The UCI Phonotactic Calculator: An online tool for computing phonotactic metrics

Connor Mayer (UCI)*, Arya Kondur (UCI), Megha Sundara (UCLA)

* cjmayer@uci.edu

Abstract

This paper presents the UCI Phonotactic Calculator (UCIPC), a new online tool for quantifying the occurrence of segments and segment sequences in a corpus. This tool has several advantages compared to existing tools: it allows users to supply their own training data, meaning it can be applied to any language for which a corpus is available; it computes a wider range of metrics than is typical; and it provides an accessible point-and-click interface that allows researchers with more modest technical backgrounds to take advantage of phonotactic models. After describing the metrics implemented by the calculator and how to use it, we present the results of a proof-of-concept study comparing how well different types of metrics implemented by the UCIPC predict human responses from eight published nonce word acceptability judgment studies across four different languages. These results suggest that metrics that do not take absolute position within the word into account are better at predicting human responses than those that do. We close by discussing the usefulness of tools like the UCIPC in experimental design and analysis and outline several areas of future research that this tool will help support.

1 Introduction

Phonotactics refers to restrictions on the sequencing of sounds into words. For example, although the word /skif/ “skeef” is not a real English word (at least in the authors’ dialects), it could in principle become an English word. It could not be a Spanish word, however, because Spanish has a general phonotactic restriction against /s/-initial complex onsets. The fact that different languages impose different restrictions on phonotactic patterns indicate that phonotactics must be learned from data (though some aspects like sonority sequencing preferences have been proposed to be innate; Selkirk 1984, Prince & Smolensky 1993; Berent et al. 2008; etc.). Speakers generally have strong intuitions about what possible words could look like in their language. The process of developing these intuitions is typically taken to correspond to forming generalizations over sound patterns in the lexicon.

One common method for getting insight into speakers' phonotactic knowledge is to perform acceptability judgment tasks where speakers are asked to rate a nonce word based on its suitability as a possible word in their language. This might involve a forced choice task, a numeric rating, or other responses like magnitude estimation. It has long been observed on the basis of such studies that phonotactic judgments are gradient: speakers do not generally think of words as being "in" or "out" but can often arrange them on a cline of acceptability. The classic example from Chomsky & Halle (1965) is the three words "blick", "bnick", and "bzick". Although none of these are real English words, speakers typically find "blick" to be acceptable, "bzick" to be unacceptable, and "bnick" to be somewhere in between. Every experimental study that has tested for gradience has found it (e.g. Chomsky & Halle 1965, Coleman & Pierrehumbert 1997, Scholes 1966, Bailey and Hahn 2001, Hayes and Wilson 2008, Daland et al. 2011, a.o.). Further, Daland (2015) makes a compelling case for how gradient knowledge of phonotactics is crucially important for speech perception (e.g., Norris & McQueen 2008, Davidson & Shaw 2012, Chodroff & Wilson 2014; Steffman & Sundara, 2023).

As researchers, we are interested in understanding what knowledge of phonotactics speakers possess, how they acquire that knowledge, and how it is deployed in other areas of language. A common approach is to develop models that allow phonotactic metrics to be computed for words given some training data sample meant to approximate a speaker's lexicon, over which phonotactic generalizations can be formed. These metrics are commonly used as predictors of experimental data such as reading time or acceptability scores. Such models are useful because they allow us to stipulate precisely (a) what phonotactic configurations speakers are sensitive to; (b) how exposure to these configurations shapes phonotactic knowledge; and (c) how phonotactics influences speech perception more generally. A key criterion for these models is that they output gradient, rather than categorical acceptability scores.

The goal of this paper is to present a new online tool used for calculating phonotactic metrics: the UCI Phonotactic Calculator (https://phonotactics.socsci.uci.edu/). This tool has several advantages compared to existing tools for computing phonotactic acceptability: it allows users to supply their own training data, meaning it can be run on any language where a corpus representing the lexicon is available; it computes a wider range of metrics than is typical; and it provides an accessible point-and-click interface that allows researchers with more modest technical backgrounds to take advantage of phonotactic models.

The structure of the paper is as follows: Section 2 describes several existing tools for computing phonotactic metrics and the limitations of these tools that motivated the development of the UCI Phonotactic Calculator. Section 3 describes the UCI Phonotactic Calculator and the metrics it implements. Section 4 shows how the UCI Phonotactic Calculator can be used to compare a variety of proposed phonotactic models against data from eight phonotactic acceptability studies across four languages. The results demonstrate that, in every case, metrics that reference the relative position of segments in words outperform more commonly used metrics that reference the absolute position of segments. This raises several other questions that would benefit from future research. Section 5 offers a brief discussion and conclusion.

2 Limitations of existing tools for computing phonotactic acceptability

In this section we will discuss existing tools that quantify the occurrence of segment or segment sequences in a corpus embodying a lexicon. Overall, such tools are available for a small number of languages and cannot be customized because the training corpora they are based on are not accessible to the user. Both these limitations severely restrict the range of questions that can be empirically investigated.

2.1 The Phonotactic Probability Calculator

Currently, the most well-known phonotactics calculator, with 641 citations in google scholar, is the Phonotactic Probability Calculator (PPC; Vitevitch and Luce 2004). The PPC allows the phonotactic metrics described in Jusczyk et al. (1994) and Vitevitch and Luce (1999) to be computed for novel words in English, Spanish, and Modern Standard Arabic (Aljasser & Vitevitch, 2018). A related tool, the Neighborhood Density Calculator, allows neighborhood density to be calculated for words in the same languages (Vitevich and Luce 2016).

The English model is trained on a phonetic transcription of the 1964 Merriam-Webster Pocket Dictionary, which consists of about 20,000 words. Information about word frequency comes from Kučera and Francis 1967). Stress is not encoded. The website does not state what training data was used for the Spanish model, though it may be the data from the Beginning Spanish Lexicon (Vitevitch et al. 2012), which consists of 3,854 words from the glossary of a first-year Spanish textbook, transcribed in Castilian Spanish. The Modern Standard Arabic model is trained on a list of the 100,000 most frequent Modern Standard Arabic lemmas purchased from https://www.sketchengine.eu/.

The metrics calculated by the PPC are *positional, frequency-weighted, unigram and bigram scores.* The mathematical implementation of this is described in detail in Section 3.1.4, but we provide some intuitive definitions of the properties of these metrics here:

- *Unigram/bigram:* Unigram metrics consider the frequency of occurrence of individual sounds. Bigram metrics consider the frequency of occurrence of adjacent pairs. Together these measures can capture how likely listeners are to hear individual sounds, as well as particular sequences of sounds.

- *Frequency-weighted:* words that occur more frequently in the training data contribute more than do less frequent words. In a model that is not frequency-weighted, all word types in the training set contribute equally, regardless of their frequency.
- *Positional:* the absolute position of the unigrams/bigrams in a word is considered when calculating metrics. This differs from standard unigram/bigram models, where absolute position is not considered (Markov 1913, Shannon 1948). For example, in a *positional* unigram model, the influence of a /t/ on the computed metric for a word can differ depending on whether it occurs in the first position, second position, etc. In *non-positional* unigram model, to be described below, the influence of /t/ on the metric will be the same regardless of where in the word it occurs. Analogously, in a positional bigram model the influence of a sequence /st/ on the computed metric can differ depending on whether this sequence occurs as the first and second sounds, the second and third, etc. In a non-positional bigram model, the influence of this sequence on the metric will be the same regardless of the positions it occurs in. In more rigorous mathematical terms, positional models are *non-stationary* while the non-positional models are *stationary*.

  Note that the positional metrics implemented in Vitevitch and Luce (2004) have two other important differences from standard implementations of non-positional metrics. First, the positional metrics do not explicitly reference word boundaries. This is not an issue for word-initial segments, since these are always in the first position, but it means these metrics cannot differentiate between segments that occur at the end of words and segments that don't. The non-positional models, on the other hand, insert word boundary symbols at the edges of words and reference them when computing the metrics (so, for example, a bigram sequence like /t#/, where # is a word boundary, refers to a /t/ in word-final position). Second, the positional bigram metrics are implemented as joint probabilities while the non-positional metrics are implemented as conditional probabilities. These issues will be discussed more in Section 3.

The name “Phonotactic Probability Calculator” is something of a misnomer. Values associated with word forms per the PPC do not constitute valid probabilities because the positional probabilities of individual unigrams and bigrams are combined by addition rather than multiplication. In addition, because training data is hardcoded into the PPC, it is available only for English, Spanish and Modern Standard Arabic, and the properties of the training data cannot be customized. Finally, the use of positional unigram and bigram metrics may lead to data sparsity issues for longer words. For example, the mean number of phonemes in an English word is about 5.77 (sd=1.93; Marian et al. 2012). Estimates for a unigram score corresponding to /t/ in the third position will be based on a large number of data points (since most English words have something in the third position), while estimates for a score corresponding to /t/ in the 10th position will be less reliable, as fewer words have ten or more segments. This is not an issue for non-positional models, because unigram and bigram values are calculated without taking absolute position into account.

2.2 Irvine Phonotactic Online Dictionary

A tool called the Irvine Phonotactic Online Dictionary (IPhOD; Vaden et al. 2009; http://www.iphod.com/) provides similar functionality to the PPC. IPhOD can compute a large range of phonotactic metrics, including both positional/non-positional and frequency-weighted/non-frequency-weighted probabilities, as well as neighborhood densities for English words provided by the user. In addition, it allows users to search for words that meet certain criteria with respect to these metrics (e.g. “find English words with fewer than three neighbors”).

There are two main limitations of the IPhOD calculator. The first is that the training dataset is limited to English, and more specifically to the approximately 54,000 words in the CMU English Pronouncing Dictionary (Weide 1994). While this database is quite extensive, words that do not exist in it cannot be used as part of the calculation process. This also limits users' ability to provide their own training dataset, which may be more practical for certain research purposes. Second, the overall usage of the IPhOD calculator is limited as it only supports a few phonotactic metrics and users must enter their testing dataset manually rather than through a file upload. These are minor issues that we aim to address with the UCI Phonotactic Calculator.

2.3 CLEARPOND

CLEARPOND is a tool maintained by Northwestern University that computes orthographic and phonological neighborhood density and other metrics like word length and frequency (Marian et al. 2012; https://clearpond.northwestern.edu/). CLEARPOND supports five languages, English, Dutch, French, German, and Spanish, and also allows cross-language neighborhood densities to be computed. CLEARPOND also supports the calculation of positional, frequency-weighted biphone and bigram scores using the same method described in Vitevitch and Luce (2004). Unlike the other programs described here, CLEARPOND has experimental functionality that computes neighborhood density and neighbors for a list of training words given a custom training data set. This functionality is limited in that it only computes a subset of the neighborhood density metrics and no phonotactic metrics. Thus, CLEARPOND is similar to the PPC but with support for a wider range of languages.

2.4 The UCLA Phonotactic Learner

The UCLA Phonotactic Learner (Hayes & Wilson 2008; https://linguistics.ucla.edu/people/hayes/Phonotactics/) is a program for calculating phonotactic probabilities. It is a maximum entropy model (Goldwater & Johnson 2003) that penalizes words that violate certain *featural n-gram constraints.* Features refer to properties of sounds like voicing, sonority, manner of articulation, etc. Examples of such constraints might be "don't have a voiced sound following a voiceless sound". The UCLA Phonotactic Learner induces from a training set both what constraints are necessary and how strongly they should be weighted.

An advantage of referring to features rather than segments is that it can capture variability in the acceptability of unattested sequences. For example, even though both 'bnick' and 'bzick' are poorly formed with respect to English phonotactics, speakers often have an intuition that the first is not as bad as the second (Chomsky & Halle 1965). Models that refer to segments alone, like all the models discussed above, cannot capture this distinction, since both /bn/ and /bz/ are unattested. Featural models, however, can capture the idea that because there are more onsets like /bn/ (e.g., onsets with a /b/ followed by non-nasal coronal sonorant, like /bl/, or an obstruent followed by a coronal nasal, like /sn/) than there are like /bz/, speakers should find the former more acceptable.

The UCLA Phonotactic Learner has been an enormously influential model of phonotactic learning, and its performance compares favorably to other models (e.g. Daland et al. 2011). Some limitations are that it is a standalone executable and cannot integrate directly into programming workflows, and that it does not output probabilities directly but rather numeric weights that correlate with them. The model has the additional task relative to the other models of discovering the constraints: there are several hyperparameters that govern how this process takes place that the model is sensitive to.

2.5 Other tools

There are also several databases that contain phonotactic or neighborhood density metrics for individual languages. EsPal allows neighborhood densities and other metrics to be calculated for both Latin American and Castilian Spanish based on both written and spoken corpora, though it does not support any phonotactic metrics (Duchon et al. 2013). It also allows Spanish words to be selected based on these properties. Diphones-fr is a simple database that contains diphone frequency information from over 50 million French words (New & Spinelli 2013).

The remaining existing tools are mainly used for word selection or generation under restrictions on phonotactic probability or neighborhood density. In this sense, they do not provide the same functionality as many of the tools discussed above, but still serve an important purpose in designing experimental stimuli. WordGen is one such example in which users can specify linguistic constraints to generate nonce words in Dutch, English, French, or German (Duyck et al. 2004). The main downside to WordGen is that it is a standalone Windows program and cannot easily be integrated into a broader programming workflow. Wuggy is a similar tool for word generation that takes the utility of WordGen a step further. It supports more languages, including Spanish and Vietnamese, and has a Python library (Keuleers & Brysbaert 2010).

3 The UCI Phonotactic Calculator

The UCI Phonotactic Calculator (henceforth UCIPC; https://phonotactics.socsci.uci.edu/) is an online tool we have developed that can be used to calculate various metrics to quantify phonotactic information. This tool has several primary differences from the existing tools discussed above.

1. It allows the user to specify their own training data set. To our knowledge, this is the only online tool supporting this functionality for phonotactic metrics (note that CLEARPOND does support

this, but for neighborhood density alone). This allows the UCIPC to be deployed on any language or input register (e.g. infant-directed speech).

2. It supports a wider range of metrics than most existing tools.
3. It can be run both via an online interface and via the command line, which allows it to be integrated into larger programming workflows.
4. It is open source: https://github.com/connormayer/uci_phonotactic_calculator

3.1 Currently supported phonotactic metrics

The UCIPC supports a range of phonotactic metrics that differ in how or whether they encode context, token frequency, and position. This section will provide a qualitative and quantitative description of each metric. The set of metrics is roughly divided into four classes based on the following two factors:

1. Unigram vs. Bigram metrics: do we consider the preceding context in which a sound occurs (bigram) or not (unigram)?
2. Positional vs. non-positional metrics: do we consider the absolute position in the word in which a unigram or bigram occurs or not? The positional metrics correspond to the calculations done in Jusczyk et al. (1994) and Vitevitch and Luce (2004), while the non-positonal metrics correspond to more standard implementations of n-gram models (Markov 1913, Shannon 1948).

The UCIPC also includes metrics produced by a neural network model of phonotactics (Mayer and Nelson, 2020). This differs substantially from the other metrics and is presented in its own section.

3.1.1 Non-positional unigram probabilities

In all of the sections below, we will use $w = x_1 \dots x_n$ to refer to a word consisting of symbols x_1 through x_n.

The unigram score reflects the probability of a word under a simple unigram model. Here, the probability of a word is defined as the product of the probabilities of its individual symbols. In the non-positional version of unigram metrics, probabilities are calculated without considering the position of symbols. Rather, only the frequencies of symbols are used in the calculation (Markov 1913, Shannon 1948). If a particular symbol exists in the test dataset but not in the training set, its probability is set to zero. Mathematically, we express the standard unigram probability as

$$P(w = x_1 \dots x_n) \approx \prod_{i=1}^{n} P(x_i)$$

where we express the probability of encountering an individual unigram as

$$P(x) = \frac{C(x)}{\sum_{y \in \Sigma} C(y)}$$

where $C(x)$ represents the number of occurrences of x in the training data and Σ is the set of all sounds in the training data. The UCIPC returns log unigram probabilities.

3.1.2 Positional unigram scores

The UCIPC also computes positional unigram scores. These differ from the unigram probabilities above in that they are sensitive to the absolute position of each segment in a word, as well as its identity. The positional-unigram score is a type-weighted variant of the unigram score from Vitevitch and Luce (2004). It is defined as follows:

$$PosUniScore(w = x_1 \ldots x_n) = 1 + \sum_{i=1}^{n} P(w_i = x_i)$$

where

$$P(w_i = x) = \frac{C(w_i = x)}{\sum_{y \in \Sigma} C(w_i = y)}$$

where refers to the i^{th} position in a word and is the number of times in the training data the symbol x occurs in the i^{th} position of a word.

Vitevich and Luce (2004) add 1 to the sum of the unigram probabilities "to aid in locating these values when you cut and paste the output [...] to another program." They recommend subtracting 1 from these values before reporting them.

Under this metric, the score assigned to a word is based on the sum of the probability of its individual symbols occurring at their respective positions. Note that the ordering of the symbols with respect to one another does not affect the score, only their relative frequencies within their given positions. For example, for a sequence /xy/, the fact that /x/ immediately precedes /y/ is not explicitly represented in the model; only that /x/ occurs in position 1 and /y/ occurs in position 2. Higher scores represent words with more probable phonotactics, but note that this score cannot be interpreted as a probability because the individual probabilities are summed together rather than multiplied.

3.1.3 Non-positional bigram probabilities

Unlike unigram metrics, the bigram metrics consider pairs of symbols instead of individual symbols. Calculating the standard bigram score is done in a similar way to that of unigram scores. That is, it is represented as the product of probabilities of consecutive symbols in each word conditioned on the previous symbol. Like the unigram model, the absolute position of the bigrams in the word is not considered, and bigrams not occurring in the training data are assigned a probability of zero. We pad the words with a special symbol at the beginning and end, which allows us to compute the probabilities of symbols starting and ending words. For example, the input /kæt/ ‘cat’ would consist of the bigrams {#k, kæ, æt, t#}, where # is a word boundary symbol. This allows the model to be sensitive to the frequencies with which certain segments begin and end words.

Mathematically, we express the standard bigram probability as

$$P(w = x_1 \dots x_n) \approx \prod_{i=2}^{n} P(x_i | x_{i-1})$$

where the probability of a particular bigram is calculated as

$$P(x|y) = \frac{C(yx)}{C(y)}$$

where the count function $C(\cdot)$ is defined as in Section 3.1.1. The UCIPC returns log bigram probabilities.

3.1.4 Positional bigram metrics

The UCIPC also computes positional bigram scores. This is a type-weighted variant of the bigram score from Vitevitch and Luce (2004). It is defined as:

$$PosBiScore(w = x_1 \dots x_n) \;=\; 1 \;+\; \sum_{i=2}^{n} P(w_{i-1} = x_{i-1},\, w_i = x_i)$$

where

$$P(w_{i-1} = x_{i-1},\, w_i = x_i) \;=\; \frac{C(w_{i-1} = x_{i-1},\, w_i = x_i)}{\sum_{z\, \in \Sigma} \sum_{v\, \in \Sigma} C(w_{i-1} = z,\, w_i = v)}$$

The same caveats apply here with respect to these scores not forming valid probabilities. These scores also differ from non-positional bigrams in that the bigram probabilities used to calculate the overall scores are *joint* probabilities rather than conditional probabilities: they tell us the probability of segment x occurring in position i *and* segment y occurring in position $i + 1$, while the non-positional bigram probabilities tell us the probability of segment y occurring in position $i + 1$ *given that* segment x occurred in position i.

3.1.5 Token frequency-weighted metrics

In the standard metrics, which we call type-weighted, the frequency of individual word types does not affect the output scores. That is, word types that occur more frequently are weighted the same as word types that occur very few times. The token-weighted variants of each metric do account for frequency of word types. Specifically, word types that occur frequently are weighted higher than less frequent word types. To account for this weighting, the count function is weighted so that each occurrence of a particular configuration is weighted by the natural log of the count of the word it occurs in.

For example, consider a corpus that contains the word type "kæt" 1000 times and "tæk" 50 times. Under a token-weighted unigram model, we would have $C(\text{æ}) = \ln(1000) + \ln(50) \approx 10.82$, whereas in a type-weighted unigram model, we would have $C(\text{æ}) = 1 + 1 = 2$.

Note that the token-weighted positional unigram and bigram scores correspond to the metrics in Vitevitch and Luce (2004). The only difference is that they use logarithms with base 10 and the UCIPC uses natural logarithms.

3.1.6 Smoothed metrics

The calculator also calculates variants of every metric with add-one smoothing. With this type of smoothing, every configuration (unigram, bigram, etc.) has a default count of one. Thus, configurations that are not encountered in the training set, but do appear in the test set, are treated as if they occurred once in the former rather than not at all. This assigns a small probability to such configurations. Without smoothing, the default count is zero and the associated probability is also zero. Mathematically, for a configuration x that originally has $C(x) = 0$, under add-one smoothing, the count function starts with $C(x) = 1$. Performing smoothing in the token-weighted versions of metrics is done by simply adding one to the log-weighted counts. Smoothed metrics provide a way to distinguish sequences with unattested segments or segment sequences, which would otherwise be assigned probabilities of 0.

3.1.7 An RNN model of phonotactics

In addition to these phonotactic metrics, the UCPIC also computes scores based on a simple recurrent neural network (sRNN) described in Mayer and Nelson (2020). Rather than considering fixed context windows of one or two phonemes, as is done with unigram and bigram metrics, the sRNN can in theory learn phonotactic constraints across arbitrarily large distances. In the sRNN architecture, any given state is solely dependent on the input and the state of the immediately preceding timepoint. A visual representation of this characteristic is given in Figure 1.

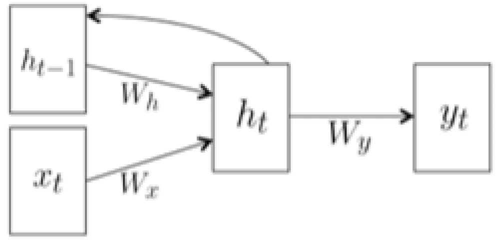

Figure 1: sRNN architecture from Mayer and Nelson (2020)

The state of the network at a given timepoint t is given by

$$h_t = \tanh(W_x x_t + W_h h_{t-1} + b_h)$$

where x_t is the vector corresponding to the phoneme input at time t , W_t and W_h are weighted matrices for the input and previous state vectors, and b_h is a fixed bias vector. The neural network then produces a probability distribution that serves as a prediction as to what segment(s) will appear in the next timepoint. Using a weight matrix W_y and the softmax function $\sigma(z)$, the probability distribution is calculated as

$$\hat{y}_t = \sigma\left(W_y h_t\right).$$

During each iteration, the weighted matrices and bias vector are updated based on the accumulation and backpropagation of the cross-entropy loss between the predicted probability distribution before encountering a particular phoneme and that phoneme's identity y . The loss function is defined as

$$L(y, \hat{y}) = -y \cdot \log(\hat{y}).$$

For a more in-depth explanation of the implementation of the sRNN, relevant hyperparameter choices, and an evaluation of the model's performance, see the original paper (Mayer and Nelson 2020).

3.1.8 Summary of UCIPC Metrics

To summarize, the UCIPC can compute the following metrics given training and test data:

1. Unigram probability
2. Smoothed unigram probability
3. Bigram probability

4. Smoothed bigram probability
5. Frequency-weighted unigram probability
6. Frequency-weighted bigram probability
7. Smoothed, frequency-weighted bigram probability
8. Positional unigram score
9. Positional bigram score
10. Smoothed positional unigram score
11. Smoothed positional bigram score
12. Frequency-weighted positional unigram score
13. Frequency-weighted positional bigram score
14. Smoothed, frequency-weighted positional unigram score
15. Smoothed, frequency-weighted positional bigram score
16. RNN score

Note that all metrics except the positional variants and the RNN score are reported as log probabilities.

3.2 A brief tutorial for the UCIPC

The UCIPC requires three inputs: a properly formatted training file, a test file, and the type of training model to use. For the training set, users have the choice of uploading their own file or selecting from the several existing datasets readily available to the UCIPC. To choose an existing dataset, users may use the dropdown menu which contains a short description of the available datasets. For a more detailed description of each file, users should view the Datasets page, which is dedicated to storing and explaining the use case of each dataset. Existing datasets include, English, Spanish, Turkish and Polish corpora (referenced below in this paper) as well as the Finnish, French and Samoan datasets used in Mayer (2020).

If uploading a personal training file, users must take care to follow a few specifications:

- The file must be in .csv format.
- The file must consist of one or two columns without headers.

In user-uploaded training files, the first column is mandatory and should consist of a word list where symbols are separated by spaces. Any transcription system is valid, so long as individual symbols are space-separated. The second column is optional and, if included, should contain the corresponding frequencies for each word, expressed as counts. If this column is included in the training file, both the type- and token-weighted variants of each metric will be computed. Otherwise, just the type-weighted metrics will have values in the output file. Note that users may not both upload their own training file and select a default training file; the UCIPC will display an error message requesting a single choice to be made.

The test file needs to be a CSV file with a single column of test words and no headers. The transcription system used in the test file should match the system used in the training file. The mechanism for uploading the test file is the same as uploading the training file.

Finally, users should select the models to train. Selecting the simple model (labeled "Unigram/Bigram Scores") will result in unigram and bigram scores being calculated, along with the positional, smoothed, and token-weighted variants. The "RNN model" runs the recurrent neural network model proposed in Mayer and Nelson (2020). This model requires significantly more time to train and run than the simple model which is why the two appear as separate options instead of being run together. Note that the RNN model is run with default hyperparameters. Users who wish to modify the hyperparameters should use the command line interface provided in the source code:

https://github.com/MaxAndrewNelson/Phonotactic_LM

Once users submit their training file, test file, and model type, the UCIPC will direct them to a separate page to download the output file. The output file will be formatted differently depending on the model type selected, but the overall structure will be the same. If the simple model is selected, users will receive a CSV file where each row contains the test word, its length, and all the calculated variations of the unigram and bigram metrics. If the RNN model is selected, users will receive a CSV file where each row contains the test word, its length, and the calculated perplexity score. Lower perplexities correspond to higher probabilities. To run the model again, users must go back to the UCIPC home page and resubmit the input form with the necessary fields (training/test file, model type).

3.3 Planned extensions of the UCIPC

Currently the UCIPC does not implement calculation of neighborhood density. This is largely due to the contexts in which it has been applied so far: we have focused primarily on the phonotactic acquisition in the first year, and previous research has indicated that infants are not sensitive to neighborhood density in this time period (Swingley & Aslin, 2002; Sundara et al. 2022). To make the UCIPC more applicable to the study of adult phonotactic knowledge, we plan to implement this functionality soon.

We also plan to add more sophisticated smoothing techniques. Currently all smoothed metrics involve add-one, or Laplace, smoothing. This technique has the virtue of being simple, but it tends to shift too much probability mass from observed to unobserved word forms. We plan to add additional smoothing techniques, such as Modified Kneser-Ney or Witten-Bell smoothing, which have been shown to perform more favorably in NLP tasks (e.g., Chen & Goodman, 1998). To our knowledge no work has looked at smoothing as it relates to modeling phonotactic acceptability judgments. More detailed study of how well different smoothing techniques correlate with empirical observations in this domain will be valuable.

Finally, we would like to emphasize that the UCIPC is an open-source project (the source code can be found at https://github.com/connormayer/uci_phonotactic_calculator). If you are interested in adding new functionality or fixing bugs, please reach out to the corresponding author.

4 Applications of phonotactic metrics in experiments with adults

In this section, we model phonotactic acceptability ratings given by human participants in published studies as a function of a variety of phonotactic metrics. Our purpose here was to determine whether positional metrics that encode absolute positional information or non-positional metrics that take word edges but no other positional information into account that are best able to predict acceptability judgements by adult native listeners. For each of the following datasets, we run the UCIPC with an appropriate training set in the same language. The calculator's outcomes are used as predictors in a regression model that attempts to predict the human ratings. All the code and data can be found at https://github.com/aryarksub/phonotactic_metrics.

The general format for each regression model is

$$Acceptability \sim UnigramScore * BigramScore$$

with random intercepts included for individual participants and items when possible. The unigram and bigram score predictors are mean-centered and scaled. Outputs from the UCIPC that are negative infinity (corresponding to a probability of zero) are adjusted to a large negative value (e.g. -50) so that scaling can be done without error. The specific type of regression used (linear or logistic) depends on the experimental design of each study.

We consider a maximum of eight models for each data set:

1. Standard: Non-positional unigram and bigram probability
2. Frequency-weighted: Token-weighted non-positional unigram and bigram probability
3. Smoothed: Smoothed non-positional unigram and bigram probability
4. Frequency-weighted, smoothed: smoothed, token-weighted non-positional unigram and bigram probability
5. Positional: Positional unigram and bigram scores
6. Frequency-weighted positional: Token-weighted positional unigram and bigram scores
7. Smoothed positional: Smoothed positional unigram and bigram scores
8. Frequency-weighted, smoothed positional: Smoothed, token-weighted positional unigram and bigram scores

Token-weighted metrics are omitted when frequency information is not available for the training data. We do not report results for the RNN model.

Each of the following subsections examines an individual test dataset and reports the results of the relevant models as run on the corresponding data. The results are formatted in a tabular manner, with the following column headers:

- Model: Specifies the metrics used as predictors in the model
- Intercept: Regression intercept
- Uni. Coef: Coefficient for the unigram score term
- Bi. Coef: Coefficient for the bigram score term
- Int. Coef: Coefficient for the interaction (between unigram and bigram score) term
- AIC: Akaike Information Criterion (estimation of prediction error; Akaike 1974)

AIC is a metric for model comparison that estimates out of sample prediction error. It rewards model fit to the data and penalizes model complexity. Lower values of AIC indicate better model performance. However, absolute AIC values are not meaningful, but differences in AIC between a model and the model with the lowest AIC can be used to evaluate their performance on a dataset. We interpret differences in AIC using the rule of thumb proposed in Burnham & Anderson (2004; p. 271): an AIC difference of <= 2 between a model M and the model with the lowest AIC score M_{min} means there is "considerable support" for M (i.e., M and M_{min} are both plausibly the best model); a difference of between 4 and 7 means M has "considerably less support" relative to M_{min}, and a difference of more than 10 indicates "essentially no support" for M relative to M_{min}.

Each table below is sorted in increasing order of AIC, with models at the top of each table having lower AIC, and therefore, better overall performance.

4.1 English

Unsurprisingly, the most numerous reports are on native English speakers phonotactic judgments. In this section, we report results from modeling data obtained from 5 published studies.

4.1.1 Albright and Hayes (2003)

The data used in Albright and Hayes (2003) consists of 58 English monosyllabic nonce verbs consisting of between 3 and 5 segments rated for phonological well-formedness by 20 native English speakers. These data correspond to the pretest portion of Experiment 1 in their paper. Participants were asked to rate forms on a Likert scale between 1 (impossible as an English word) and 7 (would be a fine English word). The data is in the form of mean ratings across participants; individual ratings are not available.

The phonotactic models were trained on the English CMU Pronouncing Dictionary (CMU Pronouncing Dictionary 2008) with frequency information from CELEX (Baayen et al. 1995). Stress location was not represented in the training data. The fitted models provided scores for the 58 nonce verbs used in the study. We used these scores as predictors in a set of linear models to model the mean rating. Because the dataset does not contain individual ratings by subject, we do not use any random effects.

The results in Table 1 show that the non-positional metrics outperform their positional counterparts in each case, and that these differences indicate considerably less support for the positional models. The differences within the classes of non-positional and positional metrics are not meaningful.

Model	Intercept	Uni. Coef.	Bi. Coef.	Int. Coef.	AIC
Smoothed	4.69160	0.15713	0.11632	-0.01857	123.115
Frequency-weighted smoothed	4.68971	0.16792	0.10045	-0.01575	123.437
Standard	4.70083	0.22337	0.07286	-0.14096	123.965
Frequency-weighted	4.69640	0.22241	0.05458	-0.11392	124.152
Positional frequency-weighted	4.72162	-0.11934	0.13720	-0.05067	129.562
Positional frequency-weighted smoothed	4.72111	-0.11911	0.13690	-0.05002	129.566
Positional	4.70156	-0.10987	0.12717	-0.02460	129.706
Positional smoothed	4.70003	-0.10882	0.12585	-0.02265	129.714

Table 1: Coefficients and AIC scores of the regression models fit to data from Albright & Hayes (2003).

4.1.2 Daland et al. (2011)

The test data obtained from Daland consists of 96 disyllabic English nonce words consisting of six segments. These words were rated on a five-point Likert rating scale and aggregated over a set of subjects. The main goal of Daland et al. (2011) was to compare the acceptability of different onsets in English. These nonce words accordingly consist of a set of 48 complex onsets and four tails. Each onset was paired with two different tails. Models were fit to the same English training data set described in the previous section, and the 96 nonce words (including tails) were scored by the fitted models. We used these scores as predictors in a linear regression model that uses the mean word ratings across participants as its output feature. Because scores were aggregated across subjects, there were no random effects in the model.

The results are shown in Table 2. The most accurate models are the smoothed and frequency-weighted smoothed non-positional metrics. Although in this case all of the positional models outperform the frequency-weighted and standard non-positional models, a difference of > 10 in AIC between the best non-positional model and the best positional model indicates very little support for the latter.

Model	Intercept	Uni. Coef.	Bi. Coef.	Int. Coef.	AIC
Smoothed	2.62606	-0.08283	0.67846	0.29493	242.270
Frequency-weighted smoothed	2.61921	-0.08328	0.68104	0.31075	244.621
Positional frequency-weighted smoothed	2.75135	-0.06492	0.66884	-0.04975	258.460
Positional frequency-weighted	2.74997	-0.06262	0.66403	-0.04779	258.719
Positional smoothed	2.73739	-0.03431	0.62606	-0.03083	259.450
Positional	2.73305	-0.02656	0.61117	-0.02467	260.176
Frequency-weighted	2.62998	-0.05006	0.42167	0.21011	284.260
Standard	10.93989	0.34735	0.32775	0.01387	284.538

Table 2: Coefficients and AIC scores of the regression models fit to data from Daland et al. (2011).

4.1.3 Needle and Pierrehumbert (2022)

The data from Needle et al. (2022) consists of ratings of 8400 English nonce words by 1440 participants. Nonce words consisted of 4-7 segments. Each participant rated 140 stimuli each, leading to 24 ratings for each individual nonce word. Ratings were provided on a five-point Likert scale.

In this case, the training dataset we use is the same one used in Needle et al. (2022) and consists of about 11,000 monomorphemic words from CELEX (Baayen et al. 1995) in the DISC transcription system. We converted the DISC transcriptions to ARPABET to stay consistent with the system used for English throughout this paper. Because this training data does not contain frequency information, the token-weighted models could not be used. The other models were fitted to this training data and used to score the experimental stimuli.

These scores were used as predictors in a linear mixed-effects model, with random intercepts for word and participant. The coefficients and AIC values of the fitted models are shown in Table 3. We see again that the non-positional models substantially outperform the positional ones. In addition, the non-smoothed non-positional model outperforms the smoothed one.

Model	Intercept	Uni. Coef.	Bi. Coef.	Int. Coef.	AIC
Standard	2.72212	0.05903	0.54662	-0.00847	566148.8
Smoothed	2.66202	-0.32281	0.69883	0.08722	566288.5
Positional smoothed	2.79281	-0.16563	0.33620	-0.10043	570030.7
Positional	2.79435	-0.15024	0.32053	-0.10272	570084.3

Table 3: Coefficients and AIC scores of the regression models fit to data from Needle et al. (2022).

4.1.4 Scholes (1966)

The test data from Scholes (1966) were obtained from the supplementary material of Hayes & Wilson (2008). It consists of 62 monosyllabic nonce words rated by 33 seventh grade students. These words varied primarily in their onsets. The students were asked whether each word was a possible word of English and asked to provide a yes/no response: thus, the data here consist of binary responses rather than Likert scores. The data were aggregated across onset, which means each of the 62 onsets is associated with a value between 0 and 1 that represents the proportion of "yes" responses across participants.

The training data used were also from the supplementary materials of Hayes & Wilson (2008) and consists of 55 English onsets and their type frequencies. This is a subset of the onsets in the CMU Pronouncing Dictionary with "exotic" onsets like [zw] and [sf] removed. This is rather different from the training data sets in previous cases because our training data consists of onsets, rather than words, and our frequency counts correspond to the number of word types each onset occurs in. The models were trained on this dataset and tested on the 62 words from Scholes (1966); this testing data also comes from the supplementary material for Hayes & Wilson (2008). The model scores were used as predictors in a logistic regression model over the proportions, weighted by the number of participants. Because we do not have individual ratings, we do not include any random effects.

The results in Table 4 show again that the non-positional metrics perform better than the positional metrics, although the overall differences between models are relatively minor. The frequency-weighted models have a small advantage over equivalent non-frequency-weighted models. We will discuss this phenomenon more in Section 4.2 below when we look at Polish onsets, where the effect is much stronger.

Model	Intercept	Uni. Coef.	Bi. Coef.	Int. Coef.	AIC
Frequency-weighted smoothed	-0.58485	0.02450	1.93349	0.20597	35.40623
Smoothed	-0.31853	-1.12792	2.83191	-0.24395	36.03306
Frequency-weighted	-0.20491	0.55417	1.60020	-0.30457	36.53359
Standard	-0.16237	0.51992	1.64383	-0.36761	36.65767
Positional frequency-weighted smoothed	-0.11198	0.70349	1.65370	-0.48528	37.56650
Positional frequency-weighted	-0.25016	0.84140	1.38754	-0.21887	38.09191
Positional	0.70141	0.06254	3.19127	-1.93961	39.76660
Positional smoothed	-0.00595	0.67075	2.32122	-1.25656	41.47684

Table 4: Coefficients and AIC scores of the regression models fit to data from Scholes (1966).

4.1.5 Hayes and White (2013)

The test data procured by Hayes and White (2013) consists of 160 English nonce words consisting of between 2 and 7 segments rated on a logarithmic scale by 29 participants. Participants were asked to perform a magnitude estimation task (Lodge 1981, Bard et al. 1996) comparing the well-formedness of each word with the reference word “poik”. The log of these magnitudes is the dependent variable we use here. The training data is the same CMU Pronouncing Dictionary data used in the analyses of Albright & Hayes (2003) and Daland et al. (2011). Models were trained on this data and used to score each nonce word. These scores were used as predictors in a linear mixed-effects model with random intercepts for participant and word.

The coefficients and AIC values of the fitted models are shown in Table 5. Again, the non-positional metrics tend to do better than positional metrics, with the smoothed frequency-weighted model performing the best.

Model	Intercept	Uni. Coef.	Bi. Coef.	Int. Coef.	AIC
Frequency-weighted smoothed	4.40106	-0.35521	0.52082	0.02889	12338.82
Smoothed	4.39708	-0.40281	0.55471	0.03242	12349.81
Standard	4.41836	-0.29086	0.44213	0.00310	12507.21
Frequency-weighted	4.41401	-0.28490	0.43514	0.01021	12519.93
Positional frequency-weighted	4.42823	-0.02101	0.18375	-0.00925	13009.03
Positional frequency-weighted smoothed	4.42809	-0.02090	0.18315	-0.00907	13009.36
Positional	4.43072	-0.03758	0.19762	-0.01205	13013.83
Positional smoothed	4.43027	-0.03711	0.19561	-0.01148	13014.93

Table 5: Coefficients and AIC scores of the regression models fit to data Hayes and White (2013).

4.2 Polish (Jarosz & Riesling 2017)

The Polish test data we use comes from Jarosz and Rysling (2017). In this paper, 81 native Polish speakers were asked to rate 159 test words consisting of 53 onsets and 3 tails (similar to the design in Daland et al. 2011) on a Likert scale of 1-5. Each participant rated each word once, leading to 12,880 responses.

Our training data set consisted of the list of Polish onsets with accompanying type frequencies from Jarosz (2017). These are generated from a corpus of child-directed speech consisting of about 43,000 word types (Haman et al. 2011). Because we trained only on onsets, we generated model predictions for the 53 onsets in isolation (meaning that the three tails corresponding to each onset receive the same score). The model scores are used as predictors in a linear mixed-effects model with random intercepts for word (including tail) and participant.

Table 6 shows again that the best models are non-positional, smoothed metrics. Similar to the Scholes data presented in Section 4.1.4, but more pronounced, frequency-weighting appears to be crucial for model performance. This may reflect some language-specific sensitivity to frequency. However, as with the Scholes data, the training data consists of onsets with type frequencies rather than words with token frequencies. When a non-frequency-weighted model is applied to this data, the training data consists of a simple list of attested onsets lacking both type and token frequency information. It is more likely therefore than the success of the frequency-weighted models here corresponds to a sensitivity to type frequency information.

Model	Intercept	Uni. Coef.	Bi. Coef.	Int. Coef.	AIC
Frequency-weighted smoothed	3.08772	0.00761	0.72491	0.07526	44609.70
Smoothed	3.09279	-0.02533	0.68918	0.06116	44799.76
Positional frequency-weighted smoothed	3.22977	0.30610	0.58109	-0.19084	44835.34
Positional frequency-weighted	3.22888	0.30468	0.58098	-0.18967	44836.69
Frequency-weighted	3.05181	0.05792	0.63124	0.18117	44849.67
Standard	3.05091	-0.03312	0.67438	0.15339	44883.97
Positional smoothed	3.14070	0.42246	0.34818	-0.05221	44907.11
Positional	3.14046	0.42175	0.34839	-0.05175	44908.04

Table 6: Coefficients and AIC scores of the regression models fit to data from Jarosz & Riesling (2017).

4.3 Spanish

This data set was collected by authors CM and MS using the methodology from Sundara & Breiss (submitted) for use in an unrelated study that is still in progress. The data consists of 576 unique CVCV Spanish nonce words rated on a discrete scale from 1 to 100 by 168 participants. Each participant rated 144 tokens, leading to 24,192 ratings. The phonotactic models were trained on a set of about 27,000 word types including citation and inflected forms taken from the EsPal database (Duchon et al. 2013) with stress encoded. The frequencies associated with these words were calculated from a large collection of Spanish subtitle data. The trained models were used to score the 576 nonce words.

We use these scores as predictors in a linear mixed-effects model with random intercepts for participants and words. Random intercepts are used for individual words and subjects. The results in Table 7 below show again that the non-positional metrics substantially outperform the positional metrics, and that the non-frequency-weighted metrics are generally more successful.

Model	Intercept	Uni. Coef.	Bi. Coef.	Int. Coef.	AIC
Smoothed	51.07835	-1.03073	8.11025	1.32290	187729.1
Standard	50.83292	-0.97408	7.08787	1.68646	187932.9
Frequency-weighted	50.82480	-1.02140	7.11876	1.72649	188059.9
Frequency-weighted smoothed	51.03021	-1.15668	8.26838	1.45804	188059.9
Positional smoothed	52.95626	-2.64322	6.81189	-2.55959	188252.1
Positional	52.95591	-2.64340	6.81094	-2.55890	189100.6
Positional frequency-weighted	52.99178	-2.25829	6.75389	-2.48905	189668.1

Positional frequency-weighted smoothed	52.99200	-2.25728	6.75381	-2.48962	189668.3

Table 7: Coefficients and AIC scores of the regression models fit to the Spanish dataset.

4.4 Turkish

The test data, described in more detail in Mayer (2024), consists of 596 Turkish CVCVC nonce words rated on a discrete scale from 1 to 100 by 90 subjects following the same methodology as the Spanish study above. Each participant rated 192 tokens, leading to 17,280 ratings. The phonotactic models were trained on a set of 18,472 citation forms from the Turkish Electronic Living Lexicon database (TELL; Inkelas et al. 2000). This training data does not contain frequency information, so we omit results from the frequency-weighted models. Fitted models were used to generate scores for the 596 nonce words.

These scores were used as predictors in a linear mixed-effects model with random intercepts for word and participant.

Again, the results in Table 8 show that non-positional metrics again do a better job of capturing human ratings than their positional counterparts, and the best model appears to be the standard smoothed model.

Model	Intercept	Uni. Coef.	Bi. Coef.	Int. Coef.	AIC
Smoothed	39.42271	6.56583	2.46451	6.26266	159545.6
Standard	39.16679	6.88337	1.84632	7.46382	159581.9
Positional	45.03984	-0.33506	11.17251	-1.29134	159628.4
Positional smoothed	45.03317	-0.31176	11.13564	-1.28345	159628.8

Table 8: Coefficients and AIC scores of the regression models fit to Turkish data from Mayer (2024).

4.5 Summary of results

Several clear trends emerge from the results presented above:

- Non-positional models outperform positional models in every case.
- Non-positional smoothed models generally outperform non-positional unsmoothed models. However, for positional models smoothing typically has little effect and sometimes reduces their performance.
- Frequency-weighted models generally performed similarly to or slightly worse than non-frequency weighted models. The exceptions to this are Hayes & White (2013), Scholes (1966) and Jarosz & Riesling (2017). The latter two cases are not really exceptions because the frequency information in the training data consists of onset type frequencies rather than token frequency: the success of the frequency-weighted models in these cases simply indicates that type frequency is important. The only true exception is Hayes & White (2013), where including token frequency improved the performance of the smoothed model.

What is even more striking about these results is that they emerge across a range of different domains and languages: some, like Scholes (1966), Daland et al. (2011), and Jarosz & Rysling (2017) focus only on onsets, while the others look at whole word forms. In all cases, non-positional models that encode information about word edges, but no other absolute position information best predict native speaker judgements.

Why should it be the case that non-positional models do better? There are several possible reasons. First, as mentioned earlier in Section 2.1, positional models have issues with data sparsity which make it difficult for them to assign accurate scores to long words: estimates for later positions will necessarily be based on less data than for earlier positions, because there are fewer words with material in those positions. Thus, we might expect scores assigned to longer words to be less useful in predicting human behavior. However, the maximum length of test words in the eight studies we looked at was 7 segments, and even in published results on monosyllabic nonce words with fewer segments or onsets alone (Albright and Hayes 2003, Scholes 1965, Jarosz and Riesling 2017), non-positional models outperformed positional ones.

Second, the non-positional metrics can capture phonotactic constraints that target both word-initial and word-final material, which are often important positions in terms of phonotactic constraints (e.g. Beckman 1997, Lombardi 1999) and an important source of information in word segmentation and allophonic learning. In native Turkish words, for example, /ɾ/ can never begin a word and words cannot end in voiced stops or affricates. The non-positional models can encode this with the use of boundary symbols: such a model trained on Turkish would assign a low probability to the sequences '#r' and 'b#', where # is a word boundary, reflecting the prohibition on word-initial /r/ and word-final voiced stops. Although the positional models can encode word initial constraints, since every occurrence of a sound in position 0 is a word-initial occurrence, they cannot encode word-final constraints, since the position of the final element in a word depends on its length. A positional model struggles to encode restrictions like Turkish's ban on final voiced stops: a [b] in the third position could have a high probability if a word is of length five (as in [babam] 'my father'), but a low probability if it's of length three.

Finally, in addition to these factors, it may simply be the case that non-positional models simply correspond better to human cognitive processes than positional ones do, either because humans do not take absolute position into account, because they compute conditional rather than joint probabilities, or some combination of these. More detailed investigation using tools such as the UCIPC will be useful in teasing these factors apart.

These results have important implications. First, positional metrics of phonotactics, at least as currently implemented, do not predict human phonotactic generalization as well as non-positional metrics: the non-positional metrics outperform the positional ones in every case. This suggests that the positional metrics used by many phonotactic calculators, including the popular Phonotactic Probability Calculator (Vitevitch & Luce 2004), may not be the most suitable for modeling human acceptability judgments.

Second, it is generally the case that models that take token frequency into account perform more poorly than models that don't: this is somewhat less clear cut, however. Further research is needed to determine the extent to which speakers are sensitive to token frequency when forming phonotactic judgments.

Finally, smoothed models generally outperform unsmoothed models. This is not surprising: speakers do not judge words containing unattested sequences as totally unacceptable. It is important to note, however, that the add-one smoothing used in these models is rather coarse, assigning each unattested sequence the same pseudo-count. It has been well established in linguistic research that speakers generalize to unattested sequences based on their similarity to existing sequences in the language (e.g. Chomsky & Halle 1965, Hayes & Wilson 2008, Wilson & Gallagher 2018, Dai et al. 2023, a.o.). More robust smoothing metrics that can capture these differences would be valuable but are beyond the scope of the current paper.

5. Applications of phonotactic metrics in stimulus construction

In addition to serving as variables of interest in experimental work or computational models of speech, the metrics calculated by the UCIPC are also useful for constructing and selecting experimental stimuli.

5.1 Experiments with infants

As shown in the previous sections, the UCIPC can be used to calculate a wide array of metrics to summarize how likely segments and segment sequences are in any given corpus. Such metrics are also extremely useful when constructing stimuli for experiments. For instance, the UCIPC can be used to quantify the extent to which some segments or segment sequences are frequent, in any language for which a phonologically transcribed lexicon or corpus of speech is available. Such quantification is necessary when manipulating segment or segment sequence frequencies as an independent variable in experiments designed to determine when, if at all, infants are sensitive to native language patterns (e.g., Archer & Curtin, 2011; Gonzalez-Gomez & Nazzi, 2012; Friederici & Wessels, 1993). Quantification is also necessary to identify and index experimental confounds when differences in segment and segment sequence likelihood are not the target of inquiry but are nonetheless likely to influence infant behavior (Gonzalez-Gomez & Nazzi, 2012; Nazzi, Bertoncini, Bijeljac-Babic, 2009; Sebastián-Gallés & Bosch, 2002; Solá-Llonch & Sundara, under review).

In addition to standardizing the calculations of metrics to promote replicability, tools like the UCIPC lower the bar for new investigators, particularly those working on under resourced languages to develop stimuli. Typically, to develop stimuli in a new language, an investigator would need access to a corpus, as well as computational skills to conduct corpus analyses to identify patterns and index the incidence of sounds and sound sequences, With UCIPC, the metrics can be obtained for any language as long as there is a dataset with all the words in a dictionary or corpus listed in a consistent transcription system. With time, we expect to increase the number of pre-existing datasets for different languages, to alleviate the challenge of identifying suitable-sized corpora in different languages.

5.2 Experiments with Artificial Languages

The outcome of artificial language experiments has been reported to differ in adults with different native languages (White, et al., 2018; Huang and Do, 2021, Do and Yeung, 2021, etc.). This is typically dealt with by either recruiting only participants who speak the same language(s), so that the same L1 biases are shared across participants, or to use language background as a control variable in the analysis. Phonotactic metrics such as those generated by the UCIPC can also be useful in designing or analyzing artificial language learning experiments. For example, if a study were to be run on both English and Spanish speakers, phonotactic models fit to English and Spanish training data could be useful to score each stimulus and identify and remove cases where the models' scores deviate substantially between languages. Alternatively, these scores could themselves be used as control variables, rather than the coarser metric of language background. This approach has the potential not only to better control for L1 effects in AGL, but also to quantify and predict them.

Finally, the artificial language itself can be used as the training data to ensure that the test items do not vary on segment and segment sequence likelihood that are themselves not the target of inquiry

6. Discussion and conclusion

In this paper we have presented the UCI Phonotactic Calculator, a new online tool that allows users to compute a suite of different phonotactic acceptability metrics. Compared to existing tools, the UCIPC has several desirable properties:

- It allows users to provide their own training data, allowing it to be applied to any language, whether natural or artificial, for which appropriate training data is available
- It computes a large suite of different types of acceptability metrics
- It has a simple point and click interface that allows it to be used by researchers with limited technical backgrounds

Section 4 provided an example of how the calculator can be applied to answer questions about what aspects of phonotactic patterns speakers encode and how they generalize to unattested patterns. This demonstrated that, overall, models that are not sensitive to absolute position in the word or to token frequency do the best at predicting human judgments across a range of studies in four different languages.

The UCIPC has several valuable research applications in addition to modeling phonotactic acceptability. It can be used in stimulus construction for lexical decision tasks, infant experiments, or artificial grammar learning studies to control for the effects of phonotactic probability in participants' native languages. It can also be used be used to model changes in phonotactic generalizations resulting from changing vocabulary size (see, e.g., Sundara and Breiss, submitted).

We hope that the UCIPC will be a valuable tool for researchers who are interested in phonotactic acceptability. We would like to close by emphasizing that the UCIPC is an open-source project: the source code can be freely examined, and we welcome contributions from researchers who would like to add additional functionality or fix existing bugs.

Declarations

Funding: This research was funded by NSF award #2214017 to CM and MS.

Conflicts of interest/competing interests: The authors have no relevant financial or non-financial interests to disclose.

Ethics approval: The Turkish and Spanish data were collected under UCI IRB #2060 "Evaluating feature-vs. segment-based phonotactic generalizations in adults.

Consent to participate: Informed consent was obtained from all participants from the Turkish and Spanish studies.

Consent for publication: All participants in the Turkish and Spanish studies provided informed consent for publication of their study results.

Availability of data and materials: The UCI Phonotactic Calculator can be accessed at https://phonotactics.socsci.uci.edu/. The data and code used in the analyses in Section 4 can be found at https://github.com/aryarksub/phonotactic_metrics.

Code availability: The code for the UCI Phonotactic Calculator can be found at https://github.com/connormayer/uci_phonotactic_calculator.

Open practices statement

Links to the data and code are provided above in the Declarations section. The Turkish and Spanish studies were not preregistered.

Acknowledgments

Thanks to Canaan Breiss for being an early adopter of the UCI Phonotactic Calculator and helping us identify and address several issues in its implementation.

Appendix A

Aside from the web interface, phonotactic metrics can also be calculated via the command-line interface for the UCIPC. To use the interface, users must download the UCIPC source code from the GitHub repository and, in their local terminal, navigate to the `src` directory which holds the `ngram_calculator.py` file. The calculator can then be run with the command

```
python ngram_calculator.py [train_file] [test_file] [results_file]
```

where the arguments refer to the local paths to the training file, test file, and output file to use, respectively. For example, using the command-line interface on sample files located in the `data` directory can be done as follows:

```
python ngram_calculator.py ..\data\english_cmu_freq.txt
..\data\sample_test_data\english_test_data.csv outfile.csv
```

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6): 716–723.

Aljasser, F., & Vitevitch, M.S. (2018) A web-based interface to calculate phonotactic probability for words and nonwords in Modern Standard Arabic. *Behavior Research Methods*, *50:* 313-322.

Archer, S. L., & Curtin, S. L. (2011). Perceiving onset clusters in infancy. *Infant Behavior and Development*, *34*, 534–540.

Baayen, R.H., Piepenbrock, R., & Gulikers, L. (1995). CELEX2 LDC96L14. Web Download. Philadelphia: Linguistic Data Consortium.

Bailey, T.M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods. *Journal of Memory and Language 44*:568–591.

Bard, E.G., Robertson, D., and Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language, 72*: 32–68.

Beckman, J. N. (1997). Positional faithfulness, positional neutralisation and Shona vowel harmony. *Phonology, 14*(1), 1-46.

Berent, I., Lennertz, T., Jun, J., Moreno, M.A., & Smolensky, P. (2008). Language universals in human brains. *Proceedings of the National Academy of Sciences, 105*:5321–5325.

Burnham, K.P., & Anderson, D.R. (2004). Multimodal inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*(2): 261-304.

Chen, S.F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language 13*(4): 359–394.

Chodroff, E. & Wilson, C. (2014). Phonetic vs. phonological factors in coronal-to-dorsal perceptual assimilation. Paper presented at LabPhon 14: the 14th Conference on Laboratory Phonology, Tokyo.

Chomsky, N., & Halle, M. (1965). Some controversial questions in phonological theory. *Journal of Linguistics,* 1(2):97–138.

CMU pronouncing dictionary (2008). *Carnegie Mellon University pronouncing dictionary*. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Coleman, J., & Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. In Coleman, J. (ed.), *Proceedings of the 3rd Meeting of the ACL Special Interest Group in Computational Phonology*. Association for Computational Linguistics, Somerset, NJ: 49-56.

Dai, H., Mayer, C., & Futrell, R. (2023). Rethinking Representations: A Log-bilinear Model of Phonotactics. *Proceedings of the Society for Computation in Linguistics*: Vol. 6, Article 24.

Daland, R. (2015). Long words in maximum entropy phonotactic grammars. *Phonology, 32*(3):353–383

Daland, R., Hayes, B., White, J., Garellek, M., Davis, A., & Normann, I. (2011). Explaining sonority projection effects. *Phonology, 28*: 197–234.

Davidson, L. & Shaw, J.A. (2012). Sources of illusion in consonant cluster perception. *Journal of Phonetics, 40*: 234-24.

Do, Y., & Yeung, P. H. (2021). Evidence against a link between learning phonotactics and learning phonological alternations. *Linguistics Vanguard, 7*(1), 20200127.

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., Carreiras, M. (2013). EsPal: One-stop Shopping for Spanish Word Properties. Behavior Research Methods, 45(4): 1246-58.

Duyck, W., Desmet, T., Verbeke, L., & Brysbaert, M. (2004). WordGen: A Tool for Word Selection and Non-Word Generation in Dutch, German, English, and French. Behavior Research Methods, Instruments & Computers, 36(3), 488-499.

Friederici, AD, & Wessels, JM. (1993). Phonotactic knowledge of word boundaries and its use in infant speech perception. Perception & Psychophysics, 54(3), 287-295.

Gonzalez-Gomez, N & Nazzi, T. (2012). Acquisition of nonadjacent phonological dependencies in the native language during the first year of life. Infancy, 17(5), 498-524.

Haman, E., Etenkowski, B., Łuniewska, M., Szwabe, J., Dąbrowska, E., Szreder, M., & Łaziński, M. (2011). Polish CDS Corpus.

Hayes, B., & White, J. (2013). Phonological naturalness and phonotactic learning. *Linguistic Inquiry, 44*:45-75.

Hayes, B., & Wilson, C. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry 39,* 379-440.

Huang, T., & Do, Y. (2021). Phonetically Grounded Structural Bias in Learning Tonal Alternations. *Frontiers in Psychology, 12*, 705766.

Inkelas, S., Küntay, A., Orgun, O., & Sprouse, R. (2000). Turkish electronic living lexicon (TELL). *Turkic Languages, 4,* 253-275.

Jarosz, G. (2017). Defying the Stimulus: Acquisition of Complex Onsets in Polish. *Phonology, 34*(2): 269-298.

Jarosz, G., & Rysling, A. (2017). Sonority Sequencing in Polish: the Combined Roles of Prior Bias and Experience. *Proceedings of the 2016 Annual Meetings on Phonology*, USC.

Jusczyk, P.W., Luce, P.A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language, 33*: 630-645.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. Behavior Research Methods 42(3), 627-633.

Kučera, H., & Francis, W. N. (1967). Computational analysis of present-day American English. Providence, RI: Brown University Press.

Lodge, M. (1981). *Magnitude scaling: Quantitative measurement of opinions*. Beverly Hills, CA: Sage.

Lombardi, L. (1999). Positional faithfulness and voicing assimilation in Optimality Theory. Natural Language & Linguistic Theory, 17(2), 267-302.

Marian, V., Bartolotti, J., Chabal, S., Shook, A. (2012). CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. PLoS ONE 7(8): e43230. doi:10.1371/journal.pone.0043230

Markov, A. A. (1913). Essai d'une recherche statistique sur le texte du roman "Eugene Onegin" illustrant la liaison des epreuve en chain ('Example of a statistical investigation of the text of "Eugene Onegin" illustrating the dependence between samples in chain'). *Izvistia Imperatorskoi Akademii Nauk (Bulletin de l'Académie Impériale des Sciences de St.-Pétersbourg), 7*: 153–162.

Mayer, C. (2020). An algorithm for learning phonological classes from distributional similarity. *Phonology, 37*(1), 91-131.

Mayer, C. (2024). One (semi)ring to rule them all: reconciling categorical and gradient models of phonotactics. Talk presented at the LSA Session on Formal Language Theory in Morphology and Phonology. 2024 Annual Meeting of the Linguistic Society of America. New York, NY.

Mayer, C., & Nelson, M. (2020). Phonotactic learning with neural language models. *Proceedings of the Society for Computation in Linguistics*. Vol. 3. Article 16.

Nazzi, T, Bertoncini, J, & Bijeljac-Babic, R. (2009). A perceptual equivalent of the labial-coronal effect in the first year of life. The Journal of the Acoustical Society of America, 126(3), 1440-1446.

Needle, J. M., Pierrehumbert, J. B., & Hay, J. B. (2022). Phonotactic and Morphological Effects in the Acceptability of Pseudowords. In A. Sims, A. Ussishkin, J. Parker, & S. Wray (Eds.), *Morphological Diversity and Linguistic Cognition*. CUP.

New, B., & Spinelli, E. (2013). Diphones-fr: A French database of diphone positional frequency. Behavior research methods, 45(3): 758-764.

Norris, D. & McQueen, J.M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review, 115:* 357

Prince, A., & Smolensky, P. (1993/2004). Optimality theory: Constraint interaction in generative grammar. Cambridge, MA: Blackwell. (Technical Report CU-CS-696-93, Department of Computer Science, University of Colorado at Boulder, and Technical Report TR-2, Rutgers Center for Cognitive Science, Rutgers University, New Brunswick, NJ, April 1993.)

Scholes, R. (1966). Phonotactic grammaticality. The Hague: Mouton.

Selkirk, E. (1984). On the major class features and syllable theory. In Aronoff, M., and Oehrle, R.T. (eds.), *Language sound structure: Studies in phonology presented to Morris Halle by his teacher and students.* MIT press, Cambridge, MA. 107-113.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Technical Journal, 27*(3):379–423.

Solá-Llonch, E., & Sundara, M. (under review). Young infants' sensitivity to vowel harmony is independent of language experience. Infant Behavior and Development

Steffman, J., & Sundara, M. (2024). Disentangling the role of biphone probability from neighborhood density in the perception of nonwords. *Language & Speech, 67* (1), 166-202.

Sundara, M., & Breiss, C. (submitted). The acquisition of native language phonotactics: Integrating insights from machine learning, and adult and infant experiments. *Cognition*.

Sundara, M., Zhou, Z.L., Breiss, C., Katsuda, H., & Steffman, J. (2022). Infants' developing sensitivity to native language phonotactics: A meta-analysis. *Cognition, 221:* 104993

Swingley, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science, 13*(5), 480–484.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language, 40*(3): 374–408.

Vitevitch, M.S., & Luce, P.A. (2004) A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, and Computers, 36:* 481-487.

Vitevitch, M.S., & Luce, P. (2016). Phonological neighborhood effects in spoken word perception and production. *Annual Review of Linguistics, 2:* 75-94.

Vitevitch, M.S., Stamer, M.K., & Kieweg, D. (2012). The Beginning Spanish Lexicon: A Web-based interface to calculate phonological similarity among Spanish words in adults learning Spanish as a foreign language. *Second Language Research, 28*, 103-112.

Weide, R.L. (1994). CMU Pronouncing Dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

White, J., Kager, R., Linzen, T., Markopoulos, G., Martin, A., Nevins, A., Peperkamp, S., Polgárdi, K., Topintzi, N., van De Vijver, R. (2018). Preference for locality is affected by the prefix/suffix asymmetry: Evidence from artificial language learning. In Sherry Hucklebridge & Max Nelson (eds.), *NELS 48: Proceedings of the Forty-Eighth Annual Meeting of the North East Linguistic Society: Vol. 3*. Amherst, MA: GLSA, 207–220.

Wilson, C., & Gallagher, G. (2018). Accidental gaps and surface-based phonotactic learning: a case study of South Bolivian Quechua. *Linguistic Inquiry, 49*(3): 610-623.