PNAS

Supporting Information for

Syllables and their beginnings have a special role in the mental lexicon

Yue Sun[1]* & David Poeppel[1,2,3]
1 Ernst Strüngmann Institute for Neuroscience, Frankfurt am Main, Germany
2 New York University, New York, USA
3 Max Planck-NYU Center for Language, Music, and Emotion (CLaME), New York, USA

Corresponding author: Yue Sun
Email: yue.sun@esi-frankfurt.de

This PDF file includes:

Supporting text 1 to 4
Figures S1
Tables S1 to S5

Supporting Text 1: Onset bias in functional load across different types of consonant clusters

As most examined languages allow multi-consonant clusters to occur at both syllable onset and coda, we further analyzed whether the distribution of functional load between the two syllabic positions differs as a function of the length of consonant clusters that host the contrasting consonant pair. For instance, consonants that distinguish the word '**p**ay' (/**p**eɪ/ - CV) from '**s**ay' (/**s**eɪ/ - CV) belongs to a single-consonant cluster at syllable onset, while those that contrast 'p**l**ay' (/p**l**eɪ/ - CCV) from 'p**r**ay' (/p**r**eɪ/ - CCV) belongs to a double-consonant cluster. This analysis aimed to examine whether the overall onset advantage in functional load observed in each language reflects a common distributional property that exists across clusters of different sizes.

In order to conduct this analysis, we first assessed, for each language, the overall contribution of each type of consonant cluster in constructing phonological wordforms (i.e., frequency of occurrence) and contrasting wordforms from each other (i.e., functional load). This assessment revealed that, across the examined languages, single-consonant clusters (i.e., C clusters) are the predominant structural unit for the construction of phonological wordforms (mean frequency of occurrence = 86.78%; SD = 6.67%) and the primary contributor to lexical contrast (mean functional load = 91.31%; SD = 5.12%) (Table S2). Among multi-consonant clusters, CC clusters showed the highest structural presence across wordforms (mean frequency of occurrence = 13.67%; SD = 5.09%) and the strongest functional contribution in lexical contrast (mean functional load = 9.28%; SD = 4.39%), although their relevance at the level of the entire lexicon is substantially smaller than C clusters. Altogether, across the examined languages, C and CC clusters jointly occupy 99.32% of the total existing structural slots of syllable onset and coda positions across all wordforms and are involved in contrasting 99.83% of the identified minimal pairs. We thus focused on these two types of clusters and examined the distribution of functional load between syllable onset and coda for each cluster type.

To compute the functional load of syllable onset and coda for C and CC clusters, we labelled each minimal pair with the syllabic position of the substituted consonants (onset or coda), and the type of consonant cluster that encapsulated the substituted consonants (C or CC). We then computed the functional load of syllable onset and coda separately for each type consonant cluster. Specifically, the functional load of syllable onset and coda for a given consonant cluster corresponded to the proportion of minimal pairs involving each syllable position among all the minimal pairs that is labelled with that specific consonant cluster. Note that, while all the 12 languages were included for the analysis of C clusters, the analysis of CC clusters only included 10 languages that allow these clusters to occur at both syllable onset and coda. The two languages excluded from the analysis were Turkish that disallows CC clusters to occur at syllable onset[1] and Korean which does not allow CC clusters to occur at either syllable onset or coda position.

Our results showed an onset advantage in functional load for both types of consonant clusters (C cluster: Mean = 63.05%; SD = 20.89%; CC cluster: 65.78%; SD = 29.42%; Figure 2B). Furthermore, our analysis did not reveal significant differences between the two types of consonant cluster in the amount of onset advantage ($t(9) = 0.03$; $p > .1$). These results demonstrated that the onset advantage in functional load is consistently present across the two major cluster types of the examined languages. Moreover, C and CC clusters exhibited similar distribution of functional load between syllable onset and coda, despite of substantial differences in their overall contribution to lexical contrast.

[1] Turkish language has a phonotactic restriction that disallows consonant clusters to occur at syllable onset. The only few words in which consonant clusters occur at syllable onset are loanwords from foreign languages (e.g., tren – 'train'; strateji – 'strategy').

Supporting Text 2: Onset bias in functional load across words with different lengths

We analyzed the distribution of functional load between syllable onset and coda positions across words with different lengths (i.e., number of syllables). This analysis aimed to examine whether an onset advantage in functional load is consistently present across words with different lengths. For this analysis, we focused on words that contain from 1 to 4 syllables, which account for 99.12% of total number of minimal pairs identified across the examined languages.

To compute the functional load of syllable onset and coda for each word length, we labelled each minimal pair with the syllabic position of the substituted consonants (onset or coda), and the lengths of the two words (measured with the number of syllables: 1 to 4). We then computed the functional load of syllable onset and coda separately for each word length. Specifically, the functional load of syllable onset and coda for a given word length corresponded to the proportion of minimal pairs involving each syllable position among all the minimal pairs labelled with that specific word length.

Our result showed that, in all the examined languages, an onset advantage in functional load was present across all word lengths (Figure 2C). Interestingly, across languages, multisyllabic words exhibited higher amount of onset advantage than monosyllabic words, which is marked by a substantial increase of onset advantage from monosyllabic words to disyllabic words (average increase = 50.38%; SD = 16.07%; $t(11) = 10.86$; $p < 0.001$). For words with three or four syllables, our data showed larger variability across languages in the amount of onset advantage. Specifically, the majority of languages (7 of 12) showed similar levels of onset advantage in tri- and quadri-syllabic words comparing to the level in disyllabic words. Meanwhile, four languages (German, English, Swedish and Norwegian) exhibited decrease of onset advantages in tri- and quadri-syllabic words with respect to disyllabic words. Finally, Turkish showed similar onset advantage in trisyllabic words as in disyllabic words, but a substantial increase in quadrisyllabic words.

Our results revealed a consistent increase of the onset bias in functional load from monosyllabic words (27%) to disyllabic words (77%) across all the examined languages. This increase may reflect a trade-off between the preferred word length and the preferred syllabic position in the construction of lexicons. As monosyllabic words are generally favored in language use due to their shortness (1), it is conceivable that all languages exhibit a tendency to create and maintain a large number of monosyllabic words. This tendency, coupled with the skeletal limitation of a single syllable that can maximally hold two syllabic slots for consonants (one onset and one coda), may create pressure for lexicons to fully utilize all available skeletal slots from both syllable onset and coda to construct and differentiate monosyllabic words. Consequently, this pressure leads to smaller onset biases in functional load for monosyllabic words. In the case of disyllabic words, the addition of a second syllable not only increases the number of potential skeletal slots for consonants but also expands the phonological space for creating distinct wordforms. Therefore, lexicons can more easily achieve a wide range of disyllabic wordforms without the necessity to fully exploit all possible skeletal slots for consonants. This, in turn, allows for a greater manifestation of the preference for syllable onset over syllable coda in both word construction and contrast.

Supporting Text 3: Extended Materials & Methods

Lexical databases

Most databases provided all types of the information that are necessary for our analysis, except for Spanish and Turkish. The Spanish database (BuscaPalabras) (2) does not provide lemma status of the phonological wordforms. Therefore, we checked the lemma status for each word in the database from another lexical source of Spanish (EsPal) (3). Phonological wordforms whose lemma status was unknown in EsPal were excluded from the analysis. In addition, the TELL database for Turkish does not provide syllabification. Since the syllabification rules of Turkish are quite strict with very few exceptions (4), we implemented these rules to obtain the syllabified transcription of each phonological wordform.

We focused the analysis on the *lemma* representation of words, which corresponds to the *canonical form* or *root form* of each word. Lemmas are commonly considered to constitute the "core" lexicon of a language, and have been used to investigate various structural properties of lexical systems, such as phonological similarity and ambiguity among wordforms (5, 6), distributional regularity of homophones (7), functional load of phonemic (8) and featural contrasts (9).

Computation of functional load of syllable onset and coda

Functional load describes the extent to which a language makes use of a particular phonological unit or structure to distinguish words from one another (10–12). Traditionally, the assessment of functional load has been focused on the contribution of individual phoneme pairs (e.g., /b/-/p/ or /o/-/i/) to distinguish minimal word pairs. Counting minimal pairs that involve each phoneme pair consists of one, and the simplest, method to measure the functional load of each phoneme pair, among other more complex methods (13, 14). Investigations beyond individual phoneme pairs usually measure the total level of functional load which takes into account all pairs of phonemes that belong to the same category or share a specific property (9, 14). It is noteworthy that, despite of different ways to calculate functional load, its assessment has always been restricted to word pairs that contrast with each other by substituting of a single phoneme (i.e., minimal pairs). That is, the definition of *minimal pairs*, on which the measurement of functional load is based, is different from the definition of *phonological neighbors*, which include word pairs differ from each other by either adding, deleting or substituting a phoneme (15).

Our measurement of functional load does not take into account word frequency. This measurement is commonly referred to as "type-based" as opposed to the "token-based" measurement which takes into account word frequencies. In lexicon research, several studies conducted type-based analyses to explore various kind of wordform regularities, including phonological similarity among wordforms (5), the occurrence of homophony in lexicons (7), and the functional load of consonants and vowels (14). One study that examined the functional load of consonants and vowels employed both type-based and token-based analyses and found quantitative and no qualitative differences between results from the two analyses (14). In addition, the type-based measurement also allowed for a more straightforward examination of phonological and phonotactic underpinnings for the onset bias in functional load in our follow-up analysis using lexicon simulation. Our lexicon simulation analysis requires generating pseudo-lexicons based on certain phonological/phonotactic regularities of the language and comparing the onset-coda asymmetry in functional load between real lexicon and simulated lexicons. While the sequencing of phonemes in the generated pseudowords can be justified by following combined regularities of CV skeletons and position-specific sound inventories, it is not straightforward to justify the assignment of word frequencies to randomly generated wordforms. A similar point was raised in a previous study, which also conducted a lexicon simulation analysis in examining the occurrence of homophony in lexicons (7).

Skeletal occurrence of syllable onset and coda

For each language, we defined the *skeletal occurrence* of syllable onset and coda as the respective frequencies of occurrence of syllable onset and coda positions across the syllabified CV skeletons of

all the words in the lexicon. One skeletal slot of syllable onset or coda may comprise a single consonant or a multi-consonant cluster. For instance, the word 'parakeet' (/CV-CV-CVC/) contains 3 slots of syllable onset and 1 slot of syllable coda, and the word 'flat' (CCVC) contains 1 slot of syllable onset slot and 1 slot of syllable coda. We first counted the total number of skeletal slots of syllable onset and coda position across all the wordforms. We then computed the frequency of occurrence of each syllabic position, which we referred to as *skeletal occurrence*, by dividing the number of skeletal slots of each position by the sum of skeletal slots from both syllable onset and coda positions.

For the computation of inventory size for syllable onset and coda, we focused on skeletal slots that contain a single consonant (i.e., C-slots) and measured the total number of unique consonants that can occur in these slots at syllable onset and coda. C-clusters represent on average 86.78% of all skeletal slots of syllable onset and coda across the 12 examined languages (see Supporting Text 1). The consonant inventories from these slots thus, reflect the occurrences regularities of consonants at the two syllabic positions across the examined languages. Although multi-consonantal slots (e.g., CC-slots, CCC-slots) contain individual consonants, it is more appropriate to describe the occurrence regularities of consonants in these slots in terms of consonant clusters instead of individual consonants.

Analyses of the variation of functional load across individual onset and coda positions within multisyllabic words

For this analysis, we focused on minimal pairs of words that contain from 2 to 4 syllables. We gave each minimal pair three labels: (i) the syllabic position of the consonants (onset or coda) that differed between the two words; (ii) the total number of syllables of each of the two words (2 or 3 or 4); and (iii) the position of syllable that encapsulated the substituted consonants within the corresponding words (for disyllabic words: 1 or 2; for trisyllabic words: 1 or 2 or 3; and for quadrisyllabic words: 1 or 2 or 3 or 4). We then computed, separately for each word length, the functional load of each syllabic position of each syllable. For instance, the inspection of the English lexicon revealed 14422 minimal pairs between disyllabic words. Among these minimal pairs, 7282 pairs (50.49%) involve substituting consonants at the onset of the first syllable, 324 pairs (2.25%) involve the coda position of the first syllable, 4932 pairs (34.20%) involve the onset position of the second syllable, and 1884 pairs (13.06%) involve the coda position of the second syllable. The percentages between the parentheses indicate the functional load of the onset and coda positions of each of the two syllables of disyllabic words.

We analyzed, separately for each word length, whether the variation of functional load across individual syllable onset and coda positions can be better accounted for by the asymmetry between syllabic positions (SP: onset *vs* coda) or by an overall decay following the global order (GO) of these positions from word beginnings to word endings. We constructed two linear mixed models using Functional Load as the dependent variable: the first model used Syllabic Position (SP) as the fixed effect; the second model used Global Order (GO) as the fixed effect. Both models included Language as the random effect for both intercept and slope of the fixed effect. These analyses were conducted using the 'fitlme' function of MATLAB (R2022a) (The MathWorks, Natick, MA, USA).

Lexicon simulation

Basic (B): for this simulation type, the consonant inventory for each syllabic position is deducted by counting all unique consonants that occur at least once in summarizing all skeletal slots of position. Note that, for all simulation types, the deducted inventories are sensitive to the number of consonants that occupy the skeletal slot of the syllabic position. For instance, the inventory for CC clusters at a given syllabic position is composed of all CC sequences (e.g., /st/, /pl/) that can occur at that syllabic position. Therefore, for Basic lexicons, one inventory was deducted for each type of C clusters at each syllabic position. Finally, since the comparison between the real lexicon and Basic simulated lexicons aimed to examine the impact of pure size differences between onset and coda inventories on the distribution of functional load between the two positions, we did not take into account the relative frequencies of consonants (or consonant clusters) within each basic inventory during the random selection of consonants (or consonant clusters) to fill in skeletal slots of the corresponding syllabic position.

Hierarchical (H): for this simulation type, the consonant inventory of each skeletal slot was deducted by counting all unique consonants in summarizing all skeletal slots from the same syllabic position (onset or coda) that share with the skeletal slot in question the same structural and positional properties of the host syllable, including the CV skeleton of the syllable, the stress and tonal status of the syllable (if used in the language) and the relative position of the syllable within the word.

Hierarchical+Transitional (H+T): for this simulation type, the consonant inventory of each skeletal slot was deducted from all skeletal slots from the same syllabic position that share with the skeletal slot in question both hierarchical (structural, stress, tone, within-word position) and transitional (syllable nucleus) properties. For both Hierarchical and Hierarchical+Transitional lexicons, we took into account the relative frequencies of consonants (or consonant clusters) within each deducted inventory when filling in skeletal slots of syllable onset and coda, in order to maximally simulate the impact of syllable-level phonotactic restrictions on the occurrence of consonants at the two syllabic positions.

In summary, our three series of simulated lexicons encompassed a comprehensive spectrum of phonological/phonotactic regularities that govern the occurrence of consonants and consonant clusters at syllable onset coda positions. It ranged from the Basic simulation which solely considered the size of consonant inventory as the source of regularity, to the Hierarchical+Transitional simulation which incorporated all major syllable-level phonotactic regularities concerning the probabilistic distribution of consonants at the two syllabic positions.

Supporting Text 4: Discrepancies among languages

Although our study showed consistent presence of onset bias in functional load across the 12 examined languages, we observed some discrepancies when examining the quantitative connection between the functional load of syllable onset and coda and the phonological/phonotactic regularities associated with the two positions.

Results from our lexicon simulations revealed two general trends that are followed by the majority of the examined languages. First, for the majority of the languages, the real lexicon exhibited significantly higher onset bias in functional load than all the three series of simulated lexicons, from the ones with the most basic phonological constraints at syllable onset and coda (i.e., Basic lexicons: 9 of 12 languages) to those with the most complex syllable-level phonotactic constraints (i.e., Hierarchical+Transitional lexicons: 7 languages). Secondly, in most languages, the successive application of hierarchical and transitional constraints increases both the onset-coda disparities in inventory size (11 languages) and functional load (9 languages).

A first group of languages that strongly followed both trends consist of Germanic languages (German, English, Dutch, Norwegian, Swedish) and Czech. These languages typically yielded relatively large disparities in the onset bias in functional load between Basic lexicons and the real lexicon (real-minus-basic = 15.26% on average) (Figure 5B), which indicates a clear inefficiency to account for the functional asymmetry between syllable onset and coda by basic scale differences in skeletal and inventorial attributes of the two positions. Moreover, these languages show relatively strong enhancement of the onset bias from Basic lexicons to Hierarchical+Transitional lexicons following the application of syllable-level phonotactic restrictions (Figure 6B, Table S4). In fact, both Germanic languages and Czech are known to intensively utilize syllable coda for word construction, which is reflected in minor onset-coda disparities in skeletal occurrence and large size of the broad consonant inventory at syllable coda (Figure 4). This phonological regularity expectedly leads to smaller onset biases in functional load when the word simulation procedure solely takes into account the broad consonant inventories. Meanwhile, our results showed that the broad consonant inventories associated with syllable coda consistently undergo stronger reductions than those at syllable onset when syllable-level phonotactic constraints are applied (Figure S1), which increases the likelihood for the generated wordforms to differ with each other at syllable onset than coda and, therefore, the level of onset bias in functional load in phonotactically more realistic lexicons. Among these 6 languages, 4 of them still show smaller onset bias in functional load in Hierarchical+Transitional lexicons than the real lexicon, which suggests the presence of additional linguistic mechanisms (e.g., cross-syllable phonotactic regularities, other morphological/semantic rules) for further enhancing the involvement of syllable onset in lexical contrast. Regarding the two exceptions, English exhibits comparable onset biases in Hierarchical+Transitional lexicons with the real lexicon, which suggests that these within-syllable phonotactic regularities could suffice to drive the functional asymmetry between syllable onset and coda. Finally, for German, while the levels of onset bias in functional load from Hierarchical lexicons are still smaller than that from the real lexicon, the levels from Hierarchical+Transitional lexicons surpass the level from the real lexicon by a rather large margin (Figure 6B). These findings indicate that additional linguistic mechanisms, as speculated above, are necessary to bring down the level of onset bias to the level observed in the real lexicon.

A second group of languages that follow the two trends to a lesser degree are the Romance languages (French, Italian, Spanish) and Greek. These languages typically exhibit less utilization of syllable coda for word construction, and hence show stronger onset-coda disparities in phonological attributes (in particular skeletal occurrence) (Figure 4) and functional load (Figure 2). The strongly biased skeletal occurrence of syllable onset over syllable coda across wordforms would unsurprisingly lead to small disparities in the onset bias in functional load between the Basic lexicons and the real lexicon (real-minus-basic = 3.66% on average) (Figure 5B). In particular, since syllable-level phonotactic constraints only affect the size of the consonant inventories associated with individual slots of syllable onset and coda and not the frequencies of occurrence of these slots across the CV skeletons of the wordforms, the application of these constraints exerts limited impact in changing the level of onset bias in functional

load among the three series of simulated lexicons (Figure 6B, Table S4). Among the 4 languages, Spanish showed slightly higher onset biases in Basic lexicons than in the real lexicon (real-minus-basic = -1.83%). The level of onset biases remains stabilized across the two series of phonotactic lexicons with a minor numeric decrease (-0.88% from Basic lexicons to the Hierarchical+Transitional lexicons). This data suggests that, for Spanish, the functional disparity between syllable onset and coda is primarily driven by the skeletal predominance of syllable onset over syllable coda across wordforms, which diminishes the contribution of syllable-level phonotactics in determining the level of the onset bias. The same interpretation would be conceivable for the results of Italian and French, which also exhibit fairly similar levels of onset biases in functional load across the 3 series of simulated lexicons (Figure 6B) with a small numeric decrease (-2.69% from Basic lexicons to the Hierarchical+Transitional lexicons) in the case of Italian and small numeric increase (3.79%) in the case of French (Table S4).

Regarding the remaining two languages, Turkish exhibits the smallest onset bias in functional load among all the examined languages. This observation is most salient for monosyllabic words (Figure 2C), which show almost no onset bias. Coherently, Turkish also shows a relatively small onset advantage in skeletal occurrence (Figure 4A) and equal inventory size at syllable onset and coda (Figure 4B), which indicates strong usage of syllable coda in the construction of phonological wordforms. In fact, Turkish even presents a skeletal preference for syllable coda than syllable onset regarding consonant clusters. Specifically, Turkish only allows consonant clusters to occur at syllable coda, while all the other languages, that use consonant clusters, exhibit greater occurrence of consonant clusters at syllable onset than at syllable coda. The small skeletal and non-existent inventorial advantages of syllable onset over syllable coda resulted in comparable onset biases between the level of functional load across Basic lexicons with the level from the real lexicon. While one might interpret this finding as potential floor effect, given the overall small onset bias in functional load from the real lexicon, the application of hierarchical and transitional constraints increased the onset bias in functional load by more than 10%, which also exceed the level observed in the real lexicon. This latter finding is indeed surprising, given that the applications of these phonological constraints gave little inventory reduction at both syllabic positions (Figure S1). Further investigations on long-term (cross syllable) dependencies of phonemes are needed to understand the mechanisms in the achievement of onset bias in functional load in the real lexicon of Turkish.

Finally, Korean stands out from all the other languages due to an exceptionally small consonant inventory at syllable coda (7 consonants in total), which also leads to the largest onset-coda disparity in inventory size (12 consonants) among all the examined languages (Figure 4B). Results from Basic simulations (using broad consonants inventories at syllable onset and coda) showed that the average onset bias in functional load across the simulated lexicons is higher than the onset bias observed in the real lexicon (Figure 5). This result may be due to the small inventory at syllable coda, which sets a rather low ceiling for the usage of syllable coda for lexical contrast. One possibility is that Korean fully involves all the 7 consonants in constructing and contrasting phonological wordforms in the real lexicon, such that simulations using the broad sound inventories could already achieve realistic distributions of functional load between syllable onset and coda. Our follow-up analyses confirmed this assumption, which showed almost no inventory reduction at syllable coda after the application of hierarchical and transitional constraints (Figure S1). In fact, Korean is the only language that exhibited larger inventory reduction at syllable onset than at syllable coda when more phonotactic restrictions are applied. In particular, the fact that onset-coda disparity in inventory size further decreases after the transitional constraints are applied indicates that syllable nucleus exerts stronger constraints on consonant occurrence at syllable onset than at syllable coda, which diverges from the other examined languages. This observation is in line with previous studies that demonstrated stronger statistical connection between onset and nucleus than between nucleus and coda in Korean (16), which resonates with view that Korean syllables adopt a body-coda structure rather than a more frequently assumed onset-rhyme structure (17). Therefore, in the case of Korean, while the substantially small consonant inventory at syllable coda limits its functional involvement in lexical contrast overall, which leads to an overall onset bias in functional load, the potentially alternative internal structure of Korean syllables allowed the language to fully exploit the consonant inventory at syllable coda. These two specific phonological

regularities of Korean syllables govern the distribution of functional load between the onset and coda positions.

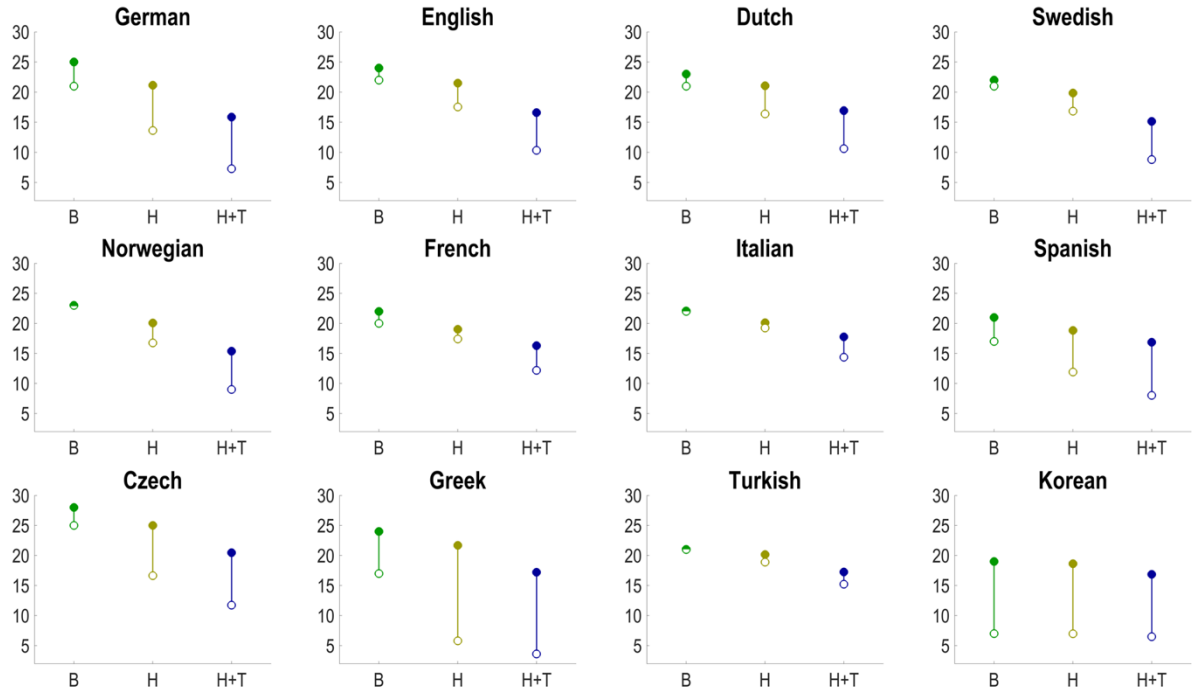

Figure S1. Average inventory size of syllable onset and coda positions under different phonotactic constraints. Each graph presents the averaged inventory size of consonant across all skeletal slots containing single consonant cluster (C cluster) at syllable onset (filled circle) and syllable coda (open circle) which are determined under three phonotactic conditions. *Basic* (B): under this condition, all skeletal slots of C clusters at syllable onset or coda were associated with a common, broad, consonant inventory of that syllabic position, which is summarized across all slots of C-clusters of that syllabic position. *Hierarchical* (H): under this condition, each skeletal slot of C clusters at syllable onset or coda was associated with a consonant inventory that was specific to positional and structural properties of the syllable that hosted the slot. *Hierarchical+Transitional* (H+T): under this condition, each skeletal slot of C clusters at syllable onset or coda was associated with a consonant inventory that was specific to positional and structural properties and the nucleus of the host syllable. (See Supporting Text 3 for more detailed description of how sound inventories under each phonotactic condition are defined.) On average, C clusters account for 86.78% of all skeletal slots at syllable onset and coda positions across wordforms in the 12 examined databases (see Supporting Text 1). Thus, the inventory size difference of C clusters among the three phonotactic conditions reflects the overall impact of different streams phonotactic restrictions on the occurrence of consonants at syllable onset and coda.

Table S1: Functional load of syllable onset (FL_{onset}) and coda (FL_{coda}) as well as the onset bias (FL_{onset} - FL_{coda}) for each language. All measurements are presented in percentages.

Language	FL_{onset}	FL_{coda}	FL_{onset} - FL_{coda}
German	76.53	23.47	53.06
English	67.54	32.46	35.08
Dutch	79.08	20.92	58.17
Swedish	77.54	22.46	55.08
Norwegian	80.10	19.90	60.21
French	77.43	22.57	54.85
Italian	92.67	7.33	85.33
Spanish	94.79	5.21	89.57
Czech	85.65	14.35	71.30
Greek	98.22	1.78	96.45
Turkish	61.33	38.67	22.66
Korean	86.21	13.79	72.41

Table S2. Frequency of occurrence (FO) and functional load (FL) of four types of consonant clusters with different number of consonants (C, CC, CCC, CCCC). Both FO and FL are presented in percentage.

Language	C		CC		CCC		CCCC	
	FO	FL	FO	FL	FO	FL	FO	FL
German	85.68	90.43	13.49	9.45	0.77	0.12	0.05	0
English	85.15	91.59	14.11	8.30	0.74	0.11	0.0007	0
Dutch	82.72	88.72	16.12	11.13	1.16	0.15	0.007	0
Swedish	80.09	86.02	18.59	13.76	1.31	0.22	0.01	0
Norwegian	81.66	87.92	17.10	11.75	1.24	0.33	0.005	0
French	84.02	89.08	15.55	10.82	0.41	0.10	0.009	0
Italian	90.61	95.33	9.00	4.56	0.39	0.10	0.0005	0
Spanish	89.89	96.46	9.89	3.43	0.22	0.10	–	–
Czech	80.36	83.88	18.49	15.34	1.12	0.78	0.03	0
Greek	82.79	87.89	16.43	12.04	0.76	0.07	0.009	0
Turkish	98.43	98.45	1.56	1.55	0.005	0	–	–
Korean	100	100	–	–	–	–	–	–

Table S3. Onset bias in functional load (in percentage) averaged across 50 simulated lexicons and their comparisons to the level of onset biases observed in the real lexicon. Δ indicates the difference in the amount of onset bias between real and the simple lexicons (real-minus-simulated). *z* indicates the statistic of the *z*-test for the comparison between onset bias observed in the simulated lexicons to the one observed in the real lexicon. *p*-values are corrected for multiple comparison using Holm–Bonferroni method.

Language	Simulated lexicon		Real lexicon			
	Mean	SD	Value	Δ	*z*	*p*
German	39.73	1.40	53.06	13.33	9.55	**<.001**
English	20.79	0.83	35.08	14.29	17.14	**<.001**
Dutch	40.51	0.83	58.17	17.66	21.23	**<.001**
Swedish	45.47	0.94	57.79	9.61	10.18	**<.001**
Norwegian	53.37	0.82	55.08	6.83	8.35	**<.001**
French	49.61	0.54	54.85	5.24	9.64	**<.001**
Italian	80.99	0.90	85.33	4.33	4.82	**<.001**
Spanish	91.41	0.58	89.57	-1.83	-3.17	**<.01**
Czech	39.05	1.96	71.30	32.25	16.45	**<.001**
Greek	89.52	0.72	96.45	2.43	9.60	**<.001**
Turkish	23.34	1.18	22.66	-0.68	-0.57	0.57
Korean	74.83	0.40	72.41	-2.42	-6.08	**<.001**

Table S4. Onset bias in functional load (in percentage) in the real lexicon (R) and the onset biases in each type of simulated lexicons averaged across the 50 simulated lexicons for each type (B: Basic; H: Hierarchical; H+T: Hierarchical+Transitional).

Language	B	H	H+T	R
German	39.73	47.14	62.96	53.06
English	20.79	23.57	34.18	35.08
Dutch	40.51	46.55	54.09	58.17
Swedish	45.47	49.74	52.79	55.08
Norwegian	53.37	55.91	56.67	60.21
French	49.61	49.75	53.40	54.85
Italian	80.99	79.33	78.30	85.33
Spanish	91.40	90.98	90.52	89.57
Czech	39.05	60.38	65.44	71.30
Greek	89.52	93.76	93.88	96.45
Turkish	23.34	34.28	34.04	22.66
Korean	74.28	73.49	71.57	72.41

Table S5. Onset bias in functional load (in percentage) averaged across 50 Hierarchical+Transitional (H+T) lexicons and their comparisons to the level of onset biases observed in the real lexicon. Δ indicates the difference in the amount of onset bias between real and the simulated lexicons (real-minus-simulated). *z* indicates the statistic of the *z*-test for the comparison between onset bias observed in the simulated lexicons to the one observed in the real lexicon. *p*-values are corrected for multiple comparison using Holm–Bonferroni method.

Language	H+T lexicon		Real lexicon			
	Mean	SD	Value	Δ	*z*	*p*
German	62.96	0.82	53.06	-9.90	-12.14	**<.001**
English	34.18	0.73	35.08	0.90	1.23	0.22
Dutch	54.09	0.49	58.17	4.08	8.40	**<.001**
Swedish	52.79	0.72	57.79	2.29	3.19	**<.001**
Norwegian	56.67	0.55	55.08	3.54	6.46	**<.01**
French	53.40	0.66	54.85	1.45	2.18	0.058
Italian	78.30	0.83	85.33	7.03	8.51	**<.001**
Spanish	90.52	0.38	89.57	-0.94	-2.47	0.055
Czech	65.44	0.93	71.30	5.86	6.30	**<.001**
Greek	93.88	0.38	96.45	2.57	6.76	**<.001**
Turkish	34.04	1.55	22.66	-11.37	-7.34	**<.001**
Korean	71.57	0.34	72.41	0.84	2.46	**<.05**

SI References

1. G. K. Zipf, *Human behavior and the principle of least effort* (Addison-Wesley, 1949).
2. C. J. Davis, M. Perea, BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behav. Res. Methods* **37**, 665–671 (2005).
3. A. Duchon, M. Perea, N. Sebastián-Gallés, A. Martí, M. Carreiras, EsPal: One-stop shopping for Spanish word properties. *Behav. Res. Methods* **45**, 1246–1258 (2013).
4. Ö. Koşaner, Ç. C. Birant, Ö. Aktaş, Improving Turkish Language Training Materials: Grapheme-to- Phoneme Conversion for adding Phonemic Transcription into Dictionary Entries and Course Books. *Procedia - Soc. Behav. Sci.* **103**, 473–484 (2013).
5. I. Dautriche, K. Mahowald, E. Gibson, A. Christophe, S. T. Piantadosi, Words cluster phonetically beyond phonotactic regularities. *Cognition* **163**, 128–145 (2017).
6. S. T. Piantadosi, H. Tily, E. Gibson, The communicative function of ambiguity in language. *Cognition* **122**, 280–291 (2012).
7. S. Trott, B. Bergen, Why do human languages have homophones? *Cognition* **205**, 104449 (2020).
8. A. Wedel, S. Jackson, A. Kaplan, Functional Load and the Lexicon: Evidence that Syntactic Category and Frequency Relationships in Minimal Lemma Pairs Predict the Loss of Phoneme contrasts in Language Change. *Lang. Speech* **56**, 395–417 (2013).
9. A. Martin, S. Peperkamp, Asymmetries in the exploitation of phonetic features for word recognition. *J. Acoust. Soc. Am.* **137**, EL307–EL313 (2015).
10. C. F. Hockett, The Quantification of Functional Load: A Linguistic Problem,. *U.S. Air Force Memo. RM-5168-PR.* (1966).
11. D. Surendran, P. Niyogi, “Measuring the usefulness (functional load) of phonological contrasts” (Technical Report TR-2003-12, Dept. of Comp. Science, Univ. of Chicago, 2003).
12. A. Wedel, A. Kaplan, S. Jackson, High functional load inhibits phonological contrast loss: A corpus study. *Cognition* **128**, 179–186 (2013).
13. L. Van Severen, *et al.*, The relation between order of acquisition, segmental frequency and function: The case of word-initial consonants in Dutch. *J. Child Lang.* **40**, 703–740 (2013).
14. Y. M. Oh, C. Coupé, E. Marsico, F. Pellegrino, Bridging phonological system and lexicon: Insights from a corpus study of functional load. *J. Phon.* **53**, 153–176 (2015).
15. M. S. Vitevitch, P. A. Luce, Phonological Neighborhood Effects in Spoken Word Perception and Production. *Annu. Rev. Linguist.* **2**, 75–94 (2016).
16. Y. Lee, M. Goldrick, The emergence of sub-syllabic representations. *J. Mem. Lang.* **59**, 155–168 (2008).
17. Y. B. Yoon, B. L. Derwing, A language without a rhyme: Syllable structure experiments in Korean. *Can. J. Linguist. Can. Linguist.* **46**, 187–237 (2001).