Richard Wiese* and Paula Orzechowska

Structure and usage do not explain each other: an analysis of German word-initial clusters

https://doi.org/10.1515/ling-2020-0030
Received March 5, 2020; accepted December 19, 2022; published online August 31, 2023

Abstract: The present study focuses on German word-initial consonant clusters and asks whether feature-based phonotactic preferences correlate with patterns of type and token frequencies in present-day usage. The corpus-based analyses are based on a comprehensive list of such clusters, representing current usage, and on a number of feature-based phonotactic preferences. Correlating the variables by means of a correlation analysis and a regression analysis leads to a number of observations relevant to the general topic of featural-segmental structures versus usage. First, out of eighteen correlations between (raw and logarithmic) type and token frequencies, and preferred feature patterns, only one significant correlation was found. Second, a regression analysis led to similar results: out of thirteen variables tested, only two contribute to logarithmic type and token frequencies. Only a limited set of cluster properties investigated in the present paper constitutes a relevant predictor of frequency measures. The study thus demonstrates, in accordance with other recent evidence, that preferred phonetic/phonological structures and their usage frequency constitute two separate domains for which distributions may not have to coincide.

Keywords: frequency; German; phonetic features; phonotactics; usage-based phonology

1 Introduction

Combinations of consonants, either prevocalic or postvocalic, are generally regarded as disfavored in relation to single consonants. Although this aversion is open to several different interpretations (e.g., in terms of structural markedness or articulatory difficulty), it constitutes a well-known fact about the phonetic and/or

***Corresponding author: Richard Wiese**, Institut für Germanistische Sprachwissenschaft, Philipps-Universität Marburg, 35041 Marburg, Germany, E-mail: wiese@uni-marburg.de. https://orcid.org/0000-0001-7333-1589
Paula Orzechowska, Faculty of English, Adam Mickiewicz University, Poznań, Poland, E-mail: paulao@amu.edu.pl. https://orcid.org/0000-0002-0416-7305

phonological structure of the world's languages. This issue has been approached from different perspectives such as linguistic universals (Greenberg 1978; Maddieson 2013), phonological theory (Hall 1992; Rochoń 2000) and linguistic development (Ingram 1978; Jakobson 1968; Levelt et al. 1999). The complexity of sequences of consonants in German has been investigated in terms of phonological theory (e.g., Féry 2003; Vennemann 1982), acquisition in child language (e.g., Leopold 1949; Lleo and Demuth 1999; Ott et al. 2006; Yavaş et al. 2018), phonological features and rankings of goodness (e.g., Orzechowska and Wiese 2011, 2015), and online processing (e.g., Celata et al. 2015; Domahs et al. 2009; Korecky-Kröll et al. 2014; Ulbrich et al. 2016).

Basically, two "philosophies" exist for the analysis of linguistic properties such as consonant clusters. One view evaluates clusters on the basis of structural properties expressed by universals (Greenberg 1978) and phonological principles (e.g., sonority; Parker 2012b, 2017). A second view starts from a usage-based perspective, according to which preferred structures tend to be most frequent, and structures, in fact, emerge from their (frequency of) use (Bybee 2001). Utterance tokens lead the language user to build up representations and abstractions which are increasingly complex. Under this view, representations and abstractions are closely tied to tokens, often called "exemplars", and their frequencies, especially in terms of lexical frequencies (Bailey and Hahn 2001). According to this interpretation, clusters obeying sonority principles are considered well-formed because language users are generally exposed to them (in perception and/or production) more often.

More recently, however, studies have questioned the exclusive role of one or the other view; see further discussion in Section 4. In the present study, we pursue the question of whether frequencies and structural preferences are closely related. On the one hand, a possible result would be that structure and usage do not correlate, on the other hand, there might be a strong relationship between them. Evidence in favor of the first approach was provided by Zydorowicz et al. (2016), and Orzechowska and Zydorowicz (2019) who analyzed correlations between different measures of markedness and frequency in Polish and English clusters, and demonstrated that there is no clear relationship (i.e., no or weak correlation) between principles of phonotactic markedness and logarithmically transformed type and token frequencies. More specifically, the majority of initial and final cluster types with the highest frequency of occurrence were shown to be structurally dispreferred in terms of different distance measures.

Similar results were documented in production studies. Jarosz (2017) argued that there are discrepancies between lexical statistics and structural preferences: while the lexical statistics of Polish onset clusters do not demonstrate a preference for a sonority-based profile, children acquiring Polish show a preference for the sonority principle. In a similar vein, Orzechowska (2019: Ch. 5) investigated factors affecting

spontaneous speech processes in Polish adults and observed that the reduction rate of clusters at word edges is primarily affected by cluster frequency, followed by manner of articulation properties such as stridency and continuance. Some of these properties outrank well-formedness in terms of the a sonority slope as a criterion motivating cluster modification. The relationship between frequency and cluster structure is also specific to a cognitive function. In a series of reaction time studies, Orzechowska (2019: Ch. 4) showed that the interaction between usage and cluster structure in terms of the manner of articulation features does not affect response latencies, while place of articulation distances facilitate processing. In contrast, usage and the sonority profile of clusters constitute a driving force at the metalinguistic level when phonological judgment is made.

As we have shown, research related to structure and frequency has focused on some type of overarching well-formedness conditions, sonority in particular. It must be noted that the present analysis employs the *Sonority Sequencing Generalization* as one of numerous factors potentially correlating with frequency measures. Other factors are related to cluster-internal or segment-internal properties, mainly expressed in terms of features. In this sense, the present work coincides with analyses in which sub-segmental cues are key for phonotactic modeling and representations (e.g., Hayes and Wilson 2008; Hirst 1980).

In this contribution, we ask a related question: Is there a relationship between feature-based phonotactic preferences and patterns of type and token frequencies? For the purpose of the present study, we start with a comprehensive list of word-initial clusters of German and then run statistical analyses to determine whether a rich set of phonetic/phonological descriptive parameters is correlated with various frequency measures. The nature of our approach is exploratory. We do not argue in favor of an *a priori* assumption regarding the relationship but test the existing approaches.

2 Deriving phonotactic preferences

2.1 German word-initial clusters

Descriptions of German agree that up to three consonants can be found in word-initial position (Hall 1992; Kohler 1990; Meinhold and Stock 1980; Wiese 1988). Viewed from a cross-linguistic perspective, present-day German thus allows for a moderate degree of phonotactic complexity (see a classification in Maddieson 2013). However, the precise list of such clusters differs between descriptions.

For the present work, we used a list originally collected for the analysis in Orzechowska and Wiese (2015), who compiled a set of 56 clusters from an extensive

Table 1: List of word-initial German clusters.

Size	Cluster types
CC	bj̯, bl, bʁ, dʁ, fj̯, fl, fʁ, gl, gm, gn, gʁ, kl, km, kn, kʁ, ks, kv, pfl, pfʁ, pl, pn, pʁ, ps, sf, sk, sl, sm, sn, sp, sʁ, st, stʂ, sv, ʃk, ʃl, ʃm, ʃn, ʃp, ʃʁ, ʃt, ʃv, tj̯, tm, tʁ, tsv, tv, vl, vʁ
CCC	skl, skʁ, skv, spl, stʁ, ʃpl, ʃpʁ, ʃtʁ

corpus of newspaper texts (*Leipziger Wortschatz-Portal*, see Section 2.3 for a detailed description), as evidenced in Table 1. This inventory is thus more comprehensive than lists given in the published accounts mentioned above. The resulting inventory includes clusters found in assimilated loans, rare words and proper nouns. Examples of such clusters include /pn/ in *Pneu*, /sm/ in *Smog*, /fj̯/ in *Fjord* and /ʃk/ in *Schkopau*.

The rationale behind the inclusion of such items is the intention to study the usage of present-day German on a broad empirical basis, rather than to base the choice of clusters on a preconceived notion of what constitutes "German" in a historical or norm-based sense, or what in fact constitutes a fully-fledged cluster (e.g., see the treatment of /s/C and /ʃ/C sequences in Goad 2011; Goad and Rose 2004; Yavaş et al. 2008). As a result, the analysis starts with the above inventory of 56 word-initial clusters, among which 48 are bisegmental (CC) and 8 are trisegmental (CCC). Except for treating the affricates /pf ts/ as monosegmental, we take no stand here on the structure of such clusters in terms of, for instance, extrasyllabicity or sub-syllabic categories: it is quite possible that the clusters are diverse in this respect.

2.2 Feature-based analysis of clusters

Based on an in-depth description of phonetic and phonological features of clusters in Table 1, Orzechowska and Wiese (2015) identified a number of new phonotactic constraints for the cluster set in German. The goal was to find an empirically grounded method of ranking CCs and CCCs in terms of these constraints. The authors proposed a method involving an analysis of sub-segmental properties of consonants within clusters, leading to a set of 15 constraints (called *parameters*).[1] This set was grouped into four broad dimensions pertaining to *Complexity*, *Place of articulation, Manner of articulation* as well as *Voicing*. Table 2 provides a complete list of the parameters. Each parameter is instantiated by a range of patterns (with further

1 Additional parameters may be conceived, but we emphasize that the present analysis covers a wide range of possible parameters. The list is an extension of the set used in previous studies from the authors.

Table 2: Dimensions and parameters of cluster descriptions, from Orzechowska and Wiese (2015).

Dimensions	Parameters	Observed patterns
Complexity	(1) Size	CC or CCC
	(2) Compositionality	fully compositional
	(3) Identity avoidance	total avoidance
Place of articulation	(4) Distance	distances 0–6
	(5) Labial C	0 or 1 labials
	(6) Coronal C	0, 1 or 2 coronals
	(7) Dorsal C	0, 1 or 2 dorsals
	(8) Initial C	labial, coronal or dorsal
	(9) Final C	labial, coronal or dorsal
Manner of articulation	(10) Distance	distances 0-4
	(11) Increase in opening	increase, decrease, plateau or mixed
	(12) Obstruent C	1 or 2 obstruents in CC, 2 or 3 in CCC
Voicing	(13) Initial C	voiced or voiceless
	(14) Final C	voiced or voiceless
	(15) Agreement	total, partial or no agreement

details given in Appendices B and C). The parameters were based on features used in the classification of consonants by the *International Phonetic Association* (2007). While there are many alternative featural descriptions (e.g., see Hall 2001; McCarthy 1988), such a list of articulatory phonetic features ensures a theory-neutral description of consonants and their clustering, as much as this seems feasible.

The dimension of *Complexity* is represented by three parameters, two of them reflecting phonotactic universals. *Size* specifies the number of consonants in a string and is related to the basic CV syllable as proposed in Greenberg (1978). In German, the parameter has two possible patterns; clusters composed of two and three segments (e.g., /kl/: CC; /ʃpʁ/: CCC). *Compositionality,* related to Greenberg's (1978) principle of 'resolvability', states that clusters are decomposable into shorter constituents with independent existence. Thus a CCC (e.g., /ʃpʁ/) is fully compositional if it is formed of two CCs (/ʃp/ and /pʁ/) and three individual consonants, all of which are attested in the phonological system of the language. The third parameter, *Identity avoidance*, refers to the occurrence of identical segments in a cluster and is an instance of the *Obligatory Contour Principle* (Yip 1988). For *Compositionality* and *Identity avoidance,* only one value is available as all clusters in German are composed of existing (i.e., fully compositional) and non-identical consonants (i.e., avoid identity).

Within the *Place of articulation* (POA) dimension, the *Distance* parameter captures the place distinctions in the IPA description of consonants and their ordering from bilabial to glottal. The distance equal to one holds between

consecutive places on the scale: bilabial – labio-dental – alveolar – post-alveolar – palatal – velar – uvular. For example, the distance between /p/ and /l/ in /pl/ has the value of 2, while the smallest and the largest distances of 0 and 6 are assigned to clusters such as /sn/ and /bʁ/, respectively. The parameter is related to phonotactic models using place distances (e.g., Dziubalska-Kołaczyk 2019). The remaining parameters specify either the number of segments with a particular feature (labial, coronal, dorsal) or the feature's position in a cluster (initial, final). For instance, the /skʁ/ cluster is assigned the following values: labial = 0, coronal = 1, dorsal = 2 for parameters 5 through 7; initial C = coronal and final C = dorsal for parameters 8 and 9.

Similarly, the *Manner of articulation* (MOA) parameters called *Increase in opening* and *Distance* evaluate clusters with respect to degrees of articulatory opening between adjacent consonants and the manner distinctions on the following scale: plosive – affricate – fricative – nasal – liquid. These parameters have an obvious relation to sonority-based principles such as the *Sonority Sequencing Generalization* (e.g., Selkirk 1984) or the *Dispersion Principle* (Clements 1990). Thus the parameters determine, for example, a cluster's rise or fall in sonority as well as its sonority distance (e.g., /kl/: increase, distance 4; /ʃp/: decrease, distance 2). All CCCs in German combine the two articulatory gestures, while their distance is a mean of the constituent distances (e.g., /spl/: mixed, distance (2 + 4)/2 = 3). The last manner parameter counts the number of *Obstruent C* in a sequence (e.g., /skl/: 2; /tm/: 1). The calculations are performed separately for CCs and CCCs to adequately embrace the proportion of obstruents and sonorants in a cluster.

Finally, the *Voicing* dimension covers the position of voiced and voiceless segments in a cluster. More specifically, *Initial C* and *Final C* specify the feature at cluster margins (e.g., /fj/: voiceless initial C and voiced final C), while *Agreement* states the voicing profile throughout a cluster (e.g., /fj/: no agreement). A detailed exposition and justification of the 15 parameters in Table 2 is offered in Orzechowska and Wiese (2015).

Given the parameters, the first step of the analysis involves tagging each cluster with the pattern it represents. Next, the number of clusters that adhere to a particular pattern is calculated. The numbers are then transformed into percentages that are calculated over the total number of clusters in the dataset. The percentages express the degree to which a given parameter holds in the inventory of clusters under scrutiny, revealing whether it represents a common (preferred) or rare (dispreferred) pattern.[2]

2 Transforming the percentage values further into an ordinal scale would mean a loss of information, as it would not make it possible to compare the degrees of parameter preferability.

Let us illustrate the computation procedure on the example of /fj/.[3] The cluster is composed of two consonants and is therefore labeled '2' on parameter (1) *Size*. The cluster inventory embraces 48 CC types, which corresponds to 86 % (=48 × 100/54) of all clusters under investigation. Thus, /fj/, similarly to each 2-member cluster, is assigned a percentage score equal to 0.86. Moreover, since 46 types (82 %) start with a voiceless consonant and 46 types end with a voiced one, /fj/ scores 0.82 on two parameters: (13) *Voicing in initial C* and (14) *Voicing in final C*. However, as a sequence of two fricatives, /fj/ displays the manner distance equal to 0 (parameter (10) *Manner of articulation distance*), which is represented by only 4 cluster types (=4 × 100/54 = 7 %) in the inventory, scoring 0.07. On this basis, we can state that in word-initial position German displays strong preferences for shorter clusters (86 %), whose constituents consonants differ in voicing (82 %) but disfavors plateaus (7 %).

As was mentioned earlier, parameters (2) *Compositionality* and (3) *Identity avoidance* are represented by a single pattern. Since all clusters in German are fully compositional and do not contain identical segments, each cluster is assigned a score equal to 1 (100 %). These parameters are included neither in Appendices B and C nor in the statistical analyses.

Summated individual percentage scores for a cluster constitute the *composite feature score* (Σ). Such a score assigned to all clusters reflects the degree to which the criteria in Table 2 are met. Next, the arithmetic mean ($\bar{x}$) is calculated over this sum in order to arrive at an overall evaluation of clusters. Composite features scores derived for the present dataset range between 38 % (for the lowest-scoring /skv/) and 56 % (for the highest-scoring /ʃʁ/), with a median of 48 %. These values make it possible to estimate the relative preferability of clusters. For instance, as for /fj/, the composite score Σ equals 5.58 ($\bar{x}$ = 43), which suggests that /fj/ represents a cluster in the mid-range of cluster preferability. Averaging over the values found for parameters (1–15) reveals the degree to which the German clusters in Table 1 follow the proposed set of parameters. The composite feature scores and the means are given in the final two columns of Appendix C.

Along with the composite feature score, the analyses to follow also look at individual constituent parameters that contribute to Σ. For this purpose, we consider two types of statistical analyses.

First, we provide a set of correlations between different frequency measures and selected parameters that are associated with the structural composition of clusters, namely *Size, Increase in opening* and *Composite feature score. Size* (CC, CCC) is an obvious structural property of clusters: starting with the work by Jakobson (1962), it

3 German palatal voiced consonant varies between an approximant [j] and a fricative [j]; see pronouncing dictionary Duden (1990) and Wiese (2000, 236–238). In the present contribution, we treat it as a fricative.

has been assumed that the markedness of clusters increases with the number of adjacent consonants, where a single consonant constitutes the unmarked case. *Increase in opening* characterizes the type of articulatory constriction from the first to the last consonant in an initial cluster (i.e., increase, decrease, plateau, mixed). *Increase in opening* is thus a close approximation to the concept of sonority. The *Composite feature score* Σ expresses the overall degree of cluster preferability in terms of a summation of all structural feature patterns.

Next, following this analysis, we investigate the relevance of selected parameters suited for regression with a view to determining whether, and to what extent, various place, manner and voice properties are related to cluster frequency in German. To reach this goal, we use the wide range of preferences presented in Table 2 and values given in Appendix B (with the exception of *Compositionality* and *Identity avoidance*).

2.3 Frequency measures

For measuring frequencies of linguistic items, a number of alternatives are available. First, types can be distinguished from tokens. Word types refer to individual words starting with a particular cluster, while word tokens represent the number of occurrences of such words in a corpus. Secondly, frequencies can be taken at face value or can be transformed logarithmically.

Most importantly, present-day corpora allow for large-scale counts. Even with a wide-ranging list of German clusters, it is possible to establish frequency counts. The frequencies to be used here were derived from the *Leipziger Wortschatz-Portal* database (www.wortschatz.uni-leipzig.de). This corpus of newspaper texts of present-day German comprised, at the time of access, 172 million tokens representing 1.65 million of word types. Words were transcribed automatically by the *Festival* software. Items containing word-onset clusters were extracted manually. Entries with very low token frequencies (<100) were excluded after visual examination because they were usually caused by spelling variants, spelling errors or misparsings by the software. This cut-off point also resulted in eliminating some rare but potential clusters such as /pt/ in *Ptah* and *Ptashnikov*.

On the basis of this corpus, the frequency counts listed in (1) were at our disposal. Raw token frequencies in the corpus ranged from 47 to 1.9 million occurrences (with those below 100 excluded). Raw type frequencies were defined by the number of word types per cluster, and ranged from 1 to 1 261. Logarithmic transformations of both token and type frequencies (base 10) change the exponential and highly skewed scales into nearly linear ones. In addition, we used rank orders for raw types and

tokens. Overall, the following six frequency measures are employed in the present study.

(1) Frequency counts for clusters
  - token frequencies
    - raw
    - logarithmic
    - rank order
  - type frequencies
    - raw
    - logarithmic
    - rank order

A detailed account of all type and token frequencies for clusters in Table 1 as used in the analyses below is provided in Appendix A.

3 Statistical analyses

3.1 Analysis 1: correlating preferences and usage

This section reports the results of correlating the three structural preferences and frequency counts introduced above. As the optimal arrangement and selection of frequency data for statistical analysis are difficult to determine *a priori*, the six different frequency counts listed in (1) were considered. For each correlation (Pearson correlation coefficients), significance levels were computed. The summary of the results is presented in Table 3 (sample size n = 56 clusters, significance level 5 %).

As can be observed, correlation coefficients are always low, and predominantly non-significant (n.s.). A single statistically significant result is marked with an asterisk '*' (the shaded cell). The widely spread non-existence of a correlation is also illustrated as a scatter plot in Figure 1, which shows, as one of many examples, the relationship between the *Composite feature scores* and *Logarithmic token frequency* (r = 0.1335) from Table 3.

Results in Table 3 demonstrate that only one out of fifteen correlations turns out to be weak (r = 0.2268) but just significant, namely that between *Logarithmic type frequency* and *Size*. At present, we do not offer an interpretation for this single case.

We also computed correlations between two measures of usage, *Raw token frequencies* and *Raw type frequencies*. The results are presented as a scatter plot in Figure 2. In contrast to the previous results comparing usage and structure, this correlation is highly significant (r = 0.810, p-value = 0).

Table 3: Correlating structural preferences and frequencies.

Frequencies	Scores		
	Composite feature score	**Size**	**Increase in opening**
token frequencies:			
Raw	$r = 0.0101$ $p = 0.4706$ n. s.	$r = 0.1685$ $p = 0.1072$ n. s.	$r = 0.1043$ $p = 0.2221$ n. s.
logarithmic	$r = 0.1335$ $p = 0.1634$ n. s.	$r = 0.2041$ $p = 0.0657$ n. s.	$r = 0.1916$ $p = 0.0786$ n. s.
rank order	$r = -0.105$ $p = 0.7793$ n. s.	$r = -0.221$ $p = 0.9492$ n. s.	$r = -0.1722$ $p = 0.8978$ n. s.
type frequencies:			
Raw	$r = -0.055$ $p = 0.658$ n. s.	$r = 0.171$ $p = 0.1038$ n. s.	$r = 0.0508$ $p = 0.355$ n. s.
logarithmic	$r = 0.1379$ $p = 0.1554$ n. s.	$r = 0.2268$ $p = 0.0464$*	$r = 0.167$ $p = 0.1093$ n. s.
rank order	$r = -0.1014$ $p = 0.7715$ n. s.	$r = -0.2336$ $p = 0.9585$ n. s.	$r = -0.1534$ $p = 0.8705$ n. s.

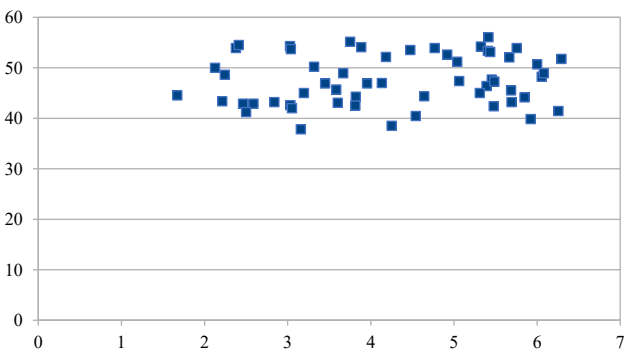

Figure 1: Scatter plot for the composite feature score (*y*-axis) and logarithmic token frequency (*x*-axis).

This finding demonstrates that, within usage, measures are not independent of each other. In other words, types and tokens are largely correlated. The one obvious outlier to be seen in the right-hand bottom of Figure 2 is /tsv/. The cluster has high token frequency due to token-frequent *zwei* 'two' (1 002 205 occurrences of *zwei* out of all 1 968 112 /tsv/ occurrences), but it is found in only 235 word types. This disproportion contrasts with the vast majority of clusters as they display a monotonic increase in both types and tokens. This is also true for the top-most

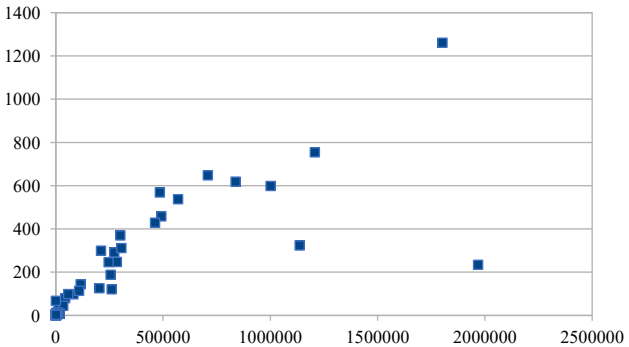

Figure 2: Scatter plot for frequency measures; raw types (*y*-axis) and raw tokens (*x*-axis).

cluster /ʃt/, which has the highest type and token frequencies (1 261 word types, 1 802 222 word occurrences).

3.2 Analysis 2: regression analysis of structures and usage

In order to investigate the relationship between a wider range of structural and phonological parameters and various frequency measures in a different manner, we also performed a linear regression analysis using the *R* software (v. 4.1.0) (R Core Team 2020 online) and the *lmtest* package (Hothorn et al. 2021). The key question here was whether frequency can be estimated from the contribution of phonetic/phonological properties. Thus, *Raw type frequency*, *Raw token frequency*, *Logarithmic type frequency* and *Logarithmic token frequency* served as dependent variables. Independent variables were the feature-based parameters presented in Table 2. A complete list of variables along with their levels (numerical or categorical) is given in Appendix B. Again, parameters (2) and (3) from Table 2 are not included in the regression models as they do not distinguish between clusters at all.

At the outset, we performed *dummy coding* to change each categorical variable into a dichotomous variable. Next, we tested a number of models. The best models for types and tokens were selected on the basis of stepwise regression (function: *step*) using the Akaike Information Criterion (AIC). These best models were obtained for logarithmic tokens and types, which are presented in Tables 4 and 5, respectively. Since models including raw frequencies yielded worse results in terms of higher AIC values, they will not be discussed in the sections to follow.

In both tables, tests of significance give strong premises to infer the significance of two variables: *MOA distance* and *Voicing agreement* between adjacent consonants. These models – as best fitted to the data – also include the following variables: the

Table 4: Linear regression for logarithmic token frequency.[4]

Parametric coefficients:					
	Estimate	**Std. error**	***t*-value**	**Pr (>|*t*|)**	
Intercept	12.5452	3.2180	3.898	0.000294	***
6. Coronal C	0.9913	0.7025	1.411	0.164549	
10. MOA distance	0.9389	0.3959	2.372	0.021672	*
13. Voicing initial C voiceless	−5.0974	3.0241	−1.686	0.098232	.
14. Voicing final C voiceless	4.8716	3.0490	1.598	0.116529	
15. Voicing agreement partial	−3.6750	1.3475	−2.727	0.008839	**
15. Voicing agreement total	−5.7068	2.9389	−1.942	0.057923	.

Residual standard error: 2.864 on 49 degrees of freedom; Multiple *R*-squared: 0.2025, Adjusted *R*-squared: 0.1049; *F*-statistic: 2.074 on 6 and 49 DF, *p*-value: 0.07339.

Table 5: Linear regression for logarithmic type frequency.

Parametric coefficients:					
	Estimate	**Std. error**	***t*-value**	**Pr (>|*t*|)**	
Intercept	4.3382	2.3706	1.830	0.07360	.
6. Coronal C	1.3343	0.6879	1.940	0.05844	.
9. POA final C dorsal	1.6297	0.8305	1.962	0.05564	.
9. POA final C labial	0.9642	0.9331	1.033	0.30675	
10. MOA distance	0.8405	0.3114	2.699	0.00963	**
13. Voicing initial C voiceless	−4.1745	2.2745	−1.835	0.07279	.
14. Voicing final C voiceless	4.3963	2.2469	1.957	0.05635	.
15. Voicing agreement partial	−3.6185	1.0512	−3.442	0.00122	**
15. Voicing agreement total	−4.9838	2.1825	−2.283	0.02697	*

Residual standard error: 2.021 on 47 degrees of freedom; Multiple *R*-squared: 0.285, Adjusted *R*-squared: 0.163; *F*-statistic: 2.342 on 8 and 47 DF, *p*-value: 0.03304.

presence of a coronal consonant, place of articulation of the final consonant in a cluster, voicing of the first and the last consonant in a cluster. In other words, more than half of the thirteen structural parameters did not contribute to the regression models, and of those included, only a few are statistically significant.

4 Our benchmark in selecting the models best fitted with the data constituted the AIC value rather than the *p*-value. A significance level weaker than 0.05 for the whole model (where all factors are considered) does not suggest that individual factors with statistically significant parametric coefficient are irrelevant. The *p*-value above the 0.05 threshold results from including two variables: Coronal C and Voice C-final voiceless, which are nonsignificant (*p*-value = 0.16 and *p*-value = 0.12, respectively). Reduced models in which the two variables were not considered yielded results with higher AIC. Therefore, the results presented in Table 4 are considered to be adequate and fully interpretable.

The results for token and type frequency measures largely coincide. First, average frequency values are higher for larger MOA distances. More specifically, logarithmic frequencies for both types and tokens increase along with an increase in MOA distance. An increase in the sonority distance equal to 1 (i.e., along the scale ranging from 0 to 5) corresponds with an increase of 0.84 in log freq for types and of 0.94 in log freq for tokens. This finding demonstrates that clusters displaying the largest sonority distances are most frequent. Such a group of clusters is represented by plosive + liquid sequences such as /pl bl kl gl/, whose logarithmic frequencies are found in the range between 2.27 and 2.63 for types (with an overall scale 0–3.1), and between 5.39 and 5.66 for tokens (scale 1.67–6.29); see Appendix A.

Second, voice agreement has a negative effect on both dependent variables. *Partial* and *total* agreement lower logarithmic frequencies of types and tokens, although the effect is stronger for *total* agreement (a higher estimate value). This pattern is primarily represented by clusters such as /s/+stop, /ʃ/+stop and C_1C_2 represented by a sequence of a voiced obstruent followed by a liquid or a nasal. The results suggest that clusters displaying no voice agreement, represented by the voiceless + voiced pattern, are positively correlated with the frequency measures, i.e., they have higher log-freq values (Appendices B and C).

As for type frequency, several other variables constitute borderline cases in terms of the statistical significance levels, namely, voicing and place of articulation properties. For instance, an increase in the number of coronal segments in a cluster corresponds with an increase in logarithmic frequency (p-value = 0.05844). A similar effect can be reported for the presence of a dorsal consonant (p-value = 0.05564) and a voiceless consonant (p-value = −0.05635) cluster-finally. In the first case, further conclusions can be drawn. In particular, the presence of a dorsal segment cluster-finally is relevant for type frequency, however, in the absence of a dorsal, there is no difference whether the position is occupied by a labial or a coronal. This observation is related mainly to the presence of /ʁ/ in prevocalic position. In order to determine the relationship between the 'borderline' place and voice variables and frequency, further studies and analyses would be required.

4 Discussion

The goal of the study was to investigate diverse frequency measures of German initial clusters in relation to universals, phonological principles, numerous phonetic/phonological preferences and numerical scores expressing cluster preferability. We conclude that, as much as structure and frequency may overlap, generally they live lives of their own. The statistical analyses have revealed that feature-based preferences, on the one hand, and usage frequency measures, on the other hand, do not necessarily coincide. First, out of 18 correlations presented in Table 3, there is only one

significant, though weak, correlation for the measures chosen: a positive correlation between *Logarithmic token frequency* and *Size*. In contrast, within the frequency measures, there is a high positive correlation between raw types and tokens, as perhaps to be expected. Of course, it should not be excluded on principled grounds that such measures may correlate, even strongly. Thus in light of the present results, it is even more surprising that no significant correlations could be found for the majority of measures used, i.e., between the three structural parameters (i.e., *Composite feature score, Size* and *Increase in opening*) and six different frequency counts.

Second, the logistic regression models reported in Tables 4 and 5 identified two variables that contribute significantly to logarithmic type and token frequencies, namely *MOA Distances* and *Voicing agreement*. Other variables contribute to the models in that they increase the overall goodness of fit of the models to the data at hand (by lowering the AIC values). However, given the set of 13 independent variables included in the regression analyses, most of which do not explain the existing patterns of frequencies, we conclude that structural and frequency distributions overwhelmingly constitute two separate domains of generalizations.

Given the analysis for logarithmic type frequencies in Section 3.2, we can observe that some preferences pertain to contrast enhancement, and are thus functionally motivated. More specifically, higher frequency clusters are likely to exhibit larger sonority distance and variation in terms of voicing. A similar set of preferences in German was identified in Orzechowska and Wiese (2015) and in more recent work. Based on the identical dataset (a list of 56 clusters and four frequency measures derived from the *Leipziger Wortschatz-Portal*), Orzechowska and Dziubalska-Kołaczyk (2022) investigated the relationship between frequency and eight types of phonetic distances pertaining to the place and manner of articulation as well as the sonorant-obstruent contrast. The study revealed a positive correlation between the manner of articulation distance and only one frequency measure: type frequency increases with an increase in the sonority distance between adjacent consonants forming a cluster. This result overlaps with the present analysis, suggesting that sonority distances constitute a relevant phonotactic primitive and motivate the core structure of clusters found in German.

The present results provide indirect evidence in favor of the preference for a sonority-based ordering of segments in German clusters. First, voiceless consonants are less sonorous compared to voiced ones, which motivates the [-voice] + [+voice] + vowel combination in word-initial or syllable-initial position. Second, the preference related to *MOA Distances* lends support to approaches that consider sonority distances in the evaluation of clusters (e.g., Clements 1990; Harris 1983; Parker 2012a). For example, Clements (1990) argues that the most natural C1C2V sequence is represented by obstruent + liquid + vowel due to a steady and gradual rise in sonority from the first consonant towards the vowel. As we have seen, this pattern involves the largest sonority distance and is frequent. However, given the overall findings of the present study, the two domains often follow their own paths and principles.

In some recent work, this view has been explored as well. Basirat et al. (2021) tested the respective roles of attestedness and well-formedness of biconsonantal onsets for speakers of French using a nonce-word acceptability task. Attested clusters were found to be accepted more easily than unattested ones, but independently, items with well-formed clusters were found to be rated higher than items containing ill-formed ones. Furthermore, participants showed individual differences. Note that attestedness was measured in terms of the lexical statistics in the French language (types), and not in terms of frequency of occurrence (tokens) as in the present work.

An EEG experiment reported in Wiese et al. (2017) also demonstrated that both structural factors (more specifically, adherence to the sonority principle) as well as frequency of use shape the online processing of Polish clusters. The experiment showed that the process of learning of nonce words is facilitated by the presence of existent (rather than hypothetical) consonant clusters in a word. The same pattern was observed in a parallel study on German by Ulbrich et al. (2016). Similarly, Jarosz (2017) demonstrated a discrepancy between the lexical statistics of Polish onsets and the fact that Polish children favor onsets following the expected sonority rise. In addition, Silva et al. (2019) presented neurolinguistic evidence on European Portuguese suggesting that well-formedness, on the one hand, and frequency differences between clusters, on the other hand, have different time signatures: EEG effects for the former factor take place earlier than effects for the latter.

It must be borne in mind that the set of preferences identified for initial clusters in this contribution may differ from preferences that are found in other word positions (e.g., final, medial) and that motivate the linguistic behavior of speakers. This idea was explored by Orzechowska (2019) who demonstrated that phonotactic preferences arise from *feature weight,* and that weight is specific not only to cluster position in a word, but also to a linguistic function activated. That is, different sets of features underlie cluster inventories (lexical preferences), while other features facilitate production (articulatory preferences), perception and processing (cognitive preferences), and within these functions, the contribution of specific features depends on cluster position in a word. Therefore, the preferences derived statistically from frequency measures in the present paper are considered to reflect the phonotactic potential of German, and cannot be taken as patterns which hold for language in general. Sonority distances and voice agreement are strongly related to frequency. Stating whether they also affect articulatory, perceptual or cognitive processes in both word positions would require a separate study.

We have also shown that a ranking of clusters based on a large set of parameters does not necessarily provide insights into usage frequency. In turn, constituent parameters employed in the analysis are better predictors of frequency. Under this view, one can argue that higher-order principles such as sonority do not suffice in the study of phonotactics, and that phonological systems might rely on more subtle preferability conditions, expressed in terms of relevant sub-segmental properties. This point was

argued for by Daland et al. (2011). According to their non-word acceptability rating study, sonority effects are predictable from lexicon-based frequency profiles, revealing sonority to be a significant predictor of acceptability for unattested clusters. However, computing the effects of the sonority principle from the lexicon without a universalist/innate principle requires the presence of phonological representations of two kinds: syllabification and sonority-related features. Note that the lexicon used here was a lexicon of types (18, 612 words). A similar result was reported by van de Vijver and Baer-Henney (2012) for the acquisition of German onset clusters. By the same token, Albright (2009) tested probabilistic and feature-based models on phonotactics and showed that natural classes expressed in terms of features and probabilities have the potential of explaining gradient judgment on well-formedness. Albright (2009) observed that token frequencies make little contribution to the predictions of the models as the majority of words in the lexicon are low frequency.

The present analysis complements the results of Daland et al. (2011) as it is based on both types and tokens of German onsets. Additionally, it is consistent with Albright (2009) in that the results for log freq tokens and types in Tables 4 and 5, respectively, are practically identical. Note, however, that the present work is based on lexical and usage statistics, and not on acceptability results.

Naturally, it would be interesting to pursue further studies related to frequency in terms of co-occurrence probabilities. The relevance of likelihood in phonotactic learning was stressed in, among others, Bailey and Hahn (2001) and Vitevitch and Luce (1999), who documented a correlation between segment position in a word and intersegmental co-occurrence, and the accuracy of judgment and response times. This area of research in German phonotactics remains open for future investigation. At this point, when considering structure and frequency, it can be stated that there is no way of determining cause and effect at the synchronic level (see Kiparsky (2008) for arguments on the diachronic perspective). We can only pursue further research on the possible relationship between the two domains. In this respect, we have seen that generalizing measures of well-formedness or rankings of clusters based on such measures are less informative than their constituent sub-measures.

5 Conclusions

The present study has related various measures of phonotactic patterns and preferences for German initial clusters to frequency patterns. The statistical results for correlations between them have revealed that feature-based preferences, on the one hand, and usage frequency measures, on the other hand, constitute two separate (but possibly overlapping) domains. We suggest that there is, in general, no simple path from structure to usage or vice versa. The latter seems to be assumed in some versions of usage-based phonology (among others, Bybee 2001; Pierrehumbert 2001),

according to which language use crucially shapes linguistic structure, or in which structures are emergent properties of usage only. Conversely, structuralist phonology often based markedness judgments on frequency of use, following the principle that "unmarked features occur more frequently than marked features" (Rice 2007: 94). For example, shorter clusters are argued to be unmarked with respect to longer clusters, because the latter are found to be rarer. In fact, the single positive correlation found between *Logarithmic token frequency* and *Size* corroborates this assumption. More generally, however, ties between structural markedness and frequencies must be scrutinized carefully.

Acknowledgments: We would like to thank Bernd Möbius and Barbara Samlowski for their assistance in the extraction of corpus frequencies, the reviewers for their helpful comments, and Andrzej Porębski for statistical help.
Research funding: This research has been financed by the National Science Centre (Poland), under grant no. 2015/18/E/HS2/00066.
Competing interests: The authors have no competing interests to declare.
Data sources: Leipziger Wortschatz-Portal. www.wortschatz.uni-leipzig.de. See the Appendices as well as the Zenodo file, https://doi.org/10.5281/zenodo.7458539.

Appendix A: Raw type and token frequencies, logarithmic type and token frequencies and rank order of type and token frequencies of clusters (CL); ordered according to raw type frequency

CL	Raw type freq.	Log 10 type freq.	Rank type freq.	Raw token freq.	Log 10 token freq.	Rank token freq.
ʃt	1261	31,007,150,866	1	1,802,222	62,558,082,869	2
pʁ	754	28,773,713,459	2	1,207,203	6,081,780,306	3
ʃp	648	28,115,750,059	3	708,589	58,503,944,061	7
gʁ	619	2,791,690,649	4	838,145	59,233,191,585	6
fʁ	599	27,774,268,224	5	1,001,796	60,007,792,933	5
kʁ	569	27,551,122,664	6	485,619	56,862,956,703	10
tʁ	538	27,307,822,757	7	570,454	57,562,206,297	8
bʁ	458	2,660,865,478	8	492,293	56,922,236,605	9
kl	429	26,324,572,922	9	463,396	56,659,522,807	11
ʃtʁ	370	25,682,017,241	10	300,725	54,781,695,336	13
dʁ	324	25,105,450,102	11	1,136,562	60,555,931,317	4
ʃv	312	2,494,154,594	12	306,218	54,860,307,157	12
fl	300	24,771,212,547	13	210,850	53,239,736,054	19

(continued)

CL	Raw type freq.	Log 10 type freq.	Rank type freq.	Raw token freq.	Log 10 token freq.	Rank token freq.
ʃl	292	24,653,828,514	14	273,296	54,366,332,753	15
bl	248	23,944,516,808	15	285,319	54,553,306,932	14
gl	247	23,926,969,533	16	246,355	53,915,613,811	18
ʦv	235	23,710,678,623	17	1,968,112	62,940,498,093	1
pl	188	22,741,578,493	18	256,185	54,085,536,976	17
kv	144	21,583,624,921	19	115,613	50,630,066,707	21
ʃpʁ	126	21,003,705,451	20	204,499	53,106,911,886	20
ʃʁ	121	20,827,853,703	21	260,420	54,156,743,346	16
ʃn	114	20,569,048,513	22	109,475	50,393,149,539	22
ʃm	99	19,956,351,946	23	58,474	47,669,628,034	24
kn	98	19,912,260,757	24	82,839	49,182,348,476	23
sk	80	1,903,089,987	25	43,767	4,641,146,779	25
tj	68	18,325,089,127	26	1069	30,289,777,052	46
pfl	53	17,242,758,696	27	29,843	44,748,424,789	27
st	45	16,532,125,138	28	34,697	45,402,919,261	26
sl	28	14,471,580,313	29	15,198	41,817,864,402	29
ps	27	14,313,637,642	30	13,575	41,327,398,383	30
sm	14	11,461,280,357	31	7646	38,834,342,937	32
sv	14	11,461,280,357	32	9041	39,562,164,692	31
sp	13	11,139,433,523	33	6604	38,198,070,646	33
gn	10	1	34	2822	34,505,570,094	39
ks	10	1	35	3831	3,583,312,152	38
stʦ	9	0.9542425094	36	17,844	42,514,922,146	28
fj	8	0.903089987	37	389	25,899,496,013	48
skʁ	8	0.903089987	38	1125	3,0511525224	43
sn	8	0.903089987	39	2079	33,178,544,893	40
bj	6	0.7781512504	40	4009	36,030,360,563	37
sʁ	6	0.7781512504	41	5656	37,525,094,008	35
stʁ	6	0.7781512504	42	6528	38,147,801,457	34
tv	5	0.6989700043	43	1092	30,382,226,384	44
skv	4	0.6020599913	44	1437	31,574,567,681	42
vʁ	4	0.6020599913	45	1570	31,958,996,524	41
skl	3	0.4771212547	46	1081	3,033,825,694	45
ʃpl	3	0.4771212547	47	693	28,407,332,346	47
gm	2	0.3010299957	48	291	2,463,892,989	50
sf	2	0.3010299957	49	316	24,996,870,826	49
ʃk	2	0.3010299957	50	47	16,720,978,579	56
vl	2	0.3010299957	51	4662	36,685,722,692	36
km	1	0	52	175	22,430,380,487	53
pfʁ	1	0	53	134	21,271,047,984	55
pn	1	0	54	239	23,783,979,009	52
spl	1	0	55	163	22,121,876,044	54
tm	1	0	56	257	24,099,331,233	51

Appendix B: German clusters (CL) with parameter patterns (adapted from Orzechowska and Wiese 2015); ordered alphabetically. Parameters (2) and (3) (*Compositionality, Identity avoidance*) are excluded from the table as all the clusters display a single pattern

CL	1 Size	4 POA dist.	5 Lab. C	6 Cor. C	7 Dor. C	8 POA C-in	9 POA C-fin	10 MOA dist.	12 Increase	11 Obstr.C	13 Voice C-in	14 Voice C-fin	15 Agreement
bj̯	2	4	1	0	1	lab	dor	2	increase	2/2	+	+	total
bl	2	2	1	1	0	lab	cor	4	increase	1/2	+	+	total
bʁ	2	6	1	0	1	lab	dor	4	increase	1/2	+	+	total
dʁ	2	4	0	1	1	cor	dor	4	increase	1/2	+	+	total
fj̯	2	3	1	0	1	lab	dor	0	plateau	2/2	–	+	no
fl	2	1	1	1	0	lab	cor	2	increase	1/2	–	+	no
fʁ	2	5	1	0	1	lab	dor	2	increase	1/2	–	+	no
gl	2	3	0	1	1	dor	cor	4	increase	1/2	+	+	total
gm	2	5	1	0	1	dor	lab	3	increase	1/2	+	+	total
gn	2	3	0	1	1	dor	cor	3	increase	1/2	+	+	total
gʁ	2	1	0	0	2	dor	dor	4	increase	1/2	+	+	total
kl	2	3	0	1	1	dor	cor	4	increase	1/2	–	+	no
km	2	5	1	0	1	dor	lab	3	increase	1/2	–	+	no
kn	2	3	0	1	1	dor	cor	3	increase	1/2	–	+	no
kʁ	2	1	0	0	2	dor	dor	4	increase	1/2	–	+	no
ks	2	3	0	1	1	dor	cor	2	increase	2/2	–	–	total
kv	2	4	1	0	1	dor	lab	2	increase	2/2	–	+	no
pfl	2	1	1	1	0	lab	cor	3	increase	1/2	–	+	no

(continued)

CL	1	4	5	6	7	8	9	10	12	11	13	14	15
	Size	POA dist.	Lab. C	Cor. C	Dor. C	POA C-in	POA C-fin	MOA dist.	Increase	Obstr.C	Voice C-in	Voice C-fin	Agreement
pfʁ	2	5	1	0	1	lab	dor	3	increase	1/2	–	+	no
pl	2	2	1	1	0	lab	cor	4	increase	1/2	–	+	no
pn	2	2	1	1	0	lab	cor	3	increase	1/2	–	+	no
pʁ	2	6	1	0	1	lab	dor	4	increase	1/2	–	+	no
ps	2	2	1	1	0	lab	cor	2	increase	2/2	–	–	total
sf	2	1	1	1	0	cor	lab	0	plateau	2/2	–	–	total
sk	2	3	0	1	1	cor	dor	2	decrease	2/2	–	–	total
skl	3	3	0	2	1	cor	cor	3	mixed	2/3	–	+	partial
skʁ	3	2	0	1	2	cor	dor	3	mixed	2/3	–	+	partial
skv	3	4	1	1	1	cor	lab	4	mixed	3/3	–	+	partial
sl	2	0	0	2	0	cor	cor	2	increase	1/2	–	+	no
sm	2	2	1	1	0	cor	lab	1	increase	1/2	–	+	no
sn	2	0	0	2	0	cor	cor	1	increase	1/2	–	+	no
sp	2	2	1	1	0	cor	lab	2	decrease	2/2	–	–	total
spl	3	2	1	2	0	cor	cor	3	mixed	2/3	–	+	partial
sʁ	2	4	0	1	1	cor	dor	2	increase	1/2	–	+	no
st	2	0	0	2	0	cor	cor	2	decrease	2/2	–	–	total
stʁ	3	2	0	2	1	cor	dor	3	mixed	2/3	–	+	partial
stʂ	2	0	0	2	0	cor	cor	1	decrease	2/2	–	–	total
sv	2	1	1	1	0	cor	lab	0	plateau	2/2	–	+	no
ʃk	2	2	0	1	1	cor	dor	2	decrease	2/2	–	–	total
ʃl	2	1	0	2	0	cor	cor	2	increase	1/2	–	+	no
ʃm	2	3	1	1	0	cor	lab	1	increase	1/2	–	+	no
ʃn	2	1	0	2	0	cor	cor	1	increase	1/2	–	+	no
ʃp	2	3	1	1	0	cor	lab	2	decrease	2/2	–	–	total
ʃpl	3	3	1	2	0	cor	cor	3	mixed	2/3	–	+	partial

(continued)

CL	1 Size	4 POA dist.	5 Lab. C	6 Cor. C	7 Dor. C	8 POA C-in	9 POA C-fin	10 MOA dist.	12 Increase	11 Obstr.C	13 Voice C-in	14 Voice C-fin	15 Agreement
ʃpʁ	3	5	1	1	1	cor	dor	3	mixed	2/3	–	+	partial
ʃʁ	2	3	0	1	1	cor	dor	2	increase	1/2	–	+	no
ʃt	2	1	0	2	0	cor	cor	2	decrease	2/2	–	–	total
ʃtʁ	3	3	0	2	1	cor	dor	3	mixed	2/3	–	+	partial
ʃv	2	2	1	1	0	cor	lab	0	plateau	2/2	–	+	no
t͡ʃ	2	2	0	1	1	cor	dor	2	increase	2/2	–	+	no
tm	2	2	1	1	0	cor	lab	3	increase	1/2	–	+	total
tʁ	2	4	0	1	1	cor	dor	4	increase	1/2	–	+	no
t͡sv	2	1	1	1	0	cor	lab	1	increase	2/2	–	+	no
tv	2	1	1	1	0	cor	lab	2	increase	2/2	–	+	no
vl	2	2	1	1	0	lab	cor	2	increase	1/2	+	+	total
vʁ	2	5	1	0	1	lab	dor	2	increase	1/2	+	+	total

Appendix C: German clusters (CL) with percentage scores for individual parameters, the resulting composite feature score (Σ) and its mean ($\bar{x}$) (adapted from Orzechowska and Wiese 2015); ordered alphabetically. Parameters (2) and (3) (*Compositionality, Identity avoidance*) are excluded from the table and do not count towards the composite feature score (Σ) and the mean ($\bar{x}$)

CL	1 Size	4 POA dist.	5 Lab. C	6 Cor. C	7 Dor. C	8 POA C-in	9 POA C-fin	10 MOA dist.	12 Incr.	11 Obstr. C	13 Voi C-in	14 Voi C-fin	15 Voi agr.	Σ	$\bar{x}$
bj	0.86	0.11	0.55	0.21	0.48	0.25	0.36	0.36	0.66	0.38	0.18	0.82	0.38	5.6	43
bl	0.86	0.25	0.55	0.57	0.46	0.25	0.39	0.2	0.66	0.63	0.18	0.82	0.38	6.2	48
bʁ	0.86	0.04	0.55	0.21	0.48	0.25	0.36	0.2	0.66	0.63	0.18	0.82	0.38	5.62	43
dʁ	0.86	0.11	0.45	0.57	0.48	0.57	0.36	0.2	0.66	0.63	0.18	0.82	0.38	6.27	48
fj	0.86	0.23	0.55	0.21	0.48	0.25	0.36	0.07	0.07	0.38	0.82	0.82	0.48	5.58	43
fl	0.86	0.2	0.55	0.57	0.46	0.25	0.39	0.36	0.66	0.63	0.82	0.82	0.48	7.05	54
fʁ	0.86	0.11	0.55	0.21	0.48	0.25	0.36	0.36	0.66	0.63	0.82	0.82	0.48	6.59	51
gl	0.86	0.23	0.45	0.57	0.48	0.18	0.39	0.2	0.66	0.63	0.18	0.82	0.38	6.03	46
gm	0.86	0.11	0.55	0.21	0.48	0.18	0.25	0.27	0.66	0.63	0.18	0.82	0.38	5.58	43
gn	0.86	0.23	0.45	0.57	0.48	0.18	0.39	0.27	0.66	0.63	0.18	0.82	0.38	6.1	47
gʁ	0.86	0.2	0.45	0.21	0.05	0.18	0.36	0.2	0.66	0.63	0.18	0.82	0.38	5.18	40
kl	0.86	0.23	0.45	0.57	0.48	0.18	0.39	0.2	0.66	0.63	0.82	0.82	0.48	6.77	52
km	0.86	0.11	0.55	0.21	0.48	0.18	0.25	0.27	0.66	0.63	0.82	0.82	0.48	6.32	49
kn	0.86	0.23	0.45	0.57	0.48	0.18	0.39	0.27	0.66	0.63	0.82	0.82	0.48	6.84	53
kʁ	0.86	0.2	0.45	0.21	0.05	0.18	0.36	0.2	0.66	0.63	0.82	0.82	0.48	5.92	46
ks	0.86	0.23	0.45	0.57	0.48	0.18	0.39	0.36	0.66	0.38	0.82	0.18	0.38	5.94	46
kv	0.86	0.11	0.55	0.21	0.48	0.18	0.25	0.36	0.66	0.38	0.82	0.82	0.48	6.16	47
pfl	0.86	0.2	0.55	0.57	0.46	0.25	0.39	0.27	0.66	0.63	0.82	0.82	0.48	6.96	54
pfʁ	0.86	0.11	0.55	0.21	0.48	0.25	0.36	0.27	0.66	0.63	0.82	0.82	0.48	6.5	50
pl	0.86	0.25	0.55	0.57	0.46	0.25	0.39	0.2	0.66	0.63	0.82	0.82	0.48	6.94	53
pn	0.86	0.25	0.55	0.57	0.46	0.25	0.39	0.27	0.66	0.63	0.82	0.82	0.48	7.01	54
pʁ	0.86	0.04	0.55	0.21	0.48	0.25	0.36	0.2	0.66	0.63	0.82	0.82	0.48	6.36	49
ps	0.86	0.25	0.55	0.57	0.46	0.25	0.39	0.36	0.66	0.38	0.82	0.18	0.38	6.11	47
sf	0.86	0.2	0.55	0.57	0.46	0.57	0.25	0.07	0.07	0.38	0.82	0.18	0.38	5.36	41
sk	0.86	0.23	0.45	0.57	0.48	0.57	0.36	0.36	0.13	0.38	0.82	0.18	0.38	5.77	44
skl	0.14	0.23	0.45	0.21	0.48	0.57	0.39	0.27	0.14	0.88	0.82	0.82	0.14	5.54	43

(continued)

CL	1 Size	4 POA dist.	5 Lab. C	6 Cor. C	7 Dor. C	8 POA C-in	9 POA C-fin	10 MOA dist.	12 Incr.	11 Obstr. C	13 Voi C-in	14 Voi C-fin	15 Voi agr.	Σ	$\bar{x}$
skʁ	0.14	0.25	0.45	0.57	0.05	0.57	0.36	0.27	0.14	0.88	0.82	0.82	0.14	5.46	42
skv	0.14	0.11	0.55	0.57	0.48	0.57	0.25	0.2	0.14	0.13	0.82	0.82	0.14	4.92	38
sl	0.86	0.07	0.45	0.21	0.46	0.57	0.39	0.36	0.66	0.63	0.82	0.82	0.48	6.78	52
sm	0.86	0.25	0.55	0.57	0.46	0.57	0.25	0.11	0.66	0.63	0.82	0.82	0.48	7.03	54
sn	0.86	0.07	0.45	0.21	0.46	0.57	0.39	0.11	0.66	0.63	0.82	0.82	0.48	6.53	50
sp	0.86	0.25	0.55	0.57	0.46	0.57	0.25	0.36	0.13	0.38	0.82	0.18	0.38	5.76	44
spl	0.14	0.25	0.55	0.21	0.46	0.57	0.39	0.27	0.14	0.88	0.82	0.82	0.14	5.64	43
sʁ	0.86	0.11	0.45	0.57	0.48	0.57	0.36	0.36	0.66	0.63	0.82	0.82	0.48	7.17	55
st	0.86	0.07	0.45	0.21	0.46	0.57	0.39	0.36	0.13	0.38	0.82	0.18	0.38	5.26	40
stʁ	0.14	0.25	0.45	0.21	0.48	0.57	0.36	0.27	0.14	0.88	0.82	0.82	0.14	5.53	43
sts	0.86	0.07	0.45	0.21	0.46	0.57	0.39	0.11	0.13	0.38	0.82	0.18	0.38	5.01	39
sv	0.86	0.2	0.55	0.57	0.46	0.57	0.25	0.07	0.07	0.38	0.82	0.82	0.48	6.1	47
ʃk	0.86	0.25	0.45	0.57	0.48	0.57	0.36	0.36	0.13	0.38	0.82	0.18	0.38	5.79	45
ʃl	0.86	0.2	0.45	0.21	0.46	0.57	0.39	0.36	0.66	0.63	0.82	0.82	0.48	6.91	53
ʃm	0.86	0.23	0.55	0.57	0.46	0.57	0.25	0.11	0.66	0.63	0.82	0.82	0.48	7.01	54
ʃn	0.86	0.2	0.45	0.21	0.46	0.57	0.39	0.11	0.66	0.63	0.82	0.82	0.48	6.66	51
ʃp	0.86	0.23	0.55	0.57	0.46	0.57	0.25	0.36	0.13	0.38	0.82	0.18	0.38	5.74	44
ʃpl	0.14	0.23	0.55	0.21	0.46	0.57	0.39	0.27	0.14	0.88	0.82	0.82	0.14	5.62	43
ʃpʁ	0.14	0.11	0.55	0.57	0.48	0.57	0.36	0.27	0.14	0.88	0.82	0.82	0.14	5.85	45
ʃʁ	0.86	0.23	0.45	0.57	0.48	0.57	0.36	0.36	0.66	0.63	0.82	0.82	0.48	7.29	56
ʃt	0.86	0.2	0.45	0.21	0.46	0.57	0.39	0.36	0.13	0.38	0.82	0.18	0.38	5.39	41
ʃtʁ	0.14	0.23	0.45	0.21	0.48	0.57	0.36	0.27	0.14	0.88	0.82	0.82	0.14	5.51	42
ʃv	0.86	0.25	0.55	0.57	0.46	0.57	0.25	0.07	0.07	0.38	0.82	0.82	0.48	6.15	47
tj	0.86	0.25	0.45	0.57	0.48	0.57	0.36	0.36	0.66	0.38	0.82	0.82	0.48	7.06	54
tm	0.86	0.25	0.55	0.57	0.46	0.57	0.25	0.27	0.66	0.63	0.82	0.82	0.38	7.09	55
tʁ	0.86	0.11	0.45	0.57	0.48	0.57	0.36	0.2	0.66	0.63	0.82	0.82	0.48	7.01	54
tsv	0.86	0.2	0.55	0.57	0.46	0.57	0.25	0.11	0.66	0.38	0.82	0.82	0.48	6.73	52
tv	0.86	0.2	0.55	0.57	0.46	0.57	0.25	0.36	0.66	0.38	0.82	0.82	0.48	6.98	54
vl	0.86	0.25	0.55	0.57	0.46	0.25	0.39	0.36	0.66	0.63	0.18	0.82	0.38	6.36	49
vʁ	0.86	0.11	0.55	0.21	0.48	0.25	0.36	0.36	0.66	0.63	0.18	0.82	0.38	5.85	45

References

Albright, Adam. 2009. Feature-based generalisation as a source of gradient acceptability. *Phonology* 26. 9–41.

Bailey, Todd M. & Ulrike Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language* 44. 568–591.

Basirat, Anahita, Cédric Patin & Jérémie Jozefowiez. 2021. Sonority projection effect in French: A signal detection theory approach. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 66(2). 255–266.

Bybee, Joan L. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.
Celata, Chiara, Katharina Korecky-Kröll, Irene Ricci & Wolfgang U. Dressler. 2015. Phonotactic processing and morpheme boundaries: Word-final /Cst/ clusters in German. *Italian Journal of Linguistics* 27. 85–110.
Clements, George N. 1990. The role of the sonority cycle in core syllabification. In John Kingston & Mary E. Beckman (eds.), *Papers in Laboratory Phonology I: Between the grammar and physics of speech*, vol. I, 283–333. New York: Cambridge University Press.
Daland, Robert N., Bruce Hayes, James White, Marc Garellek, Andrea Davis & Ingrid Norrmann. 2011. Explaining sonority projection effects. *Phonology* 28. 197–234.
Domahs, Ulrike, Wolfgang Kehrein, Johannes Knaus, Richard Wiese & Matthias Schlesewsky. 2009. Event-related potentials reflecting the processing of phonological constraint violations. *Language and Speech* 52(4). 415–435.
Duden, Aussprachewörterbuch. 1990. *Duden – Wörterbuch der deutschen Standardaussprache*. Dudenverlag: Mannheim & Wien.
Dziubalska-Kołaczyk, Katarzyna. 2019. On the structure, survival and change of consonant clusters. *Folia Linguistica Historica* 40(1). 107–127.
Féry, Caroline. 2003. Onsets and nonmoraic syllables in German. In Caroline Féry & Ruben van de Vijver (eds.), *The syllable in Optimality Theory*, 213–237. Cambridge: Cambridge University Press.
Goad, Heather. 2011. The representation of sC clusters. In Marc van Oostendorp, Colin J. Ewen, Elizabeth Hume & Karen Rice (eds.), *The Blackwell companion to phonology*, 898–923. Malden, MA & Oxford: Wiley Blackwell.
Goad, Heather & Yvan Rose. 2004. Input elaboration, head faithfulness and evidence for representation in the acquisition of left-edge clusters in West Germanic. In René Kager, Joe Pater & Wim Zonneveld (eds.), *Constraints in phonological acquisition*, 109–157. Cambridge: Cambridge University Press.
Greenberg, Joseph H. 1978. Some generalizations concerning initial and final consonant clusters. In Joseph H. Greenberg (ed.), *Universals of human language*, vol. 2, 243–279. Stanford/CA: Stanford University Press.
Hall, Tracy A. 1992. *Syllable structure and syllable-related processes in German*. Tübingen: Max Niemeyer Verlag.
Hall, Tracy A. (ed.). 2001 *Distinctive feature theory*. Berlin: Walter de Gruyter.
Harris, James W. 1983. *Syllable structure and stress in Spanish: A nonlinear analysis*. Cambridge, MA: The MIT Press.
Hayes, Bruce & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39(3). 379–440.
Hirst, Daniel. 1980. Linearisation and the single-segment hypothesis. In Jacqueline Guéron, Hans-Georg Obenauer & Jean-Yves Pollock (eds.), *Grammatical representation*, 87–99. Dordrecht: Foris.
Hothorn, Torsten, Achim Zeileis, Richard W. Farebrother, Clint Cummins, Giovanni Millo & David Mitchell. 2021. Testing linear regression models. R package version 0.9–39.
Ingram, David. 1978. The role of the syllable in phonological development. In Alan Bell & Joan Hooper (eds.), *Syllables and segments*, 143–155. New York: North Holland.
International Phonetic Association. 2007. In *Handbook of the international phonetic association: A guide to the use of the international phonetic alphabet*, 9th edn. Cambridge: Cambridge University Press.
Jakobson, Roman. 1962. *Selected writings I: Phonological studies*. The Hague: Mouton.
Jakobson, Roman. 1968. *Child language, aphasia, and phonological universals*. The Hague: Mouton.
Jarosz, Gaja. 2017. Defying the stimulus: Acquisition of complex onsets in Polish. *Phonology* 34(2). 269–298.

Kiparsky, Paul. 2008. Universals constrain change; change results in typological generalizations. In Jeff Good (ed.), *Language universals and language change*, 23–53. Oxford: Oxford University Press.
Kohler, Klaus J. 1990. German. *Journal of the International Phonetic Association* 20. 48–50.
Korecky-Kröll, Katharina, Wolfgang U. Dressler, Eva Maria Freiberger, Eva Reinisch, Karlheinz Mörth & Gary Libben. 2014. Morphonotactic and phonotactic processing in German-speaking adults. *Language Sciences* 46. 48–58.
Leopold, Werner F. 1949. *Speech development of a bilingual child*. Evanston: Northwestern University Press.
Levelt, Clara C., Niels O. Schiller & Willem J. M. Levelt. 1999. A developmental grammar for syllable structure in the production of child language. *Brain and Language* 68. 291–299.
Lleo, Conxita & Katherine Demuth. 1999. Prosodic constraints on the emergence of grammatical morphemes: Crosslinguistic evidence from Germanic and Romance languages. In *Proceedings of the 23rd Annual Boston University Conference on Language Development*, 407–418. Somerville, MA: Cascadilla Press.
Maddieson, Ian. 2013. Syllable structure. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structure online*. Munich: Max Planck Digital Library. http://wals.info/.
McCarthy, John J. 1988. Feature geometry and dependency: A review. *Phonetica* 45. 84–108.
Meinhold, Gottfried & Eberhard Stock. 1980. *Phonologie der deutschen Gegenwartssprache*. Leipzig: VEB Bibliographisches Institut Leipzig.
Orzechowska, Paula. 2019. *Complexity in Polish phonotactics: On features, weights, rankings and preferences*. Singapore: Springer Nature.
Orzechowska, Paula & Katarzyna Dziubalska-Kołaczyk. 2022. Gradient phonotactics and frequency: A study of German initial clusters. *Italian Journal of Linguistics* 34(1). 103–138.
Orzechowska, Paula & Richard Wiese. 2011. Reconstructing the sonority hierarchy. In *Proceedings of the 17th International Conference of the Phonetic Sciences*, 1542–1545. Hongkong.
Orzechowska, Paula & Richard Wiese. 2015. Preferences and variation in phonotactics: A multi-dimensional evaluation of German and Polish. *Folia Linguistica* 49. 439–486.
Orzechowska, Paula & Paulina Zydorowicz. 2019. Frequency effects and markedness in phonotactics. *Poznań Studies in Contemporary Linguistics* 55. 157–179.
Ott, Susan, Ruben van de Vijver & Barbara Höhle. 2006. The effect of phonotactic constraints in German-speaking children with delayed phonological acquisition: Evidence from production of word-initial consonant clusters. *International Journal of Speech Language Pathology* 8(4). 323–334.
Parker, Steve. 2012a. Sonority distance versus sonority dispersion–A typological survey. In Steve Parker (ed.), *The sonority controversy*, 101–166. Berlin & Boston: De Gruyter Mouton.
Parker, Steve (ed.). 2012b. *The sonority controversy*. Berlin & Boston: De Gruyter Mouton.
Parker, Steve. 2017. Sounding out sonority. *Language and Linguistics Compass* 11. 1–197.
Pierrehumbert, Janet. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In Joan L. Bybee & Paul Hopper (eds.), *Frequency and the emergence of linguistic structure*, 137–158. Amsterdam & Philadelphia: John Benjamins.
R Core Team. 2020. *R: A language and environment for statistical computing*. Vienna, Austria: R foundation for statistical computing. http://www.R-project.org/.
Rice, Karen D. 2007. Markedness in phonology. In Paul de Lacy (ed.), *The Cambridge handbook of phonology*, 79–97. Cambridge: Cambridge University Press.
Rochoń, Marzena. 2000. *Optimality in complexity: The case of Polish consonant clusters*. Berlin: Akademie-Verlag.
Selkirk, Elisabeth O. 1984. On the major class features and syllable theory. In Mark Aronoff & Richard T. Oehrle (eds.), *Language sound structure*, 107–136. Cambridge, MA: The MIT Press.

Silva, Susana, Marina Vigário, Barbara L. Fernandez, Rita Jerónimo, Kai Alter & Sónia Frota. 2019. The sense of sounds: Brain responses to phonotactic frequency, phonological grammar and lexical meaning. *Frontiers in Psychology* 10. 681.

Ulbrich, Christiane, Phillip Alday, Knaus Johannes, Orzechowska Paula & Richard Wiese. 2016. The role of phonotactic principles in language processing. *Language, Cognition and Neuroscience* 31(5). 662–682.

Vennemann, Theo. 1982. Zur Silbenstruktur der deutschen Standardsprache. In Theo Vennemann (ed.), *Silben, Segmente, Akzente*, 261–305. Tübingen: Max Niemeyer.

van de Vijver, Ruben & Dinah Baer-Henney. 2012. Sonority intuitions are provided by the lexicon. In Steve Parker (ed.), *The sonority controversy*, 195–218. Berlin & Boston: De Gruyter Mouton.

Vitevitch, Michael & Paul Luce. 1999. Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language* 40(3). 374–408.

Wiese, Richard. 1988. *Silbische und Lexikalische Phonologie: Studien zum Chinesischen und Deutschen*. Tübingen: Max Niemeyer.

Wiese, Richard. 2000. *The phonology of German*. Oxford: Clarendon Press.

Wiese, Richard, Paula Orzechowska, Phillip Alday & Christiane Ulbrich. 2017. Structural principles or frequency of use? An ERP experiment on the learnability of Polish consonant clusters. *Frontiers in Psychology – Auditory Cognitive Neuroscience* 7. 2005.

Yavaş, Mehmet, Avivit Ben-David, Ellen Gerrits, Kristian E. Kristoffersen & Hanne G. Simonsen. 2008. Sonority and cross-linguistic acquisition of initial s-clusters. *Clinical Linguistics and Phonetics* 22(6). 421–441.

Yavaş, Mehmet, Annette Fox-Boyer & Blanca Schaefer. 2018. Patterns in German /ʃC/-cluster acquisition. *Clinical Linguistics and Phonetics* 32(10). 913–931.

Yip, Moira. 1988. The obligatory Contour principle and phonological rules: A loss of identity. *Linguistic Inquiry* 19. 65–100.

Zydorowicz, Paulina, Orzechowska Paula, Katarzyna Dziubalska-Kołaczyk, Michał Jankowski, Piotr Wierzchoń & Dawid Pietrala. 2016. *Phonotactics and morphonotactics of Polish and English: Theory, description, tools and applications*. Poznań: Wydawnictwo Naukowe UAM.