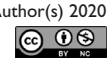
*Article*

Language *and* Speech

# Scaling Laws for Phonotactic Complexity in Spoken English Language Data

Language *and* Speech
2021, Vol. 64(3) 693–704



DOI: 10.1177/0023830920944445
journals.sagepub.com/home/las


**Andreas Baumann**
Department of English and American Studies, University of Vienna, Austria

**Kamil Kaźmierski**
Department of Contemporary English Language, Faculty of English, Adam Mickiewicz University, Poznań, Poland

**Theresa Matzinger**
Department of English and American Studies & Department of Behavioral and Cognitive Biology, University of Vienna, Austria

**Abstract**
Two prominent statistical laws in language and other complex systems are Zipf's law and Heaps' law. We investigate the extent to which these two laws apply to the linguistic domain of phonotactics—that is, to sequences of sounds. We analyze phonotactic sequences with different lengths within words and across word boundaries taken from a corpus of spoken English (Buckeye). We demonstrate that the expected relationship between the two scaling laws can only be attested when boundary spanning phonotactic sequences are also taken into account. Furthermore, it is shown that Zipf's law exhibits both high goodness-of-fit and a high scaling coefficient if sequences of more than two sounds are considered. Our results support the notion that phonotactic cognition employs information about boundary spanning phonotactic sequences.

**Keywords**
Heaps' law, Zipf's law, phonotactics, diversity, inventory size

## Introduction

Naturally emerging complex systems typically obey characteristic statistical laws. For instance, preferential attachment in emerging networks ("the rich get richer" principle) leads to a link distribution that follows a decreasing power law (Barabási & Pósfai, 2016). One of the most prominent

**Corresponding author:**
Andreas Baumann, Department of English and American Studies, University of Vienna, Spitalgasse 2-4, 8.3, Vienna, 1090, Austria.
Email: andreas.baumann@univie.ac.at

statistical laws ubiquitous in complex systems is Zipf's law (Zipf, 1949). It accounts for distributional patterns in demographics, economy, bibliometrics, and language (Corominas-Murtra & Solé, 2010; Ferrer-i-Cancho, 2016; Li, 2002). In linguistics, it models the skewed distribution of frequency of occurrence in a population of types (e.g., words) or equivalently the inverse relationship between token frequency $f$ and rank $r$, that is, $f \propto r^{-\alpha}$ with exponent $\alpha$. Similarly, Zipf's law describes the inverse relationship between the complexity of a given type (e.g., in terms of its length, size, or duration) and its frequency. Another widely attested statistical scaling law, related to Zipf's law, is Heaps' law (also known as Herdan's law; Heaps, 1978; Lü et al., 2010). It is a model of a system's complexity (typically the number of types) depending on the number of tokens in it, that is, the sample size. According to Heaps' law, complexity grows with the number of tokens in a sample in a sublinear fashion.

Most linguistic accounts of Zipf's law and Heaps' law operate on the lexical level: that is, they address the relationship between a word's frequency and its rank, and the relationship between corpus size and the number of word types in it. Its validity has been demonstrated cross-linguistically (Baayen, 2001). The picture seems to be less straightforward with regard to the phonological level. For 95 languages it has been shown that the relationship between phoneme frequency and phoneme rank only roughly follows Zipf's law (Tambovtsev & Martindale, 2007).

In this study, we focus on the domain of phonotactics, being systems of sound sequences also referred to as *n*-phones (where *n* stands for the number of phonological constituents in the sequence). We test to what extent Heaps' law and Zipf's law apply in phonotactics by analyzing *n*-phones (for a range of different lengths *n*) within words and across word boundaries.

Many mechanisms have been proposed to explain Zipf's law and Heaps' law in linguistics, and most of them can arguably be transferred to the phonotactic level as well. First, just like any linguistic subsystem, systems of sound sequences can be conceived of as complex systems which have evolved over time (e.g., Barabási & Pósfai, 2016). If this process is driven by preferential attachment, so that the choice of a sound sequence depends on how established (or "entrenched") it is already, this entails a skewed distribution of token frequencies of *n*-phone types. The extent to which preferential attachment applies is then reflected in the degree to which this distribution is skewed and translates into the power-law exponent $\alpha$ in $f \propto r^{-\alpha}$. No preferential attachment corresponds to a vanishing exponent (Baek et al., 2011).

A second hypothetical driving force, in fact brought into play by Zipf and colleagues (Newman & Zipf, 1936), is the principle of least effort, which has been suggested to explain the characteristic distributional properties on the phonological level. Some sounds are easier to produce, being for instance less complex and more well-formed, than others, and the more complex a sound is, the less frequently it is used (e.g., Deng, 2016). Similarly, sequences of sounds differ as to their complexity. For example, sequences of consonants and vowels are generally considered articulatorily and perceptually less complex than sequences of consonants (Levelt & Van De Vijver, 2004). Related to this, using graphemics as a proxy for phonotactics, Mahowald et al. (2018) have demonstrated a positive correlation between phonotactic probability as a measure of well-formedness and lexical frequency.

Third, it was argued that Zipf's law may be a consequence of multiple underlying and interacting processes, which in isolation would not necessarily give rise to power-law distributions (Aitchison et al., 2016). On the lexical level, it was argued by Lestrade (2017) on computational grounds that an interaction of syntactic and semantic factors provides a better explanation of Zipf's law in the lexicon than each of these two domains in isolation. Clearly, phonotactics is influenced by various linguistic domains as well: sound sequences are brought about through concatenating phonemes within morphemes (i.e., lexical phonotactics in the narrow sense), through morphology (e.g., affixation, ablaut), or through syntax (across word boundaries).

With regard to Heaps' law it was shown rigorously by Lü et al. (2010) that in evolving complex systems with finite size, Zipf's law entails Heaps' law. In that sense, they argue Zipf's law to be the more fundamental one of the two. Furthermore, they derive the interesting relationships that Heaps' exponent is positively correlated with system size and non-negatively correlated with Zipf's exponent. So, if a grown linguistic subsystem—such as phonotactics—obeys Zipf's law, these relationships are expected to hold.

Although the applicability of scaling laws to phonotactics is not implausible, as argued above, phonotactic scaling laws are rather understudied. Ha et al. (2009) investigate Zipf's law in systems of *n*-phones of length 2 up to 13, but they limit their analysis to word-internal phonotactics. Excluding word-boundary spanning *n*-phones from phonotactic research is potentially problematic for several reasons.

To begin with, this is because low-probability and/or complex phonotactic sequences have been suggested to fulfill the function of signaling word boundaries (Daland & Pierrehumbert 2011), and thus assist the listener in the decomposition of the speech stream into words.

Similarly, low-frequency diphones can signal morpheme boundaries or the boundaries between the individual parts of compound words (Finley & Newport, 2011). For example, in English, the 2-phone consonant cluster /kf/ does not occur word-internally and therefore strikes listeners as being ill-formed when they encounter it within a word. In contrast, the same sequence occurs frequently and sounds perfectly fine across word boundaries (e.g., "pink flamingo") or across the boundaries of the parts of compound words (e.g., "workflow", cf. Daland & Pierrehumbert 2011). Likewise, in English, the word-final cluster /ts/ does only rarely surface word internally (as in "blitz") but mostly occurs across morpheme boundaries (e.g., "cats" or "hits"). Thus, when listeners hear such sequences, they can interpret them as boundary signals (Jusczyk, 1999; Mattys & Jusczyk, 2001; Saffran et al., 1996).

In a similar vein, the "naïve discriminative learning" approach (Baayen et al., 2016; Milin et al., 2017) argues that lexical structure can arise as the result of phonotactic distribution frequencies, so that actual morphemes, words and phrases indeed emerge from phonotactics (Divjak, 2019).

Taking this research into account, across-word phonotactics seem to fulfill a non-negligible role in the interface between phonology, morphology, syntax and the lexicon. As a consequence, accounts for scaling laws in phonotactics should not exclude boundary-spanning phonotactic items per se. Furthermore, if analyses are limited to word-internal phonotactics it cannot be ruled out that potential scaling laws in phonotactics are mere epiphenomena of corresponding scaling laws in the lexicon.

Finally, and relevant from a methodological perspective, the notion of what counts as a word (at least in corpus linguistics) is biased by graphemics among other factors (Haspelmath, 2011) so that word boundaries are usually equated with whitespaces in written text. This creates a multitude of problems: compounds are graphemically realized differently across languages; some writing systems do not show clear word boundaries; words are not the primary building blocks in polysynthetic and incorporating languages.

The goal of this study is to provide an empirical analysis of Heaps' law and Zipf's law in phonotactics and to test whether they indeed hold in this domain. We compare phonotactics within word boundaries ("within-word phonotactics") to phonotactic systems that also allow for boundary-spanning *n*-phones ("within-and-across-word phonotactics"). We analyze *n*-phones of phonotactic length 2 to 6 in a corpus of spoken American English and test to what extent scaling-law characteristics (exponents) are related with phonotactic length and system size.

We demonstrate (a) that measures of phonotactic complexity taking frequency into account are more reliable than complexity measures only based on inventory size (see Rama, 2013, and discussion below) and (b) that within-and-across-word phonotactics (but not so much within-word phonotactics) shows behavior typical of emerging complex systems, indicating that phonotactic

cognition also covers information about boundary spanning *n*-phones. Finally, we argue that (c) cognitively plausible phonotactic systems consist of sequences with more than two sounds and (d) we highlight differences between scaling laws in phonotactics and the lexicon.

## 2 Data and methods

We used the Buckeye Speech Corpus (Pitt et al., 2005), which contains about 300,000 phonologically transcribed word tokens produced by a total of 40 speakers of American English. We extracted 10 nested sub-corpora from the corpus so that the smallest sub-corpus counts about 10,000 tokens, and each subsequent sub-corpus is a superset of and about 10,000 tokens larger than its predecessor.

From each sub-corpus, we extracted all *n*-phones with different length *n* from 2 to 6. In theory, (across-word) *n*-phones can be arbitrarily long, but for practical considerations, we had to set an upper limit for our investigations. Our choice is motivated by results from research on the human working memory, which has been argued to be limited to processing 3 to 5 segments at a time (Green, 2017; Mathy & Feldman, 2012). Applying this notion to phonotactics, we set our limit slightly above this cognitively grounded figure and analyzed *n*-phones up to a length of 6. In addition, this limit ensures that we do not consider sequences which are much longer than the average number of phonemes in a word, which is about $7 \pm 2$ (a cross-linguistic estimate, for example as reported by Nettle, 1995; a similar mean length in the lexicon can be derived from English spoken frequencies in the CELEX database, namely 7.25 phonemes; Baayen et al., 1995).

The extraction of *n*-phones was done in two different ways: (1) In the within-word condition, all *n*-phones (for a given *n*) occurring within word forms (i.e., word tokens separated by whitespaces) were extracted. For instance, the sequence *Heaps' law* features four within-word 2-phones /hi, ip, ps, lɔ/, two within-word 3-phones /hip, ips/, and a single within-word 4-phone /hips/. For each *n*-phone type *i*, the overall frequency of occurrence $f_i$ was computed; (2) In the within-and-across-word condition, word (and sentence) boundaries were ignored, so that the sequence *Heaps' law* shows five 2-phones /hi, ip, ps, sl, lɔ/, four 3-phones /hip, ips, psl, slɔ/, etc.

Next, we computed two complexity measures for each sub-corpus and each *n*. First, we retrieved phonotactic inventory size $d_0$, that is, the number of *n*-phone types. Second, we computed phonotactic diversity as $d_1 = \exp H$, where $d_1 = \exp H$ , where $H = -\sum p_i \log p_i$ is Shannon entropy, and where $p_i = f_i / \sum f_j$ is the probability of type *i*. That is, phonotactic diversity is high if all *n*-phone types are roughly equally frequent and low if some types are relatively frequent while others are rare.

Both sizes are special cases of the more general diversity number of order *q*, defined as $d_q = \left(\sum p_i^q\right)^{1/(1-q)}$ (see Hill, 1973, for a formal derivation). Thus, $d_0$ ($q = 0$) provides a relatively rough measure of phonotactic complexity, while $d_1$ ($q = 1$) is more fine-grained since it also captures token frequency. We have chosen these two measures of complexity, since $d_0$ is a standard measure of the complexity of phonological systems (Nettle, 1995, 2012; Rama, 2013; Wichmann et al., 2011) and functions as a straightforward analogue of lexicon size. Moreover, $d_1$ is closely related with Shannon entropy, which itself was studied on the lexical level in relation with corpus size (Febres, Jaffé, & Gershenson, 2015). Also note that there is a close relationship between $d_1$ and Yule's characteristic *K*, the latter of which, as a measure of lexical repetition, has been used as a relatively size-independent operationalization of lexical diversity (Miranda-García & Calle-Martín, 2005): $d_1 = \exp H$ is an upper bound of the reciprocal of Yule's *K* (Herdan, 1958).

To measure system size, we determined the number of *n*-phone types *t* in the overall corpus for each *n* respectively.

Heaps' law describes the sublinear growth of complexity $d_q$ (usually the number of types) as a function of corpus size s, given by $d_q(s) = c \cdot s^{\beta}$ , where $c > 0$ is a constant and $\beta < 1$ is Heaps' exponent. We consider two different exponents: $\beta_0$ for the growth of inventory size $d_0$ and $\beta_1$ for

the growth of diversity $d_0$. For each $n$, we estimated both exponents by means of non-linear least-squares regression (Figure 1(a)).

Zipf's law can be formalized as $f(r) = f_1 \cdot r^{-\alpha}$ so that a type's token frequency $f$ is a function of its rank $f$. Here, $f_1$ is the frequency of the most frequent type (with $r = 1$) and $\alpha$ is Zipf's exponent. Again, we estimated the exponent for each $n$ via non-linear least-squares regression (Figure 1(b)). For both laws, goodness-of-fit (GoF) was assessed by means of adjusted $R^2$.

Since we consider $n$-phones with length 2 to 6, this resulted in a series of 5 estimates for each of the variables (number of tokens; number of types; exponents $\beta_0$; $\beta_1$; $\alpha$) and each condition (within, within-and-across). We computed pairwise correlation coefficients (Pearson's $R$) for all combinations of two variables (Figure 2). All computations were done in R (R Development Core Team, 2017).

# 3 Results

The computed values for number of tokens, number of types, and the exponents $\beta_0$, $\beta_1$, and $\alpha$ for all lengths $n$ are shown in Tables 1 and 2, respectively.

Growth curves for Heaps' law are shown in Figure 1(a). The curves flatten more slowly as $n$ gets larger. This is reflected in the exponent estimates, which increase with $n$. Similarly, it is evident that the growth curves of inventory size $d_0$ flatten more slowly than the growth curves of diversity $d_1$ indicating that diversity is less susceptible to changes in sample size than it is the case for inventory size. It can be seen from Table 1- that growth exponents are on average higher in the within-and-across-word condition than in the case of exclusively lexical phonotactics (within word condition), which may be attributed to larger system size in within-and-across phonotactics. Thus, phonotactics is more sensitive to changes in sample size if boundary spanning $n$-phones are admitted, in particular for high $n$. Interestingly, $\beta_0$ approaches estimates of Heaps' exponent for lexical items. In the case of 6-phones, $\beta_0 = 0.82$, which comes close to the range for lexical estimates for $\beta_0$ reported, for example by Torre et al. (2017). Note that lexical items have—cross-linguistically—on average about $7 \pm 2$ phonemes (Nettle, 1995).

The relationship between frequency and rank, as described by Zipf's law, can be seen in Figure 1b. In contrast to Heaps' exponent, there is no clear relationship between $\alpha$ and phonotactic length $n$ (neither for the within nor the within-and-across-word condition). It is remarkable that Zipf's coefficients measured for phonotactic items are considerably lower than estimates in the lexical regime which is about $\alpha = 1$ (Zipf, 1949). This entails that phonotactic frequency distributions are less skewed than lexical ones.

Figure 2 shows pairwise correlations for all combinations of variables under investigation. The measures for system size show interesting interactions with $n$. While the number of types $t$ is positively correlated with $n$ in within-and-across-word phonotactics, this is not the case in within-word phonotactics. Similarly, both Heaps' coefficients increase with the number of types in within-and-across-word phonotactics, but not in the within-word condition. Zipf's exponent correlates negatively with all other measures in the within-and-across-word condition, a relationship which cannot be observed in within-word phonotactics. What both conditions have in common is that both Heaps' exponents correlate positively with $n$ and that Zipf's exponent peaks at $n = 3$. Finally, goodness-of-fit is low for $n = 2$ in both conditions but reasonably high for $n > 2$.

# 4 Discussion

In this study, we investigated scaling laws of phonotactic complexity in a phonologically transcribed corpus. We considered $n$-phones—that is, phonotactic constituents—with different length

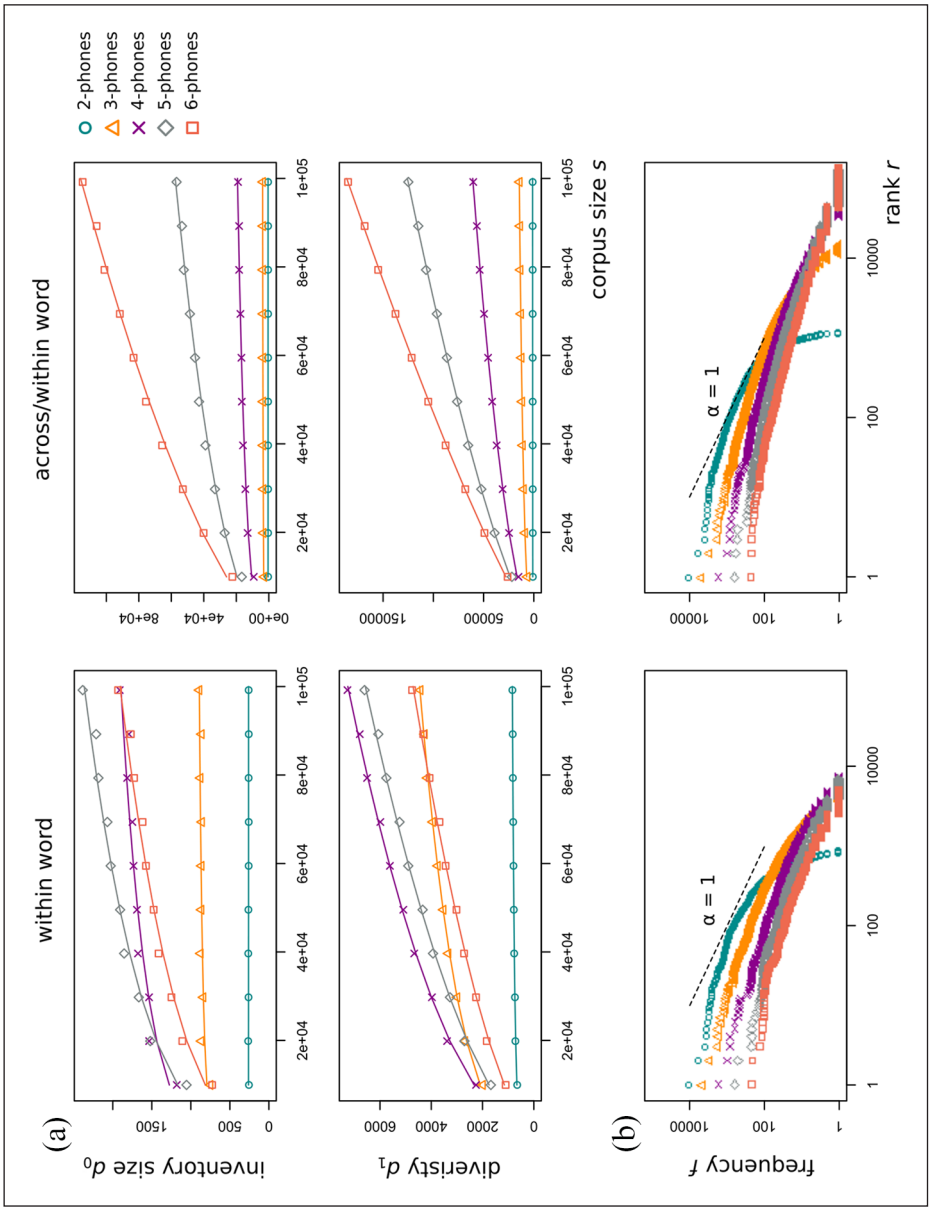



**Figure 1.** Heaps' law (a) and Zipf's law (b) for within-word phonotactics (left) and within-and-across-word phonotactics (right). (a) Phonotactic complexity ($d_0$, $d_1$) depending on corpus size for different phonotactic lengths (see legend in the upper-right corner). Lines represent non-linear least-squares regression models of Heaps' law. (b) Token frequency by rank for different phonotactic lengths. Dashed line represents Zipf's law with exponent $\alpha = 1$ (i.e., a slope of 1 on a log-log scale).

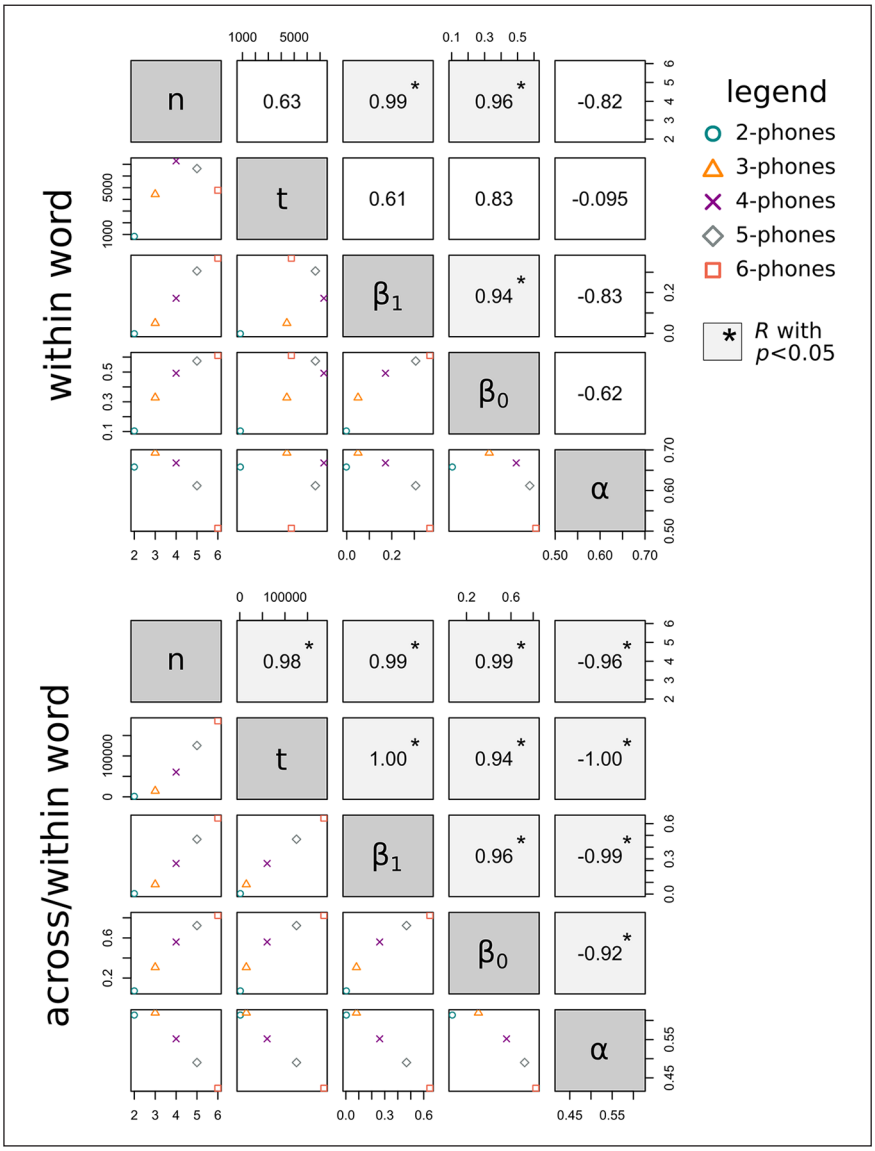



**Figure 2.** Pairwise correlations among measures in within-word and within-and-across-word phonotactics. See Tables 1 and 2.

*n* under two different conditions: within-word phonotactics and unconstrained phonotactics where *n*-phones may span word boundaries. In our analysis, we focused on Heaps' law and Zipf's law. For the former, we operationalized phonotactic complexity by means of two different measures, inventory size and (frequency sensitive) diversity. For each phonotactic length *n*, and each condition Heaps' and Zipf's exponents were estimated.

A number of observations can be made which have implications for our understanding and analysis of phonotactic systems. These concern (a) different ways of measuring phonotactic complexity, (b) the difference between within-word phonotactics and within-and-across-word, that is,

**Table 1.** Within word measures: phonotactic length (*n*), system size (the number of types (*t*)), and estimated exponents for Heaps' law and Zipf's law (together with respective 95% confidence intervals and goodness-of-fit measures).

| *n* | *t* | $\beta_1$ | 95% CI | GoF | $\beta_0$ | 95% CI | GoF | $\alpha$ | 95% CI | GoF |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 837 | −0.002 | (–0.012,0.008) | −0.095 | 0.104 | (0.098,0.11) | −0.094 | 0.658 | (0.652,0.665) | −0.094 |
| 3 | 4433 | 0.05 | (0.025,0.075) | 0.605 | 0.328 | (0.314,0.342) | 0.605 | 0.693 | (0.692,0.695) | 0.605 |
| 4 | 7291 | 0.172 | (0.141,0.203) | 0.928 | 0.492 | (0.476,0.508) | 0.928 | 0.668 | (0.667,0.669) | 0.928 |
| 5 | 6637 | 0.306 | (0.273,0.339) | 0.976 | 0.574 | (0.554,0.594) | 0.976 | 0.612 | (0.61,0.613) | 0.976 |
| 6 | 4781 | 0.369 | (0.332,0.406) | 0.980 | 0.611 | (0.586,0.636) | 0.980 | 0.507 | (0.504,0.509) | 0.980 |

**Table 2.** Across/within word measures: phonotactic length (*n*), system size (the number of types (*t*)), and estimated exponents for Heaps' law and Zipf's law (together with respective 95% confidence intervals and goodness-of-fit measures).

| *n* | *t* | $\beta_1$ | 95% CI | GoF | $\beta_0$ | 95% CI | GoF | $\alpha$ | 95% CI | GoF |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 1176 | 0.004 | (–0.004,0.012) | 0.024 | 0.072 | (0.062,0.082) | 0.024 | 0.614 | (0.609,0.619) | 0.024 |
| 3 | 14367 | 0.082 | (0.058,0.106) | 0.821 | 0.307 | (0.287,0.327) | 0.821 | 0.62 | (0.619,0.621) | 0.821 |
| 4 | 60487 | 0.261 | (0.224,0.298) | 0.958 | 0.56 | (0.542,0.578) | 0.958 | 0.552 | (0.552,0.553) | 0.958 |
| 5 | 125540 | 0.467 | (0.43,0.504) | 0.989 | 0.723 | (0.709,0.737) | 0.989 | 0.49 | (0.49,0.49) | 0.989 |
| 6 | 186062 | 0.647 | (0.618,0.676) | 0.996 | 0.824 | (0.812,0.836) | 0.996 | 0.423 | (0.423,0.424) | 0.996 |

unconstrained, phonotactics, (c) the relationship between the inspected scaling laws and phonotactic length *n*, and (d) the difference between scaling laws in phonotactics and the lexicon. In what follows, we discuss these observations (a—d) in more detail.

First (observation 1), a comparison of the two measures of complexity $d_0$ and $d_1$ reveals that frequency dependent diversity is much less strongly dependent on sample size than is phonotactic inventory size in the sense that the Heaps' coefficients corresponding to phonotactic inventory size are considerably higher. This implies that complexity measures which also take token frequency into account are much more robust with respect to comparisons across corpora and small sample sizes. Measuring differences in phonotactic inventory size (for 2-phones and 3-phones) was suggested as a way of estimating linguistic time depth (Rama, 2013). We argue that frequency-based measures are potentially more reliable tools for these matters (although they require more fine-grained quantitative analyses).

In order to assess the difference between word-internal and unconstrained phonotactics (observation 2) with respect to Heaps' and Zipf's law, let us first consider what would be expected based on formal evidence. Lü et al. (2010) have demonstrated on computational grounds that in emerging complex systems which obey Zipf's law, (a) Heaps' law is also supposed to hold, (b) Zipf's exponent shows a non-positive monotone relationship with Heap's exponent, and (c) that Heaps' exponent increases with system size. If we see (a)–(c) as indicators for how well-behaved a complex system, such as the system of *n*-phones, is, we find that only unconstrained phonotactic systems behave like typical complex systems.

This is because although both laws can be argued to hold in both conditions (based on the estimated coefficients and goodness-of-fit values), we do not find statistically robust support for (b) and (c) in the within-word condition (cf. Figure 2). We argue that it is a reflex of the fact that phonotactic system size decreases with phonotactic length since word length obviously constrains the number of possible word-internal phonotactic items, in particular if phonotactic length is high.

Because it is rather within-and-across-word phonotactics—as opposed to phonotactics restricted to word-internal sequences—which shows the behavior expected for emerging complex systems, this indicates that phonotactic cognition is organized in such a way that it also covers information about boundary spanning items (e.g., conceptualized as phonotactic representations or transition probabilities; Ernestus, 2014). This supports results from research on phonotactically driven speech segmentation, which suggest that listeners infer morpheme and word boundaries from phonotactic information (Daland & Pierrehumbert, 2011; Dressler & Dziubalska-Kołaczyk, 2006; Jusczyk, 1999; Saffran et al., 1996). Clearly, this requires listeners to have access to information on boundary spanning phonotactic sequences.

Going further, word boundaries can be considered as artifacts imposed by the lexical domain in a top-down manner, which should a priori not be taken for granted in phonotactic research. For example, in naïve discriminative learning approaches (Baayen et al., 2016; Milin et al., 2017) words are effectively epiphenomena of distributional properties of phonotactics rather than phonotactics being defined by what is allowed within words. In this model, *n*-phones are input cues and the learning process consists of finding weights to predict outcomes (i.e., lexical items). So, during learning a system of phonotactic sequences emerges which consists of predictive (discriminative) and less predictive cues. Crucially, it is the boundary spanning sequences with low transition probabilities which have high predictive power. This goes in line with our conclusion that these items represent integral parts of phonotactic cognition.

Lastly (observation 3), we can evaluate the relationship between phonotactic length *n* and the two scaling laws. A pattern that we find in the within-and-across condition is that while Heaps' exponent increases and Zipf's exponent decreases with *n*, respectively, goodness-of-fit is generally better for long phonotactic sequences (for $n = 2$, goodness-of-fit is significantly worse than for $n > 2$). As argued before, large Zipf exponents can be interpreted as indicator for preferential attachment (i.e., "the rich get richer") during the evolution of the—in our case, phonotactic—system. At the same time, high goodness-of-fit with respect to Zipf's law hints at multiple interacting processes at work (Aitchison et al., 2016; Lestrade, 2017). Taken together, the behavior of phonotactic length *n* seems puzzling. The question is this: based on the present results, which phonotactic length *n* is cognitively most plausible? It seems that phonotactic sequences in the mid-regime at $n = 3$ or $n = 4$ strike a reasonable balance between coefficient size and goodness-of-fit (for example, this can be seen by looking at the product of Zipf's $\alpha$ and GoF, which takes its maximum at about $n = 4$). This suggests that it is systems of 3-phones or 4-phones that are affected to the largest extent by factors associated with Zipf's law (multiple interacting processes, preferential attachment, least effort). While this conclusion is speculative to a certain extent, it is worthwhile to point out that Baayen et al. (2016) suggest 3-phones to work best as input units for discriminative learning in English. Another potentially related observation is that the expected value of word length (i.e., average word length weighted by frequency) in English speech is about $n = 3.26$ (estimate based on spoken lemma frequencies in CELEX; Baayen et al., 1995) and a hypothesized upper limit of about 3 to 4 segments in short-term memory (Mathy & Feldman, 2012).

The significantly increasing relationship between Heaps' exponent and phonotactic length $n$ is in contrast with findings by Torre et al. (2017). In their study, scaling laws were investigated for acoustic units. These units were defined by means of acoustic energy thresholds so that low thresholds led to short units while high thresholds led to longer units. Interestingly, they found that Heaps' exponent does not change substantially with the size of the energy threshold. Based on our findings, a positive correlation would be expected. Note, however, that the way in which phonemes (or phones) are defined crucially differs from Torre et al.'s (2017) operationalization of units of sound.

As to our final observation (observation 4), we find that Zipf's exponent in phonotactics is much lower than on the lexical level. This suggests that phonotactics is affected less by factors which are thought to give rise to Zipf's law than this is the case in the lexical domain. One reason might be that cognitive constraints related to memory and semantic organization (Piantadosi, 2014) are much weaker in phonotactics than in the lexicon. Clearly, phonotactic sequences carry less meaning than lexical items do (but see Topolinski et al., 2015, and Dressler & Dziubalska-Kołaczyk, 2006, for sound-symbolic and functional properties, respectively).

In addition, we found that Heaps' exponent in phonotactics is generally lower than its lexical counterpart. An important methodological consequence of this is that phonotactic studies do not require the same amount of linguistic data as lexical studies do. This is because a relatively large share of the complexity of the phonotactic system is covered already in small samples. This difference between phonotactics and the lexicon holds particularly true for short within-word phonotactic sequences, such as word-internal 2-phones, but vanishes as phonotactic length approaches the average length of words.

# 5 Conclusion

We have shown that phonotactic systems obey well-established scaling laws that can be found in many complex systems. This implies—not quite surprisingly—that the architecture of phonotactics is far from random but must rather be governed by similar laws of self-organization as other complex systems in linguistics (such as multiple interacting processes, preferential attachment, or the principle of least effort).

In particular, phonotactics exhibits behavior typical of complex systems following Zipf's law if phonotactic items spanning word boundaries are also considered. One advantage of the latter operationalization is that it does not require word boundaries and hence no top-down conceptualization of what linguistic unit counts as a word (which in turn requires a notion of semantics). In that sense, the present paper aligns with the acoustic study by Torre et al. (2017). Consequently, the method used here can be applied to languages with various morphosyntactic structures (such as incorporating or polysynthetic languages). Likewise, it would be interesting to study phonotactic scaling laws in animal vocalizations for which segmented data is available (Kershenbaum et al., 2016).

Moreover, our results tentatively suggest that cognitively plausible phonotactic systems consist of sequences with more than two sounds and that differences between processes operating in phonotactics and the lexicon, respectively, are reflected in properties of their respective scaling laws (coefficients; goodness-of-fit). This, however, needs to be validated on experimental and computational grounds in order to draw robust conclusions.

## Acknowledgements

We would like to thank three anonymous reviewers for various valuable comments that helped to improve our manuscript, analysis, and code.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## ORCID iDs

Andreas Baumann https://orcid.org/0000-0003-4595-2497

Theresa Matzinger https://orcid.org/0000-0001-5414-7962

## Supplemental material

The R code for doing all analyses in this contribution can be found in the script "analysis.R." It requires the script "utils.R" comprising a set of helper functions. All data are collected in the binary "data.RData." Run "analysis.R" to run the computation and to generate human readable data files. The scripts were created under R version 3.4.3. All code and data can be found in the following GitLab project: https://gitlab.com/andreas.baumann/phonotactic_scaling_laws

## References


Aitchison, L., Corradi, N., & Latham, P. E. (2016). Zipf's law arises maturally when there are underlying, unobserved variables. *PLoS Computational Biology*, *12*(12). https://doi.org/10.1371/journal.pcbi.1005110

Baayen, R. H. (2001). *Word frequency distributions* Kluwer Academic Publishers.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *CELEX2*. Linguistic Data Consortium.

Baayen, R. H., Shaoul, C., Willits, J., & Ramscar, M. (2016). Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience*, *31*(1), 106–128.

Baek, S. K., Bernhardsson, S., & Minnhagen, P. (2011). Zipf's law unzipped. *New Journal of Physics*, *13*(4), 043004. https://doi.org/10.1088/1367-2630/13/4/043004

Barabási, A.-L., & Pósfai, M. (2016). *Network science*. Cambridge: Cambridge University Press.

Corominas-Murtra, B., & Solé, R. V. (2010). Universality of Zipf's law. *Physical Review E*, *82*(1), 011102. https://doi.org/10.1103/PhysRevE.82.011102

Daland, R., & Pierrehumbert, J. (2011). Learning diphone-based segmentation. *Cognitive Science*, *35*(1), 119–155.

Deng, Y. (2016). Some statistical properties of phonemes in standard Chinese. *Journal of Quantitative Linguistics*, *23*(1), 30–48. https://doi.org/10.1080/09296174.2015.1071148

Divjak, D. (2019). Frequency in language. Cambridge University Press. https://doi.org/10.1017/9781316084410

Dressler, W. U., & Dziubalska-Kołaczyk, K. (2006). Proposing Morphonotactics. *Wiener Linguistische Gazette*, *73*, 69–87.

Ernestus, M. (2014). Acoustic reduction and the roles of abstractions and exemplars in speech processing. *Lingua*, *142*, 27–41.

Febres, G., Jaffé, K., & Gershenson, C. (2015). Complexity measurement of natural and artificial languages. *Complexity*, *20*(6), 25–48. https://doi.org/10.1002/cplx.21529

Ferrer-i-Cancho, R. (2016). Compression and the origins of Zipf's law for word frequencies. *Complexity*, *21*(2), 409–411. https://doi.org/10.1002/cplx.21820

Finley, S., & Newport, E. L. (2011). Morpheme segmentation in school-aged children. In A. Fine (Ed.), *Rochester working papers in the language sciences*. MIT Press.

Green, C. (2017). Usage-based linguistics and the magic number four. *Cognitive Linguistics*, *28*(2), 209–237. https://doi.org/10.1515/cog-2015-0112

Ha, L. Q., Hanna, P., Ming, J., & Smith, F. J. (2009). Extending Zipf's law to n-grams for large corpora. *Artificial Intelligence Review*, *32*(1–4), 101–113. https://doi.org/10.1007/s10462-009-9135-4

Haspelmath, M. (2011). The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, *45*(1), 31–80. https://doi.org/10.1515/flin.2011.002

Heaps, H. S. (1978). *Information retrieval: Computational and theoretical aspects*. Academic Press.

Herdan, G. (1958). An inequality relation between Yule's characteristic K and Shannon's entropy H. *Zeitschrift für Angewandte Mathematik und Physik ZAMP*, *9*, 69–73. https://doi.org/10.1007/BF01596857

Hill, M. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology*, *54*(2), 427–432.

Jusczyk, P. W. (1999). How infants begin to extract words from speech. *Trends in Cognitive Sciences*, *3*(9), 323–328. https://doi.org/10.1016/S1364-6613(99)01363-7

Kershenbaum, A., Blumstein, D. T., Roch, M. A., Akçay, Ç., Backus, G., Bee, M. A., . . . Zamora-Gutierrez, V. (2016). Acoustic sequences in non-human animals: A tutorial review and prospectus. *Biological Reviews of the Cambridge Philosophical Society*, *91*(1), 13–52. https://doi.org/10.1111/brv.12160

Lestrade, S. (2017). Unzipping Zipf's law. *PloS One*, *12*(8): e0181987. https://doi.org/10.1371/journal.pone.0181987

Levelt, C. C., & Van De Vijver, R. (2004). Syllable types in cross-linguistic and developmental grammars. In *Constraints in Phonological Acquisition*. https://doi.org/10.1017/CBO9780511486418.006

Li, W. (2002). Zipf's law everywhere. *Glottometrics*, (5), 14–21.

Lü, L., Zhang, Z. K., & Zhou, T. (2010). Zipf's law leads to heaps' law: Analyzing their relation in finite-size systems. *PLoS ONE*, *5*(12). https://doi.org/10.1371/journal.pone.0014139

Mahowald, K., Dautriche, I., Gibson, E., & Piantadosi, S. T. (2018). Word forms are structured for efficient use. *Cognitive Science*, *42*(8), 3116–3134. https://doi.org/10.1111/cogs.12689

Mathy, F., & Feldman, J. (2012). What's magic about magic numbers? Chunking and data compression in short-term memory. *Cognition*, *122*(3), 346–362. https://doi.org/10.1016/j.cognition.2011.11.003

Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, *78*(2), 91–121. https://doi.org/10.1016/S0010-0277(00)00109-8

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (2017). Discrimination in lexical decision. *PLoS ONE*, *12*(2). https://doi.org/10.1371/journal.pone.0171935

Miranda-García, A., & Calle-Martín, J. (2005). Yule's characteristic K revisited. *Language Resources and Evaluation*, *39*(4), 287–294. https://doi.org/10.1007/s10579-005-8622-8

Nettle, D. (1995). Segmental inventory size, word length, and communicative efficiency. *Linguistics*, *33*(2), 359–367. https://doi.org/10.1515/ling.1995.33.2.359

Nettle, D. (2012). Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society B*, *367*, 1829–1836.

Newman, S. S., & Zipf, G. K. (1936). The psycho-biology of language. *American Speech*, *21*. https://doi.org/10.2307/451704

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin and Review*, *21*(5), 1112–1130. https://doi.org/10.3758/s13423-014-0585-6

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*(1), 89–95. https://doi.org/10.1016/j.specom.2004.09.001

Rama, T. (2013). Phonotactic diversity predicts the time depth of the world's language families. *PloS One*, *8*(5). https://doi.org/10.1371/journal.pone.0063238

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science (New York, N.Y.)*, *274*(5294), 1926–1928.

Tambovtsev, Y., & Martindale, C. (2007). Phoneme frequencies follow a Yule distribution. *SKASE Journal of Theoretical Linguistics*, *4*(2), 1–11.

R Development Core Team. (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Topolinski, S., Zürn, M., & Schneider, I. K. (2015). What's in and what's out in branding? A novel articulation effect for brand names. *Frontiers in Psychology*, *6*, 585. https://doi.org/10.3389/fpsyg.2015.00585

Torre, I. G., Luque, B., Lacasa, L., Luque, J., & Hernández-Fernández, A. (2017). Emergence of linguistic laws in human voice. *Scientific Reports*, *7*. https://doi.org/10.1038/srep43862

Wichmann, S., Rama, T., & Holman, E. (2011). Phonological diversity, word length, and population sizes across languages: The ASJP evidence. *Linguistic Typology*, *15*(2), 177–197.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press.