \in \{\tau, \cdots, T\}}$ iterates over non-initial vocalic positions within the word-form $\boldsymbol{w}$. The feature surprisal of a disharmonic group $\overline{\eta}(\neg\mathcal{H})$ is computed in the same way as in equation (7) but summing over the term $\eta(\neg\mathcal{H}, t)$ instead.

Finally, we quantify the strength of a vowel harmony constraint in a language as the difference of feature surprisal of the harmonic and disharmonic vowels

$$\Delta_\eta = \overline{\eta}(\mathcal{H}) - \overline{\eta}(\neg\mathcal{H}) \quad (8)$$

If feature surprisal in harmonic phoneme sequences is lower than feature surprisal in disharmonic phoneme sequences, Δ_η is negative, indicating that harmonic sequences are assigned higher probability. It is worth pointing out that our grouping of the vowels into harmonic groups is only used to obtain feature surprisal values from the model after it has been trained. That is, our PLMs for all languages in our study are trained without an explicit signal that informs the model about the features of the vowels.

3 Experimental Data and Setup

3.1 Data

Previous research has made use of large corpora of inflected word-forms (Goldsmith and Riggle, 2012) or running text (Mayer et al., 2010) to infer vowel harmony patterns. This is mainly because vowel harmony constraints often surface in inflectional suffixes, especially in highly agglutinating languages such as Finnish, Hungarian or Turkish. Though this approach is not in itself problematic, it relies on data that may not exist for the majority of the world's languages. It is also not applicable for languages that have a different grammatical structure, for example, reduced or fusional morphology. On the other hand, if a language has vowel harmony as a phonologically conditioned rather than a purely grammatical phenomenon, the relevant vowel harmony patterns should also be recoverable from lexical data with little or no inflection at all.

We use parts of the NorthEuraLex database (`http://www.northeuralex.org/`, Dellert et al.

	Maximum	Minimum	Average	Median
Phoneme inventory size	72 (Skolt Sami)	23 (Turkish)	38.9	37
Number of word-forms	1513 (Manchu)	677 (Italian)	1136.6	1142

Table 3: Inventory sizes and word list lengths in the data sampled from NorthEuraLex.

2020) as experimental data to train our phoneme language models and quantify the effect of vowel harmony in languages that are known to exhibit this linguistic phenomenon. NorthEuraLex offers a large multilingual word list consisting of 1005 concepts translated into 107 language varieties from North Eurasia with translations provided in a unified transcription following the International Phonetic Alphabet (IPA). Moreover, NorthEuraLex contains a larger number of diverse language varieties from various language families that are known to exhibit vowel harmony, as well as language varieties that are known to lack the phenomenon.

As there is no clear definition of what constitutes vowel harmony in languages, and linguistic resources such as the World Atlas of Language Structures (Dryer et al., 2014) do not provide this information, we concentrate on a subset of 10 language varieties from NorthEuraLex, with five varieties traditionally known to exhibit vowel harmony, and five known to not exhibit the phenomenon. When selecting the languages, we tried to obtain a rather diverse sample of languages from different language families. Table 2 gives an overview over the languages and their active harmony processes (where present).

The NorthEuraLex data is available in the form of Cross-Linguistic Data Formats (CLDF `https://cldf.clld.org`, Forkel et al. 2018), following the recommendations underlying Lexibank (List et al., 2022a), a large collection of lexical word lists (`https://github.com/lexibank/northeuralex`). A core feature of CLDF is the integration of *reference catalogs*. Reference catalogs are metadata collections that offer basic information on major linguistic constructs, such as languages (Glottolog, `https://glottolog.org`, Hammarström et al. 2022) or concepts (Concepticon, `https://concepticon.clld.org`, List et al. 2022b). In addition to offering word lists standardized with respect to language names and concept elicitation glosses, Lexibank offers standardized phonetic transcriptions as specified by Cross-Linguistic Transcription Systems (CLTS, `https://clts.clld.org`, List et al. 2021), a reference catalog that offers a transcription system that conforms to the IPA but resolves ambiguities encountered in the original IPA specification (Anderson et al., 2018).

Since NorthEuraLex is available in CLDF, this means that we have direct access to standardized phonetic transcriptions segmented into individual sounds in each word form along with an underlying set of distinctive features provided by CLTS. The resulting data set provides on average 1136 unique word-forms per language (with several concepts having two or more word-forms as translational equivalents), with larger differences between individual languages. We decided against downsampling word lists to a common size due to the already small number of samples. The word list sizes range from 971 (Ainu) to 1513 (Manchu).

3.2 Preprocessing

For each of the languages, identical word-forms are collapsed to a single item, such that each sequence of phonemes is presented only once to the model. In addition, word-forms which are a substring of another word form are also ignored. Thus, if the word list of a language contains the sequences { [s i l m æ], [s i l m æ], [s i l m æ sː æ], [s i l m æ d æ] }, only the latter two sequences are kept: { [s i l m æ sː æ], [s i l m æ d æ] }. This procedure ensures that only unique sequences are presented to the model, and that train and test splits do not contain identical forms, which might otherwise lead to unjustified higher weights for sound sequences recurring across the vocabulary of individual language varieties.

3.3 Training

For each language, we randomly split the data into 60%, 10% and 30% subsets for train, validation and test splits respectively. The models were trained with the Adam optimizer (Kingma and Ba, 2015) on the task of minimizing the cross entropy of the predicted distribution and the true probability distributions over the vowel inventory. This is equivalent to minimizing the negative log-likelihood of the true phoneme at each position. 25% of the inputs were randomly replaced by a mask token to prevent overfitting on the relatively small sample. Note that the output probability distribution of the model is restricted to the vowel inventory of the language plus the end-of-sequence token, since only the vowel positions are of interest for the analysis.

A separate model was trained for each language in our subset of 10 languages from NorthEuraLex. The same hyperparameters were used for training as in Pimentel et al. (2021b), with batch size reduced to 32 since NorthEuraLex wordlists are considerably smaller than the datasets used in that paper. Table 4 in Appendix A shows the exact configuration of the hyperparameters. After each epoch the models were evaluated on a validation set, and all models were trained until validation loss converged. Training the models on unique sequences derived from word lists ensures that the model sees each sequence only once per epoch, and minimizes overlaps between train, test and validation set.

3.4 Significance Tests

As the expected behavior of vowel harmony languages is that the vowels are not evenly distributed over their words, average feature surprisal is likely to not be normally distributed. The Shapiro-Wilk test (Shapiro and Wilk, 1965) was used to check whether the surprisal values for every comparison. For every pairing of conditions at least one of them was not normally distributed with $p < 0.01$. Thus, the Wilcoxon signed-rank test was conducted to test the significance of a paired contrast (as in the example above). Effect size was calculated as the rank-biserial coefficient using the common language effect size $f = \frac{U}{n_1 \cdot n_2}$ as $r = f - (1 - f)$, with U being the test statistic and $n_1 \cdot n_2$ being the number of possible comparisons between two conditions. For an unpaired contrast (e.g. the contrast between average feature surprisal for $+$ROUND after a $-$ROUND vowel and average feature surprisal for $+$BACK after a $-$BACK vowel) a Mann-Whitney U-test was conducted, with effect size calculated as the rank-biserial coefficient using the T statistic and the sum of ranks S as $r = \frac{T}{S}$. All significance tests were conducted using the SciPy Python package (Virtanen et al., 2020).

3.5 Implementation

The methods described here are implemented in Python. The PyTorch library (Paszke et al., 2019) is used to train and evaluate our neural models. CLDF data are accessed with the help of CL Toolkit (https://pypi.org/project/cltoolkit, List and Forkel 2021), a Python package that provides convenient access to lexical word lists in CLDF.

4 Experimental Results

4.1 Feature Surprisal

All vowel harmony languages show significant differences in feature surprisal between harmonic and disharmonic conditions with negative Δ_η; individual results can be retrieved from the result tables 6-10 in Appendix C. Feature surprisal in the $+$BACK disharmonic condition was found to be higher than feature surprisal in the $-$BACK disharmonic condition for Finnish ($\Delta_\eta = -0.2148$, $p < 0.01$), Hungarian ($\Delta_\eta = -1.0806$, $p < 0.01$) and Turkish ($\Delta_\eta = -0.8602$, $p < 0.01$), which confirms the findings of Goldsmith (1985). Note that if the $+$BACK and $-$BACK harmony were equally strong, one would expect no difference in surprisal if the harmony is violated. Three out of four languages with $\pm$BACK harmony show this tendency, indicating that the relative strength of $+$BACK harmony over $-$BACK harmony is the usual case rather than an exception. A possible explanation for this difference in strength is the existence of neutral vowels, with 3 of the 4 $\pm$BACK harmony languages in our sample having at least one neutral vowel, and Turkish, the only language without neutral vowels, also showing the largest difference between the two disharmonic conditions . The probabilities of the neutral vowels are not included in the feature surprisal calculation, causing feature surprisal to be higher in the $+$BACK disharmonic condition while lowering feature surprisal in the $-$BACK disharmonic condition. For Hungarian feature surprisal was lowest in the neutral harmonic condition, meaning that neutral vowels are most likely to occur after another neutral vowel. Even though Hungarian neutral vowels trigger $-$BACK harmony, the low number of forms containing both $-$BACK vowels and neutral vowels makes it difficult for the neural language model to learn the pattern, leading to the highest feature surprisal occurring in the harmonic condition (i.e. for the $-$BACK feature). Figure 1 gives an overview of the relative strength of vowel harmony for all languages and harmonic features in the sample used in this study. For this figure the sign of Δ_η was reversed in order to quantify the reduction of feature surprisal in the harmonic sequences as compared to the dishar-

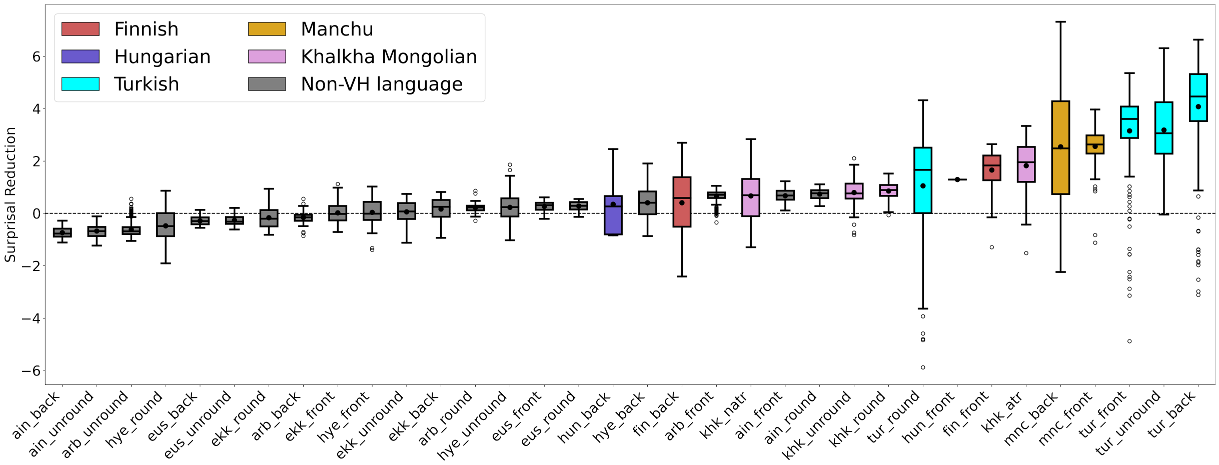

Figure 1: Surprisal reduction for the 10 varieties from NorthEuraLex. Best viewed in color.

monic sequences for each combination of feature and language. The boxplots of languages without vowel harmony are located towards the left of the plot with small differences between harmonic and disharmonic sequences, with some vowel harmony languages showing similar, yet still positive surprisal reduction (e.g. Finnish +BACK vowels, Hungarian +BACK vowels)

4.2 The Case of Turkish

For Turkish the difference in feature surprisal between harmonic and disharmonic conditions was large. Figure 2 shows that for both the ±BACK and ±ROUND conditions, the disharmonic condition displays a much higher surprisal value as compared to the harmonic condition ($\Delta_\eta = -3.6816$, $p < 0.01$ and $\Delta_\eta = -2.7061$, $p < 0.01$ respectively). A small but significant bias towards +BACK harmony was detected ($\Delta_\eta = -0.8602$, $p < 0.01$). There is one obvious reason for the relative strength of ±BACK harmony over ±ROUND, namely the parasitic nature of ±ROUND harmony in Turkish: while all morphemes have different forms for ±BACK, allowing for ±ROUND disharmony, only a subset also has separate forms for ±ROUND (Tab. 1). Thus, there are more instances of ±BACK harmony to be observed by the model, and this is expected to result in higher surprisal values for the ±BACK disharmonic conditions.

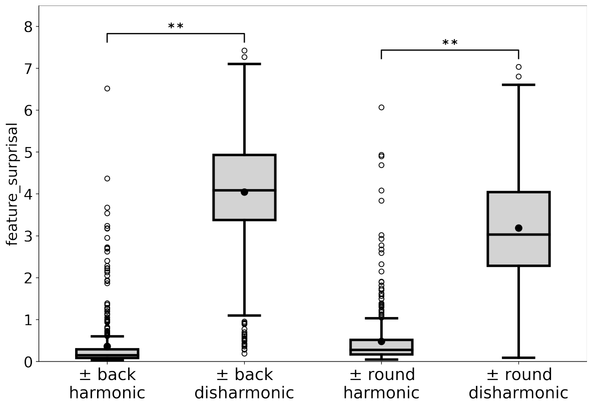

Figure 2: Feature surprisal for Turkish back harmonic/disharmonic sequences (left) and round harmonic/disharmonic sequences (right). The difference between harmonic and disharmonic conditions is significant with $p < 0.01$ in both cases.**: $p < 0.01$, *: $p < 0.05$, **ns**: $p > 0.05$

After ±ROUND vowels feature surprisal was also much higher in the disharmonic conditions, with feature surprisal in the round disharmonic condition being higher than in the unrounded disharmonic condition ($\Delta_\eta = -1.5827$, $p < 0.01$). In other words, +ROUND harmony seems to be stronger than −ROUND harmony in Turkish. When combining the disharmonic conditions within a harmonic feature and comparing them to the disharmonic conditions in the other harmonic feature, the combined back disharmonic condition (both front disharmonic and back disharmonic) yields slightly higher feature surprisal than the combined rounded disharmonic condition ($\Delta_\eta = 0.8555$, $p < 0.01$); see Table 8 in the appendix. This is in line with earlier research (Baker, 2009) that found a bias towards ±BACK harmony over ±ROUND harmony. This is also the expected result when taking into account that many suffixes do not have +ROUND forms and therefore introduce noise to the data.

4.3 Neutral Vowels

Learning vowel dependencies across neutral vowels turned out to be difficult: For Manchu and

Khalkha Mongolian the number of test items in this category was so low that no meaningful result could be produced. This is again caused by the nature of the data which consists of lemma forms. For Finnish and Hungarian the number of items was sufficient to conduct the appropriate significance tests, but the numbers are still small (102 and 63 respectively). The neural language model did not learn the association of neutral vowels with $-$BACK as assumed for Finnish and Hungarian, with significant $\Delta_\eta > 0$ between the neutral harmonic and neutral disharmonic condition only for Khalkha Mongolian and ±ATR sequences. In Hungarian, neutral vowels are most likely to occur after other neutral vowels, but this is not the case for Finnish, Manchu and Khalkha Mongolian. On the other hand, Turkish as the only language in the sample without neutral vowels showed the largest difference between harmonic and disharmonic conditions for both ±BACK and ±ROUND (see App. C for results).

It may be noted that Turkish, the language with the strongest vowel harmony effect in terms of Δ_η, has no neutral vowels both for ±BACK and ±ROUND harmony. This could have facilitated the generalization on the ±BACK and ±ROUND harmony patterns for the neural language model, at least proving that the neural language model does indeed assign higher surprisal to disharmonic sequences, since there the harmony system is symmetrical and the number of vowels is the same for each feature.

5 Discussion and Conclusion

Prior work in the (computational) linguistics community has adopted information theory as a framework for the study of human language structure across different linguistic levels including phonology (e.g., Pimentel et al., 2020, 2021c), morphology (e.g., Rathi et al., 2021; Wu et al., 2019), and syntax (e.g., Hahn et al., 2018; Futrell et al., 2015). Following the same spirit, we have introduced an information-theoretic metric to quantify vowel harmony based on feature surprisal. Our experiments have demonstrated that feature surprisal is a good indicator of whether a certain feature participates in vowel harmony patterns in a language, producing significant differences between harmonic and disharmonic conditions for most harmonic features in five vowel harmony languages. The effect was found on a very small sample of lemma forms with little to no morphological information, showing that large amounts of inflectional data are not necessary to identify some, but not all vowel harmony constraints. When calculated for ±BACK and ±ROUND features for five non-vowel harmony languages, the difference in surprisal was close to zero, meaning the neural language model did not detect any preference for harmony constraints in the languages evaluated.

We showed that neural language models can capture non-local harmony constraints over neutral vowels, which is not possible with count-based methods as employed by Mayer et al. (2010) or bigram models as in (Goldsmith and Riggle, 2012). Here the resolution of the analysis is more fine-grained with respect to the features underlying the harmonic groups. The advantage of the modeling approach presented here over both count-based and probabilistic models is that it can be used with a small dataset (word lists of about 1000 word-forms, of which ca. 300 are in the test set as the basis of the actual analysis).

The analysis presented could be extended to other types of phonological constraints, since neural language models in theory are able to learn all types of dependencies over sequences of arbitrary length. However, analysing Finnish, Hungarian, Manchu and Khalkha Mongolian required prior knowledge about harmonic vowels and the split of vowels into harmonic groups, either because the groups are not defined by the value of a feature as is the case for languages with neutral vowels, or because the feature representation in our standardized data itself might not describe a sound with the feature that is assumed to participate in vowel harmony.

If it is not known which vowels participate in vowel harmony, it seems best to use information on distinctive features in the data in order to find out which effects can be observed. However, if the vowel harmony patterns are as complex as in Khalkha Mongolian, the approach presented here would probably find its limits in corpus size. Identifying the approximate number of distinct word-forms needed to infer vowel harmony systems of individual language varieties (similar to previous studies inferring the number of words needed to get an approximate account of phoneme numbers, Dockum and Bowern 2019) would be an interesting topic for future analysis.

Limitations

The limiting factor in the analysis of Hungarian and Khalkha Mongolian was the low number of items with more than two vowels in the test data. Although this was less of a problem in the other three languages (Finnish, Turkish and Manchu all have 400+ items with three or more vowels), this is likely the case for many of the languages in NorthEuraLex. Figure 3 in Appendix B shows that many languages have an even lower number of items with more than three vowels than Finnish and Khalkha Mongolian. Given a train-valid-test split of 60%-10%-30%, the number of items available to the analysis of long-range dependencies (including, but not restricted to, the operation of vowel harmony across neutral vowels) will be very low for these languages. This is an inherent property of the data, and could only be amended by using larger word lists or a larger corpora that are not restricted to lemma forms.

Ethics Statement

The authors foresee no ethical concerns about the work presented in the paper.

Supplementary Material

The supplementary material accompanying this study was archived with Zenodo (`https://doi.org/10.5281/zenodo.7782090`). It contains all data and code needed to replicate this study, along with extensive instructions.

Acknowledgements

We thank the anonymous reviewers for their constructive comments. This research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project ID 232722074 – SFB 1102 (Julius Steuer, Badr Abdullah), by the Max Planck Society Research Grant *CALC³* (JML, `https://digling.org/calc/`), and the ERC Consolidator Grant *ProduSemy* (JML, Grant No. 101044282, see `https://doi.org/10.3030/101044282`). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency (nor any other funding agencies involved). Neither the European Union nor the granting authority can be held responsible for them.

References

Cormac Anderson, Tiago Tresoldi, Thiago Costa Chacon, Anne-Maria Fehn, Mary Walworth, Robert Forkel, and Johann-Mattis List. 2018. A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznań Linguistic Meeting*, 4(1):21–53.

Stephen R. Anderson. 1980. Problems and perspectives in the description of vowel harmony. In Robert M. Vago, editor, *Issues in Vowel Harmony*, volume 6, pages 1–48. John Benjamins Publishing Company, Amsterdam.

Adam C. Baker. 2009. Two statistical approaches to finding vowel harmony. Technical report, University of Chicago.

Terry Crowley. 2014. *Bislama reference grammar*. University of Hawaii Press.

Johannes Dellert, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, Mohamed Balabel, Hizniye Isabella Boga, Zalina Baysarova, Roland Mühlenbernd, Johannes Wahle, and Gerhard Jäger. 2020. NorthEuraLex: a wide-coverage lexical database of Northern Eurasia. *Language Resources and Evaluation*, 54(1):273–301.

Rikker Dockum and Claire Bowern. 2019. Swadesh lists are not long enough: Drawing phonological generalizations from limited data. *Language Documentation and Description*, 16:35–54.

Matthew Dryer, Martin Haspelmath, and Robert Forkel. 2014. *WALS Online [Dataset, Version 2014.2]*. Zenodo, Geneva.

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5(180205):1–10.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Quantifying word order freedom in dependency corpora. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 91–100, Uppsala, Sweden. Uppsala University, Uppsala, Sweden.

John Goldsmith. 1985. Vowel harmony in Khalkha Mongolian, Yaka, Finnish and Hungarian. *Phonology Yearbook*, 2(1):253–275.

John Goldsmith and Jason Riggle. 2012. Information theoretic approaches to phonological structure: the case of Finnish vowel harmony. *Natural Language & Linguistic Theory*, 30(3):859–896.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM networks. In *Proceedings. 2005 IEEE International*

Joint Conference on Neural Networks, 2005., volume 4, pages 2047–2052, Montreal, Que., Canada. IEEE.

Michael Hahn, Judith Degen, Noah Goodman, Daniel Jurafsky, and Richard Futrell. 2018. An information-theoretic explanation of adjective ordering preferences. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, pages 1766–1772, Madison, WI. Cognitive Science Society.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2022. *Glottolog [Dataset, Version 4.7]*. Zenodo, Geneva.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Johann-Mattis List, Cormac Anderson, Tiago Tresoldi, and Robert Forkel. 2021. *Cross-Linguistic Transcription Systems [Dataset, Version 2.1.0]*. Max Planck Institute for the Science of Human History, Jena.

Johann-Mattis List and Robert Forkel. 2021. *CL Toolkit. A Python library for the processing of cross-linguistic data [Software package, Version 0.1.1]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Johann-Mattis List, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2022a. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9(1):316.

Johann-Mattis List, Annika Tjuka, Christoph Rzymski, Simon J. Greenhill, and Robert Forkel. 2022b. *CLLD Concepticon [Dataset, Version 3.0.0]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Thomas Mayer, Christian Rohrdantz, Miriam Butt, Frans Plank, and Daniel A. Keim. 2010. Visualizing vowel harmony. *Linguistic issues in language technology*, 4(2):1–33.

John J. Ohala. 1994. Towards a universal, phonetically-based, theory of vowel harmony. In *3rd International Conference on Spoken Language Processing (ICSLP 1994)*, pages 491–494. ISCA.

Avery Ozburn. 2019. A segment-specific metric for quantifying participation in harmony. *Proceedings of the Annual Meetings on Phonology*, 7.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Tiago Pimentel, Ryan Cotterell, and Brian Roark. 2021a. Disambiguatory signals are stronger in word-initial positions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 31–41, Online. Association for Computational Linguistics.

Tiago Pimentel, Ryan Cotterell, and Brian Roark. 2021b. Disambiguatory signals are stronger in word-initial positions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 31–41, Online. Association for Computational Linguistics.

Tiago Pimentel, Clara Meister, Elizabeth Salesky, Simone Teufel, Damián Blasi, and Ryan Cotterell. 2021c. A surprisal–duration trade-off across and within the world's languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 949–962, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tiago Pimentel, Brian Roark, and Ryan Cotterell. 2020. Phonotactic Complexity and Its Trade-offs. *Transactions of the Association for Computational Linguistics*, 8:1–18.

Krisztina Polgárdi. 1999. Vowel harmony and disharmony in Turkish. *The Linguistic Review*, 16(2):187–204.

Neil Rathi, Michael Hahn, and Richard Futrell. 2021. An information-theoretic characterization of morphological fusion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10115–10120, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Sam S. Shapiro and Martin B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611.

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

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew

Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, Ilhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, SciPy 1.0 Contributors, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272.

Shijie Wu, Ryan Cotterell, and Timothy O'Donnell. 2019. Morphological irregularity correlates with frequency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5117–5126, Florence, Italy. Association for Computational Linguistics.

A LSTM Hyperparameters

Hyperparameter	Value
Embedding Size	32
Hidden Size	256
LSTM Layers	2
Dropout	0.33
Batch Size	32

Table 4: Model and Training Hyperparameters as taken from (Pimentel et al., 2021b)

B Abbreviations of Harmonic Features

Abbreviation	Feature	
b	back	+BACK
f	front	−BACK
r	round	+ROUND
u	unround	−ROUND
atr	advanced tongue root	+ATR
natr	retracted tongue root	−ATR
n	neutral	
h	harmonic	
dish	disharmonic	

Table 5: Explanation of the abbreviations used in the result tables. The condition column refers to the type of harmony tested, with vowel successions abbreviated in the way described in this table. The sequence "f_n_f" represents sequences starting with a front/−BACK vowel, followed by a neutral/BACK neutral vowel and another front/BACK vowel. If more than one harmonic feature is present (as in Turkish, Manchu and Khalkha Mongolian), the magnitude of the effect on feature surprisal is compared between the two features in the disharmonic condition only (compare row "f_r/dish" in Table 8).

C Result Tables

Table 6: P-values, Δ_η and effect size for Finnish feature surprisal

Condition	Δ_η	Statistic	p-value	Effect Size	Test
f_f/f_b	-0.8298	71.0	2.e-12	0.0263	Wilcoxon
b_b/b_f	-0.8469	415.0	3.8e-17	0.0572	Wilcoxon
n_f/n_b	0.0009	4800.0	0.1723	0.4353	Wilcoxon
f_b/b_f	-0.2148	3148.0	0.001	-0.2813	Mann-Whitney
f_n_f/f_n_b	-0.563	59.0	7.57e-05	0.1052	Wilcoxon
b_n_b/b_n_f	-0.6077	236.0	0.0009	0.2183	Wilcoxon
n_n_f/n_n_b	-0.1206	85.0	0.1114	0.308	Wilcoxon
f_n_b/b_n_f	-0.1188	688.0	0.4834	-0.0935	Mann-Whitney

Table 7: P-values, Δ_η and effect size for Hungarian feature surprisal

Condition	Δ_η	Statistic	p-value	Effect Size	Test
f_f/f_b	-0.0917	270.0	0.64	0.4538	Wilcoxon
b_b/b_f	-2.1995	2.0	9.46e-21	0.0003	Wilcoxon
n_f/n_b	0.7951	1270.0	2.47e-14	0.1287	Wilcoxon
f_b/b_f	-1.0806	364.0	5.36e-13	-0.8154	Mann-Whitney
f_n_f/f_n_b	0.0864	27.0	1.0	0.4909	Wilcoxon
b_n_b/b_n_f	-1.6036	0.0	0.0078	0.0	Wilcoxon
n_n_f/n_n_b	0.4453	243.0	0.0019	0.2348	Wilcoxon
f_n_b/b_n_f	-0.674	24.0	0.1728	-0.4	Mann-Whitney

Table 8: P-values, Δ_η and effect size for Turkish feature surprisal

Condition	Δ_η	Statistic	p-value	Effect Size	Test
f_f/f_b	-3.1502	429.0	1.65e-29	0.0244	Wilcoxon
b_b/b_f	-4.0729	258.0	4.25e-42	0.008	Wilcoxon
f_b/b_f	-0.8602	14301.0	9.15e-13	-0.3978	Mann-Whitney
r_r/r_u	-1.0516	1107.0	1.8e-06	0.2236	Wilcoxon
u_u/u_r	-3.185	10.0	9.0e-58	0.0002	Wilcoxon
r_u/u_r	-1.5827	6339.0	2.48e-21	-0.6256	Mann-Whitney
f_h/dish	-3.6816	1348.0	4.71e-70	0.0138	Wilcoxon
r_h/dish	-2.7061	3473.0	4.5e-64	0.0356	Wilcoxon
f/r_dish	0.8555	132794.0	5.55e-21	0.3656	Mann-Whitney

Table 9: P-values, Δ_η and effect size for Manchu feature surprisal

Condition	Δ_η	Statistic	p-value	Effect Size	Test
f_f/f_b	-2.5563	6.0	1.68e-24	0.0006	Wilcoxon
b_b/b_f	-3.4993	209.0	1.16e-20	0.0253	Wilcoxon
n_f/n_b	0.354	14803.0	0.0086	0.4076	Wilcoxon
f_b/b_f	0.1359	9167.0	0.6778	0.0305	Mann-Whitney
f_n_f/f_n_b	-1.3331	43.0	3.58e-05	0.0814	Wilcoxon
b_n_b/b_n_f	-1.5021	259.0	1.61e-11	0.0743	Wilcoxon
n_n_f/n_n_b	0.1291	3941.0	0.7673	0.4849	Wilcoxon
f_n_b/b_n_f	-0.0086	1273.0	0.7338	-0.0414	Mann-Whitney

Table 10: P-values, Δ_η and effect size for Khalkha Mongolian feature surprisal

Condition	Δ_η	Statistic	p-value	Effect Size	Test
atr_atr/atr_natr	-1.8211	27.0	1.55e-13	0.0095	Wilcoxon
natr_natr/natr_atr	-0.6621	1819.0	2.55e-12	0.1672	Wilcoxon
n_atr/n_natr	-0.6531	91.0	0.0185	0.2407	Wilcoxon
atr_natr/natr_atr	-1.5526	7395.0	3.21e-05	0.3415	Mann-Whitney
r_r/r_u	-1.8211	2.0	4.37e-07	0.0034	Wilcoxon
u_u/u_r	-0.6621	2.0	8.35e-13	0.0009	Wilcoxon
n_r/n_u	-0.6531	371.0	0.148	0.3747	Wilcoxon
r_u/u_r	-1.5526	170.0	2.64e-12	-0.8529	Mann-Whitney
atr_h/dish	-1.0537	2337.5	1.09e-25	0.0944	Wilcoxon
r_h/dish	-1.6815	6.0	2.18e-18	0.0011	Wilcoxon
atr/r_dish	-0.3697	8941.0	0.0024	-0.2103	Mann-Whitney

D Vowel Counts in Test Set

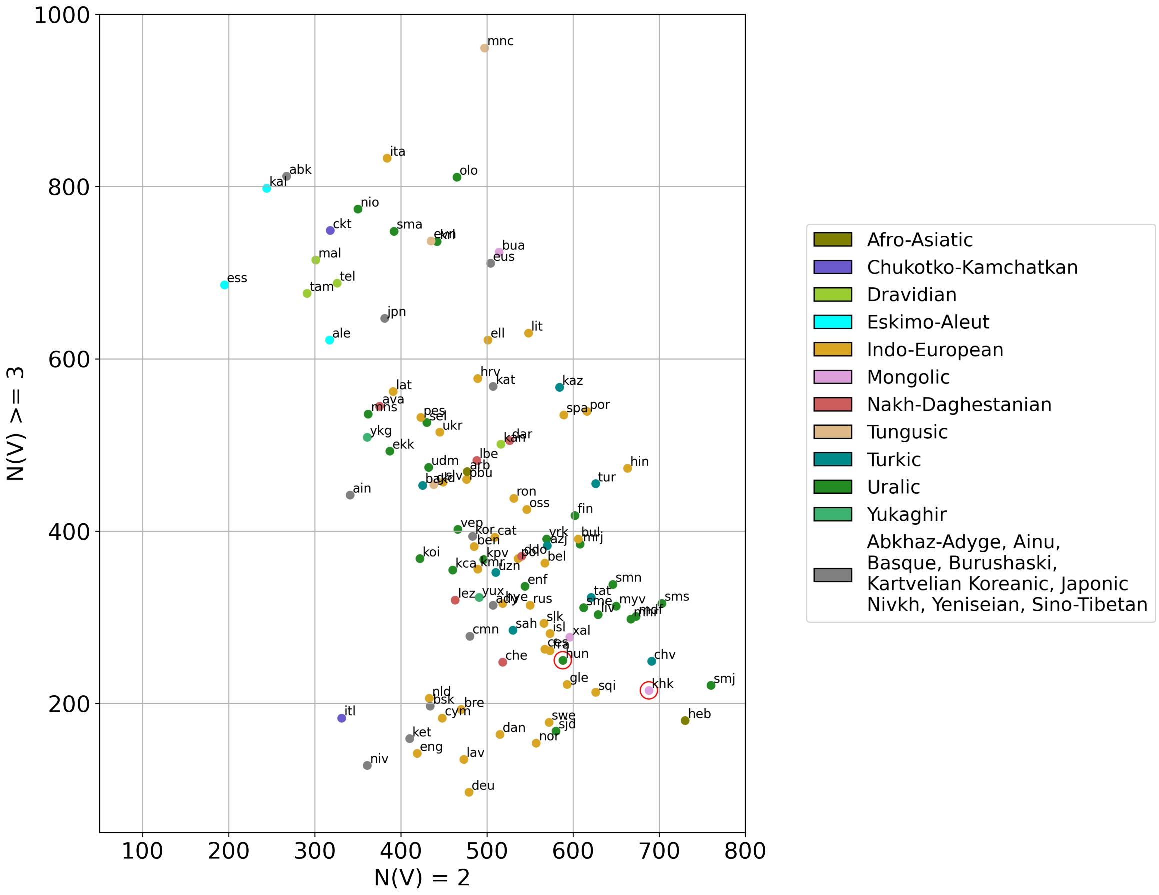

Figure 3: Number of items with 2 vowels (x-axis) and 3 or more vowels (y-axis) in all languages in NorthEuraLex. Hungarian and Khalkha Mongolian in red circles. Languages were coded for language family (see legend) and identified by ISO codes. For a mapping of ISO codes to language see the NorthEuraLex website http://www.northeuralex.org/languages.