Structural and accidental phonotactic gaps*

Kyle Gorman
University of Pennsylvania

Revised January 2013 (comments welcome)

1 Introduction

It is indisputable that many phonotactic restrictions are easily described with reference to prosodic primitives. In some cases, prosodic factors constrain the system of contrast. For instance, Latin has contrastive vowel and consonant length (e.g., *os* 'bone' vs. *ōs* 'mouth', *anus* 'ring' vs. *annus* 'year'), but the latter contrast is suspended in codas preceded by a diphthong or long monophthong; a syllable may contain a long vowel, or be checked by the first half a geminate consonant, but not both. However, constraints on underlying representations may also involve references to non-contrastive prosodic structures such as the syllable (e.g., Hooper 1973, Kahn 1976).[1] For instance, as noted by Haugen (1956), numerous restrictions on word-medial consonant clusters have a unified statement in prosodic terms:

(1) Medial Cluster Law:

A medial cluster can consist maximally of a well-formed medial coda and a well-formed medial onset

As an illustration of this tautology, consider languages like Yokuts, which forbid complex codas and complex onsets, and which enforce these restrictions in complex words via processes such as vowel epenthesis. Newman (1944:26f.) notes that this imposes an an upper bound on the size of medial clusters: no cluster of more than two consonants can be parsed into a simple coda and simple onset. In the case of Yokuts, it is certainly possible to state this restriction without reference to syllable structure, as *CCC, a constraint on trisyllabic clusters (e.g., Ettlinger 2008:92f., Zuraw 2003:820f.). However, the aforementioned constraints on complex onsets and codas find independent motivation from the total absence of initial and final clusters in Yokuts;[2] with these

*Thanks to Steve Anderson, Gene Buckley, Constantine Lignos, Rolf Noyer, Hilary Prichard, Charles Yang, and audiences at the University of California, Santa Cruz, the University of Delaware, and New York Unversity.

[1]See Blevins 1995 for the claim that syllable structure is universally non-contrastive, and Elfner 2006 for arguments that a putative counterexample derives from an underlying vowel length contrast.

[2]The tendency of consonants to pattern with word boundaries and morph junctures has been long been noted (e.g., Hill 1954, Lass 1971, Moulton 1947); Kahn (1976:24f.) takes this as evidence for the syllable.

two restrictions in place, a further constraint on medial triconsonantal clusters is otiose.[3] While the medial cluster law is certainly consistent with the hypothesis that syllable structure may be present in underlying representations (e.g., Anderson 1974:255, Vaux 2003), this need not be the case under Stampean occultation. If an underlying medial consonant sequence appears on the surface, it satisfies the medial cluster law by definition. If, however, an underlying cluster is modified by consonant deletion, coalescence, or vowel epenthesis, then it need not consist of a licit medial onset and medial coda.

Pierrehumbert (1994) begins a study of English word-medial consonant clusters with a restatement of the medial cluster law:

> That is, in the absence of additional provisos, any concatenation of a well-formed coda and a well-formed onset is predicted to be possible medially in a word. (Pierrehumbert 1994:168)

However, Pierrehumbert reports that the vast majority of the "possible" clusters (i.e., those which conform to the medial cluster law) are in fact unattested. So as to acount for this, Pierrehumbert presents "provisos" in the forms of static co-occurrence restrictions, unrelated to any phonological alternation in English.[4] This chapter will argue, however, the static constraints proposed by Pierrehumbert are unnecessary, and that the only restrictions on the inventory of medial clusters in English (beyond the medial cluster law) are those which derive from well-known phonological processes.

2 English syllable contact clusters

The aforementioned study by Pierrehumbert, as well as further investigations of this domain by Duanmu (2009: chap. 8) and Hammond (1999: chap. 3), argue for the necessity of admitting static phonotactic constraints, and illustrate proposals for the architecture of the phonotactic system. However, there are a number of reasons to reconsider the findings of these authors in light of the proposals made in previous chapters.

2.1 The role of phonological processes

Duanmu, Hammond, and Pierrehumbert do not generally take into account the effects of phonological processes which target medial clusters in English. As a consequence, some of the static constraints these authors identify may be in fact the product of English morphophonemics and Stampean occultation. For instance, Pierrehumbert writes that "nasal-stop sequences agree in

[3] Côté (2000:31f.) alludes to another critique of constraints like *CCC, namely that this constraint requires "counting": see Isac and Reiss 2008:64f. for further discussion of the comparative merits of "counting" and "grouping" analyses in phonology.

[4] Given the arguments presented in the previous chapter, that not all lexical tendencies have a synchronic basis, one may question the intuition that any gaps in the English cluster inventory (or those that go beyond the medial cluster law) must be accounted for by the synchronic grammar. In addition to some informal statistical evidence (of the sort problematized in the previous chapter), Pierrehumbert administers a wordlikeness task to validate the static constraints she proposes. However, this experiment is of a quite informal nature and the results are given only a superficial analysis, so it is less than probative.

labiality" (175) and posits a static constraint to account for this fact. However, this generalization is merely a narrower form of a restriction deriving from a process of Nasal Place Assimilation (see §3.2.2). Other derived constraints are simply not mentioned; for instance, Pierrehumbert does not discuss the highly reliable tendency of obstruent-obstruent clusters to agree in voicing (see §3.2.2); while Hammond (1999) does allude to this restriction, it is dismissed in light of a few apparent counterexamples (though see §3.1.1 below). In contrast, this chapter attempts to evaluate derived and static constraints on an equal footing.

2.2 The role of sparsity

Pierrehumbert (1994) infers static constraints from near-exceptionless gaps in the lexicon, but little effort is made to show that the patterns of lexical underrepresentation are not due to chance. Consequently, it is possible to suggest that some of these gaps are accidental rather than structural in nature. This is made all the more likely given the tendency of segment and cluster frequency distributions to be highly skewed (e.g., Pande and Dhami 2010, Sigurd 1968, Tambovtsev and Martindale 2007, Weiss 1961) so that it is difficult to distinguish between structural and accidental gaps. Furthermore, Pierrehumbert considers only triconsonantal clusters, but medial clusters may be as short as two consonants, as in *a*[n.t]*ics*, or as long as four, as in *mi*[n.str]*el*, and no justification is given for ignoring clusters of other lengths. If there is any effect of this focus, it is presumably to produce further sparsity in the distribution observed.

2.3 The role of morphological segmentation

Many components of Pierrehumbert's study cannot be replicated. Pierrehumbert limits her study to words she judges to be "morphologically simple" and "reasonably familiar"; the author's sensations thereof are not replicable, nor are they available to other researchers in any form. It has been suggested (e.g., Labov 1975, Schütze 1996) that the sensations (as well as cognitive limitations) of concerned parties should not be granted evidential status in the first place, given the potential for implicit bias; Labov calls the *Experimenter Principle*.

It is not uncommon for analysts to propose otherwise-unmotivated morphological junctures simply to preserve phonological or phonotactic generalizations. This is done by Chomsky and Halle (1968), for instance, to simplify principles of English stress assignment. Similarly, Rice (2009:546) analyses many words in Slave as compounds simply because they contain consonant clusters that rarely occur in morph-internal contexts. Applied indiscriminately, however, this heuristic trivializes both morphological segmentation and phonotactic generalization. For these reasons, the wordlist used in this study is derived from a publicly available database, and no experimenter intuitions are used.

3 Evaluation

After constructing a sample of syllable contact clusters in English simplex words, this sample is used to evaluate the coverage of static and derived constraints.

3.1 Method

3.1.1 Materials and procedure

Following Duanmu (2009: chap. 8) and Hammond (1999: chap. 3), who also consider restrictions on English medial clusters, a wordlist is generated using the English portion of the CELEX database (Baayen et al. 1996). Only words marked in CELEX as "monomorphemic" are used, and all words labeled in CELEX as non-native are excluded.[5] These more-stringent criteria exclude many words labeled exceptions in the studies by Duanmu or Hammond in their studies. For instance, nearly all the exceptions to Obstruent Voice Assimilation (see §3.2.2 below) noted by Hammond (1999:74) are excluded either as complex words (e.g., *jurisdiction*, *madcap*, *tadpole*, *scapegoat*, *magpie*) or non-native (e.g., *vodka*, *smorgasbord*).

In contrast to prior studies, these criteria also exclude words which consist of a Latinate prefix and a bound stem (e.g., *inspect*, *excrete*). While Pierrehumbert rejects this analysis as as unmotcontroversial, it in fact has extensive formal and experimental support. First, Latinate prefixes simplify the statement of many morphophonemic details in English. For instance, Aronoff (1976:11f.) observes that Latinate forms which share the same bound stem also share irregular allomorphs of that stem under derivation.

(2) Bound stem-specific allomorphy:

a.	adhere	adhesion
	cohere	cohesion
b.	conceive	conception
	perceive	perception

Aronoff takes this to be evidence that *adhere* and *cohere*, for instance, share a bound stem. There is also an interaction between Latinate prefixes on verbs and the complements they select. Latinate verbs do not generally allow ditransitive, verb participle, or adjectival resultative constructions, all of which are acceptable with similar Anglo-Saxon verbs (e.g., Gropen et al. 1989, Harley 2009).

(3) Restrictions on Latinate verbal complements:

a.	show him the painting	$\sim$	*exhibit him the painting
b.	drink himself stupid	$\sim$	*imbibe himself stupid
c.	break it off	$\sim$	*terminate it off

Lexical decision also provide evidence for the segmentation of Latinate prefixed forms. Taft and Forster (1975, 1976) and Taft et al. (1986) find that nonce words like **re-sert*, which appear to be composed of a prefix and a bound stem, take longer to reject that non-words which lack apparent morphological structure, such as **refant*. Bound stems also show frequency effects independent of whole word frequency (Taft 1979, Taft and Ardasinski 2006). Finally, Emmorey (1989) and Forster and Azuma (2000) report facilitative priming, thought to implicating morphological relatedness, between pairs like *permit-submit*, which appear to share a bound stem.

[5] The first criterion results in the exclusion of proper names, which have long been noted to push the bounds of native language phonotactics (e.g., Trubetzkoy 1958:254).

	attested	unattested	saturation	p-value
conforming	25	91	22%	.106
violating	4	40	9%	

Table 1: Dorsal-labial cluster attestation in the lexical sample

	attested	unattested	saturation	p-value
conforming	56	304	15%	.430
violating	37	243	13%	

Table 2: Coda coronal obstruent cluster attestation in the lexical sample

3.2 Results

Filtering the CELEX data according to the above criteria results in a list of 6,619 simplex words. The full set of clusters and their frequencies are listed in Appendix A. The CELEX transcriptions of these words are then syllabified and phonologized using a procedure described in Appendix B. In all, the sample contains 23 different medial coda and 40 different medial onsets. Of the 920 ($= 21 \times 40$) medial clusters that would result from free combination of medial coda and medial onset, 174 (19%) are attested.

3.2.1 Static constraints

To account for the 81% of "possible" but unattested clusters, Pierrehumbert (1994) proposes three static constraints on English medial clusters.

Dorsal-labial clusters Pierrehumbert (1994:173) writes that "velar obstruents occurred only before coronals in the clusters studied, never before labials or other velars", while noting that absence of velar-velar clusters is expected due to a separate constraint on geminate clusters (see §3.2.2). However, biliteral velar-labial clusters are found in words such as *a*[k.m]*e*, *ru*[g.b]*y*, or *pi*[g.m]*ent*. Velar-labial clusters are somewhat less common than velar-coronal clusters (e.g., *ve*[k.t]*or*), but such underrepresentation is not unlikely to occur by chance according to the Fisher exact test (Table 1).

Coronal obstruent codas Pierrehumbert (1994:175) claims that "clusters with a coronal obstruent in the coda do not occur", but at the same time observes exceptions like *a*[nt.l]*er*, *ke*[s.tr]*el* and *oi*[nt.m]*ent*. In the CELEX sample (Table 2), coda coronal obstruent clusters are not significantly less likely to occur than non-coronal obstruent clusters (e.g., *re*[p.t]*ile*). While not shown in tabular form, the same is true if attention is restricted to triconsonantal clusters ($p = .129$).

ABA clusters Pierrehumbert (1994:176) observes a "lack of clusters with identical first and third elements", ignoring presence or absence of voicing. Despite the fact that there are no exceptions to this generalization, these ABA clusters are not significantly less common than any other triconsonantal and quadraconsonantal clusters (Table 3).

	attested	unattested	saturation	*p*-value
conforming	47	512	8%	.250
violating	0	25	0%	

Table 3: ABA cluster attestation in the lexical sample

	attested	unattested	saturation	*p*-value
conforming	35	329	10%	.002
violating	11	305	3%	

Table 4: Obstruent voice assimilation cluster attestation in the lexical sample

Summary There is no statistical support for any of Pierrehumbert's static constraints.

3.2.2 Derived constraints

In *SPE*, Chomsky and Halle (1968) describe three phonological processes which target medial consonant clusters. As will be shown, these three processes have a profound influence on the English cluster inventory.

Obstruent voice assimilation Voice assimilation alternations are evidenced by the non-syllabic allomorphs of the regular past (e.g., *nap*[t]-*nab*[d]) and noun plural (e.g., *lap*[s]-*lab*[z]), which take the voicing specification of a preceding obstruent;[6] voice assimilation is also claimed to operate across prefix and compound junctures (Davidsen-Nielsen 1974).

(4) Obstruent Voice Assimilation:

$$\left[-\textsc{Son}\right] \longrightarrow \left[=\textsc{Voi}\right] \,/\, \_\_ \begin{bmatrix} =\textsc{Voi} \\ -\textsc{Son} \end{bmatrix}$$

Pierrehumbert (1994) does not discuss a constraint against adjacent obstruents disagreeing in voice. However, the vast majority of medial obstruents clusters in simplex words are either uniformly voiced, as in *hu*[z.b]*and*, or uniformly voiceless, as in or *rha*[p.s]*osdy* (Table 4). Hetero-voiced clusters, like those in *a*[b.s]*inth* and *a*[s.b]*estos*, are far rarer than would be expected from chance.

Nasal place assimilation Nasal Place Assimilation (e.g., Borowsky 1986:65f., *SPE*:85, Halle and Mohanan 1985:62) permits [ŋ] to be described as an allophone of /n/ (see §B.3), and furthermore accounts for allomorphy in certain Latinate prefixes.

[6] Underlying /-d, -z/ are assumed here (e.g., Anderson 1973, Baković 2005:284f., Basbøll 1972, Chomsky and Halle 1968:210, Hockett 1958:282, Pinker and Prince 1988:102, Shibatani 1972); alternative analyses are put forth by Bloomfield (1933:210f.), Borowsky (1986:135), Hoard and Sloat (1971), Kiparsky (1985), Lightner (1970), Luelsdorff (1969), Miner (1975), Nida (1948:426), and Zwicky (1975).

	attested	unattested	saturation	p-value
conforming	31	2	94%	3.4E−07
violating	11	22	33%	

Table 5: Nasal place assimilation cluster attestation in the lexical sample

(5) *im-*/*in-* allomorphy:

a.	polite	i[m.p]olite
	balance	i[m.b]alance
b.	tangible	i[n.t]angible
	decent	i[n.d]ecent

The rule is formalized below.

(6) Nasal Place Assimilation:

$$\left[+\textsc{Nas}\right] \longrightarrow \begin{bmatrix} = \textsc{Lab} \\ = \textsc{Cor} \\ = \textsc{Dor} \end{bmatrix} \; / \_\_ \; \begin{bmatrix} = \textsc{Lab} \\ = \textsc{Cor} \\ = \textsc{Dor} \\ -\textsc{Son} \end{bmatrix}$$

Virtually all clusters consisting of a nasal coda followed by a homorganic obstruent (e.g., *pi*[m.p]*le*, *sta*[n.z]*a*, *mo*[ŋ.k]*ey*) are attested (Table 5). As Pierrehumbert (1994:175) observes, heterorganic clusters, like those *pli*[m.s]*oll* or *scri*[m.ʃ]*aw*, do occur, but in this sample they are significantly more rare.

Degemination The final alternation found in English medial clusters is the simplification of geminates which is characteristic of "level I" morphology, and is found in the irregular /-t/ past tense (e.g., *bend*/*ben*[t], *build*/*buil*[t]), in *-ly* deadjectival derivatives (e.g., *norma*[l]*y*, cf. *calm*[l]*y*), and Latinate prefix allomorphy (Borowsky 1986:102, *SPE*:148). Degemination is formalized here as a rule deleting the first of two segments agreeing on all feature values (except for voice, possibly).

(7) Degemination:

$$\begin{bmatrix} = \textsc{Lab} \\ = \textsc{Cor} \\ = \textsc{Dor} \\ = \textsc{Son} \\ \dots \end{bmatrix} \longrightarrow \emptyset \; / \_\_ \; \begin{bmatrix} = \textsc{Lab} \\ = \textsc{Cor} \\ = \textsc{Dor} \\ = \textsc{Son} \\ \dots \end{bmatrix}$$

The sample contains no sequences of identical segments, or of identical segments differing only in voice, something highly unlikely to arise by chance (Table 6); Degemination and Stampean occultation provide a natural explanation for this gap. It is interesting to compare the absense

	attested	unattested	saturation	p-value
conforming	173	643	21%	$1.2\text{E}{-10}$
violating	0	104	0%	

Table 6: Degemination in the lexical sample

of geminates to the constraint against "ABA" clusters proposed by Pierrehumbert in this regardl: both are exceptionless, but only the former imposes a lexical tendency unlikely to arise by chance.

Summary All three of the *SPE* rules targeting medial clusters have a robust effect in constraining the inventory of possible word-medial syllable clusters; possible clusters which are surface exceptions to these three rules are much less likely to be attested than those which conform to them.

3.2.3 Computational models

Current computational models of phonotactic knowledge can "rate" possible clusters, assigning a numerical wellformedness score to any input. These models can be applied to a task of predicting which clusters are and are not attested by transforming these numerical values into a categorical prediction of either attestation or non-attestation. This is accomplished here with a soft-margin support vector machine (Cortes and Vapnik 1995) with a linear kernel, which attempts to find a single optimal numerical value about which to split attested and unattested clusters. This classifier is not intended to correspond to any component of a cognitively plausible model of phonotactic learning: it simply represents an upper bound for predicting the cluster inventory from positive data.

The models are scored using a "leave-one-out" scheme, in which each observation (a cluster) is scored using a model trained on all other observations. Four metrics are used to evaluate model fit. *Accuracy* represents the probability that a cluster is correctly classified as attested or unattested. Two additional metrics break down accuracy into constituent parts; *precision* represents the probability that a cluster which is predicted to be unattested is in fact unattested, and *recall* is the probability that an unattested cluster is predicted as such. It is possible to increase precision at the expense of recall, by predicting non-attestation for a greater number of clusters, or to maximize recall at the expense of precision by predicting all clusters to be unattested. F_1, the harmonic mean of these two measures, is a standard metric for quantifying this tradeoff; any increase in either precision or recall will result in an increase in F_1. The results are summarized in Table 7.

Null baseline In a classification task, the simplest baseline is one which uniformly predicts the most common outcome. Since only 19% of clusters are attested, 81% accuracy can be achieved simply by predicting all clusters to be unattested.

Expected frequency Pierrehumbert (1994) proposes that the well-formedness of a syllable contact cluster is proportional to the product of the independent probabilities of the coda and of the onset that make it up; this is the cluster's *expected frequency*. Pierrehumbert reports that this is

	accuracy	precision	recall	F_1
Baseline	0.812	0.812	1.000	0.896
Expected frequency	0.835	0.837	0.960	0.894
Derived constraints	0.838	0.835	0.967	0.897
DC & EF	0.861	0.866	0.969	0.914
Hayes and Wilson (2008)	0.835	0.964	0.833	0.894

Table 7: Results for the cluster classification task

an excellent predictor of which complex clusters occur and which do not. This model does not impose any constraints which span the syllable boundary; rather, it is a model of which clusters might be expected to represent accidental gaps in the sample. This produces a small but significant improvement in accuracy over the null baseline (sign test, $p = 4.5\text{E}{-}05$).

Derived constraints By hypothesis, Obstruent Voice Assimilation, Nasal Place Assimilation, and Degemination rule out a large number of possible clusters. In all, they target 316 out of 920 possible clusters (34%) for neutralization; of these clusters, only 11 (3%) occur in the sample. Together, these three processes define a simple classifier in which a cluster is predicted to be attested only if it would not be neutralized by one of these processes. This results in increased precision and a small but significant improvement in accuracy compared to the null baseline (sign test, $p = 4.0\text{E}{-}09$).

Expected frequency and derived constraints It is possible to combine into a single classifier the intuitions of the expected frequency and derived constraint models, the former accounting for accidental gaps and the latter for structural gaps imposed by neutralizing phonological processes. Simultaneously accounting for both sources of cluster inventory gaps, this model outperforms all others in accuracy and F_1.

Maximum entropy phonotactics Hayes and Wilson (2008) present a model using the principle of maximum entropy to weigh a large number of competing phonotactic constraints. In one sense, this is isomorphic to the expected frequency model in that the constraint discovery mechanism is sensitive to the expected frequency of clusters: it favors constraints which rule out clusters with high expected frequency (but which are unattested) over those which have low expected frequency. Also like the expected frequency model, alternations play no role and static constraints like those proposed by Pierrehumbert (1994) may be posited.

Since this model has numerous experimenter-defined parameters, a close replication of Hayes and Wilson's original study is attempted: both their implementation and phonological feature specifications are used here. Following Hayes and White (in press), dictionary entries are syllabified using the procedure described in Appendix B, and a novel feature [± Coda] is added to allow the model to distinguish coda and onset consonants. Also, following Hayes and Wilson, constraints are limited to those s panning as many as three segments and the suggested "accuracy schedule" is used. Since the maximum entropy model produces slightly different scores on each run, the worst-performing of 10 runs is reported here, following Hayes and Wilson. This model has the poorest recall of any model; compared to the derived constraints baseline, the constraints

induced by the maximum entropy model are narrower. This is particularly clear regarding possible clusters with a nasal coda followed by a non-homorganic obstruent, like *[m.kl]: the vast majority of such clusters, which would be neutralized by NASAL PLACE ASSIMILATION, are unattested, but many are erroneously assigned the highest possible score by the maximum entropy model.

Summary Expected frequency and derived constraints effectively account for accidental and structural gaps in the cluster inventory. The Hayes and Wilson (2008) computational model only provides an inferior approximation of the derived constraints.[7]

3.3 Discussion

What then is to be said of the 366 clusters which do not violate a derived constraint, but which are yet unattested, like [b.z] or [z.n]? Insofar as the "discovery procedure" used by linguists (e.g., Pierrehumbert) and computers (e.g., the Hayes and Wilson model) are based on principled phonological primitives, yet fail to find meaningful gaps, the absence of these clusters appears to be phonologically arbitrary. It appears that further gaps in the cluster inventory cannot be described in phonological, structural terms. The remainder of this chapter is concerned with the nature of these gaps.

3.3.1 The probability of accidental gaps

Good (1953) proposes a method for estimating the probability of accidental gaps in a sample distribution. This takes the form of an estimate of p_0, the probability of outcomes with have zero frequency in the sample.

$$p_0 = \frac{n_1}{N}$$

In prose, the value of p_0 is the ratio of clusters that occur exactly once in the sample (n_1) to the size of the sample (N). In the data here, $n_1 = 67$ and there are 997 clusters in all, so were it possible to extend the lexical sample, there is an approximately 7% chance that the next cluster would "fill in" what is now a gap. This alone indicates the non-trivial amount of "missingness" and the high likelihood of accidental gaps in such a sample.

3.3.2 Simulating medial clusters

It can be shown that the large number of possible-but-unattested clusters is a logical necessity given the sparse distribution of codas and onsets. The rank and type frequency (i.e., frequency in the lexicon) of medial codas and onsets in the lexical sample described below are displayed in log-log space in Figure 1. Both codas and onsets show a linear relationship between log rank and log frequency characteristic of Zipfian distributions (see Appendix C). As a result, an enormous

[7] McGowan (in press) claims that the frequency of individual syllable contact clusters in English is proportional to the change of sonority over the syllable boundary. However, McGowan reports that the change in sonority in fact accounts for only a small portion of the variance in cluster frequency. A pilot study showed that sonority change was not useful as a predictor of cluster attestation.

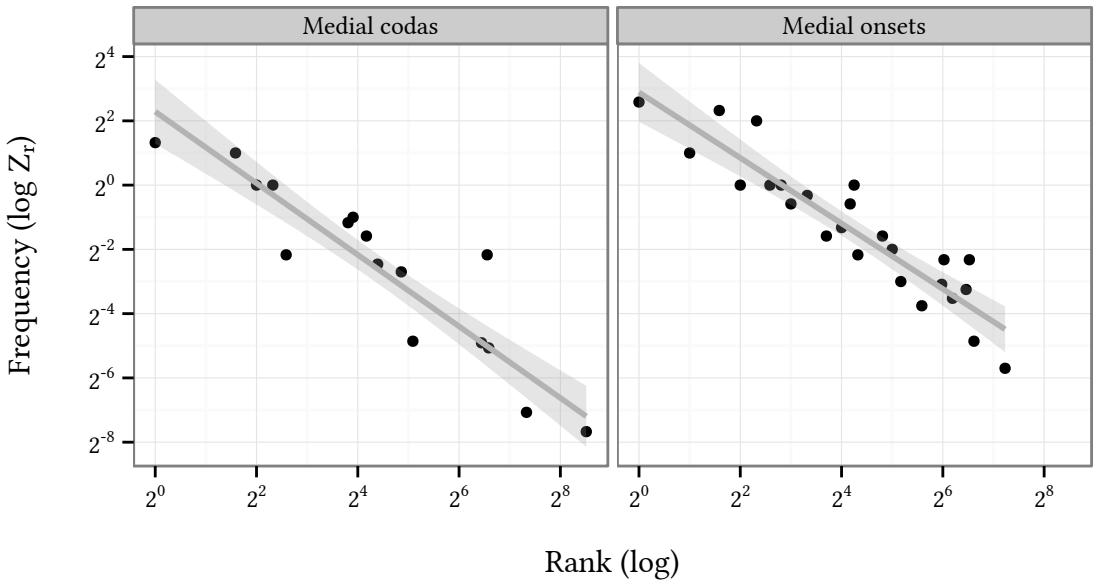

Figure 1: Medial coda and medial onset type frequencies in the lexical sample show a Zipfian distribution; frequencies have been smoothed using the Z_r transform (see Appendix C)

lexical sample would be needed to realize all clusters predicted by the medial cluster law, even if there were no constraints on the combination of medial codas and onsets in English. To illustrate this fact, a simulation is used to create new "samples" of the same size as the lexical sample used here. The following procedure is repeated so as to generate new "observations" for the simulated sample.

(8) Simulation procedure:

a. Sample a medial coda according to the observed probabilities
b. Sample a medial onset according to the observed probabilities
c. Apply the *SPE* rules to the cluster formed by their concatenation

This procedure corresponds to the assumption that the medial cluster law the derived constraints impose the only structural restrictions on the cluster inventory. Cluster frequencies in one simulated sample are shown in Figure 2; points represent simulated frequencies and the line actual cluster frequencies. As can be seen, the observed and simulated frequencies are quite similar (i.e., $R^2 = 0.712$; $p = 4.5\text{E}{-}05$). To summarize, the sparse cluster inventory, which Pierrehumbert takes as evidence for static constraints on syllable contact clusters, would result even if said static constraints do not exist.

4 Conclusions

The foregoing results suggest that the only structural constraints on English syllable contact are derived from the phonological system: this study finds no evidence for static constraints. The

many other unattested clusters can only be understood as accidental gaps which are a consequence of the finite nature of the English lexicon.

A Appendix: English syllable contact clusters

cluster	cluster frequency	coda frequency	onset frequency	−log p (MaxEnt)	rule exceptions
N . D	79	365	88	0.000	
N . T	69	365	150	0.000	
M . P	60	161	73	0.000	
M . B	53	161	65	0.000	
N . G	44	365	48	0.000	
S . T	37	87	150	0.000	
K . S	34	94	92	0.000	
N . S	33	365	92	0.000	
N . K	30	365	63	0.000	
N . Y	23	365	98	0.000	
S . K	18	87	63	0.000	
K . T	17	94	150	0.000	
L . Y	16	96	98	0.000	
K . Y	14	94	98	0.000	
N . V	12	365	18	0.000	NPA
L . K	11	96	63	0.000	
L . T	11	96	150	0.000	
N . JH	11	365	16	0.000	
M . Y	9	161	98	0.000	
G . M	9	29	36	0.000	
T . S	9	34	92	0.000	
S . P	9	87	73	0.000	
T . R	8	34	32	0.000	
N . D R	8	365	10	0.000	
N . T R	8	365	19	0.000	
S . T R	8	87	19	0.000	
K . N	7	94	28	0.000	
Z . M	7	18	36	0.000	
T . Y	7	34	98	0.000	
L . M	7	96	36	0.000	
N . Z	7	365	13	0.000	
N . F	7	365	20	0.000	NPA
M . B R	6	161	6	0.000	
Z . L	6	18	28	0.000	
G . N	6	29	28	0.000	
L . B	6	96	65	0.000	
L . D	6	96	88	0.000	
L . S	6	96	92	0.000	
G . Y	5	29	98	0.000	
P . Y	5	21	98	0.000	
L . V	5	96	18	0.000	
L . F	5	96	20	0.000	
B . Y	5	14	98	0.000	
F . R	5	15	32	0.000	
K . R	4	94	32	0.000	
M . Z	4	161	13	0.000	NPA
M . F	4	161	20	0.000	
P . T	4	21	150	0.000	
N . K W	4	365	5	0.000	
D . L	4	14	28	0.000	
F . T	4	15	150	0.000	
S . Y	4	87	98	0.000	
K . S T	3	94	10	0.000	
K . SH	3	94	7	0.000	
M . L	3	161	28	0.000	
M . P R	3	161	5	0.000	
M . P L	3	161	3	0.000	
M . N	3	161	28	0.000	

	M	.	S		3	161	92	0.000	NPA
	G	.	L		3	29	28	0.000	
	P	.	R		3	21	32	0.000	
	P	.	S		3	21	92	0.000	
	T	.	L		3	34	28	0.000	
	V	.	R		3	6	32	0.000	
N	K	.	T		3	3	150	0.000	
	L	.	G		3	96	48	0.000	
	L	.	P		3	96	73	0.000	
	N	.	F	R	3	365	5	0.000	NPA
	N	.	K	R	3	365	4	0.000	
	N	.	S	T	3	365	10	0.000	
	N	.	TH		3	365	5	2.831	
	D	.	N		3	14	28	0.000	
	F	.	Y		3	15	98	0.000	
	S	.	M		3	87	36	0.000	
	K	.	W		2	94	8	0.000	
	K	.	M		2	94	36	0.000	
	K	.	L		2	94	28	0.000	
	K	.	D		2	94	88	3.007	VAssim
	Z	.	Y		2	18	98	0.000	
	Z	.	B		2	18	65	3.193	
M	P	.	T		2	4	150	0.000	
	G	.	Z		2	29	13	3.193	
	G	.	R		2	29	32	0.000	
	P	.	N		2	21	28	0.000	
	T	.	N		2	34	28	0.000	
	T	.	F		2	34	20	2.298	
	V	.	Y		2	6	98	0.000	
	L	.	W		2	96	8	0.000	
	L	.	JH		2	96	16	0.000	
	L	.	S	T	2	96	10	0.000	
	L	.	N		2	96	28	2.113	
	L	.	T	R	2	96	19	0.000	
	N	.	CH		2	365	3	0.000	
	N	.	S	T R	2	365	3	0.000	
	N	.	G	W	2	365	2	0.000	
	N	.	HH		2	365	3	2.368	
	N	.	G	R	2	365	3	0.000	
	N	.	K	L	2	365	2	0.000	
	N	.	SH		2	365	7	0.000	
	B	.	L		2	14	28	0.000	
	B	.	R		2	14	32	0.000	
	B	.	JH		2	14	16	3.193	
	B	.	S		2	14	92	3.193	VAssim
	D	.	M		2	14	36	0.000	
	D	.	Y		2	14	98	0.000	
	S	.	F		2	87	20	2.214	
	K	.	S	W	1	94	1	0.000	
	K	.	S	T R	1	94	3	0.000	
	K	.	B		1	94	65	3.007	VAssim
	K	.	T	R	1	94	19	0.000	
	M	.	D	R	1	161	10	0.000	NPA
	M	.	K		1	161	63	0.000	NPA
	M	.	F	R	1	161	5	0.000	
	M	.	K	W	1	161	5	0.000	NPA
	M	.	R		1	161	32	0.000	
	M	.	T		1	161	150	0.000	NPA
	M	.	B	L	1	161	1	0.000	
	M	.	SH		1	161	7	2.250	NPA
	M	.	F	L	1	161	1	0.000	
	M	.	D		1	161	88	0.000	NPA
	Z	.	JH		1	18	16	3.193	
N	G	.	HH		1	5	3	5.561	VAssim
N	G	.	R		1	5	32	0.000	
N	G	.	T		1	5	150	3.193	VAssim
N	G	.	S	T	1	5	10	3.193	VAssim
N	G	.	S		1	5	92	3.193	VAssim
M	P	.	K		1	4	63	0.000	
M	P	.	S		1	4	92	0.000	
	G	.	W		1	29	8	0.000	
	G	.	B		1	29	65	3.193	

	P	K		1	21	63	0.000	
	P	M		1	21	36	3.112	
	P	L		1	21	28	0.000	
	P	S	T	1	21	10	0.000	
	T	W		1	34	8	1.998	
	T	K		1	34	63	0.000	
	T	M		1	34	36	0.000	
N	S	K	R	1	1	4	0.000	
	V	L		1	6	28	0.000	
N	CH	B		1	1	65	5.841	VAssim
	L	CH		1	96	3	0.000	
	L	D	R	1	96	10	0.000	
	L	F	R	1	96	5	0.000	
	L	R		1	96	32	2.113	
	L	G	R	1	96	3	0.000	
	L	P	R	1	96	5	0.000	
	L	SH		1	96	7	0.000	
	SH	M		1	3	36	2.835	
	SH	R		1	3	32	2.835	
	SH	T		1	3	150	2.835	
	N	S	L	1	365	1	0.000	
	N	L		1	365	28	2.113	
	N	S	K	1	365	1	0.000	
	N	TH	R	1	365	1	2.831	
	TH	M		1	3	36	2.831	
	TH	L		1	3	28	2.831	
	TH	Y		1	3	98	2.831	
N	T	M		1	1	36	0.000	
	B	N		1	14	28	0.000	
	D	P		1	14	73	3.193	VAssim
	D	R		1	14	32	0.000	
	D	V		1	14	18	3.193	
	DH	M		1	1	36	2.831	
D	Z	W		1	1	8	3.193	
	F	G		1	15	48	3.007	VAssim
	F	N		1	15	28	2.182	
	F	TH		1	15	5	5.046	
	S	W		1	87	8	0.000	
	S	L		1	87	28	0.000	
	S	P	R	1	87	5	0.000	
	S	N		1	87	28	2.182	
	S	TH		1	87	5	5.046	
	S	B		1	87	65	3.007	VAssim
	K	CH		0	94	3	2.005	
	K	S	L	0	94	1	0.000	
	K	D	R	0	94	10	3.007	VAssim
	K	K		0	94	63	1.928	Degem
	K	Z		0	94	13	3.007	VAssim
	K	F	R	0	94	5	2.298	
	K	G		0	94	48	4.935	Degem, VAssim
	K	P		0	94	73	2.298	
	K	G	W	0	94	2	4.935	Degem, VAssim
	K	HH		0	94	3	2.368	
	K	K	W	0	94	5	1.928	Degem
	K	JH		0	94	16	5.011	VAssim
	K	G	R	0	94	3	4.935	Degem, VAssim
	K	P	R	0	94	5	2.298	
	K	B	L	0	94	1	3.007	VAssim
	K	V		0	94	18	3.007	VAssim
	K	K	L	0	94	2	1.928	Degem
	K	K	R	0	94	4	1.928	Degem
	K	B	R	0	94	6	3.007	VAssim
	K	P	L	0	94	3	2.298	
	K	S	K	0	94	1	0.000	
	K	TH		0	94	5	2.831	
	K	F	L	0	94	1	2.298	
	K	TH	R	0	94	1	2.831	
	K	F		0	94	20	2.298	
	M	S	W	0	161	1	0.000	NPA
	M	W		0	161	8	2.570	
	M	CH		0	161	3	2.250	NPA
	M	S	L	0	161	1	0.000	NPA

	M . M		0	161	36	2.570	Degem
	M . S	T R	0	161	3	0.000	NPA
	M . G		0	161	48	1.862	NPA
	M . G	W	0	161	2	1.862	NPA
	M . HH		0	161	3	2.368	
	M . JH		0	161	16	0.000	NPA
	M . G	R	0	161	3	1.862	NPA
	M . V		0	161	18	1.777	
	M . K	L	0	161	2	0.000	NPA
	M . K	R	0	161	4	0.000	NPA
	M . S	T	0	161	10	0.000	NPA
	M . S	K	0	161	1	0.000	NPA
	M . TH		0	161	5	2.831	NPA
	M . TH	R	0	161	1	2.831	NPA
	M . T	R	0	161	19	0.000	NPA
	Z . S	W	0	18	1	5.407	Degem, VAssim
	Z . W		0	18	8	0.000	
	Z . CH		0	18	3	3.193	VAssim
	Z . S	L	0	18	1	5.407	Degem, VAssim
	Z . D	R	0	18	10	3.193	
	Z . K		0	18	63	3.193	VAssim
	Z . Z		0	18	13	5.407	Degem
	Z . F	R	0	18	5	5.407	VAssim
	Z . S	T R	0	18	3	5.407	Degem, VAssim
	Z . G		0	18	48	3.193	
	Z . P		0	18	73	3.193	VAssim
	Z . G	W	0	18	2	3.193	
	Z . HH		0	18	3	7.775	VAssim
	Z . K	W	0	18	5	3.193	VAssim
	Z . R		0	18	32	2.142	
	Z . G	R	0	18	3	3.193	
	Z . T		0	18	150	3.193	VAssim
	Z . P	R	0	18	5	3.193	VAssim
	Z . B	L	0	18	1	3.193	
	Z . V		0	18	18	5.407	
	Z . K	L	0	18	2	3.193	VAssim
	Z . K	R	0	18	4	3.193	VAssim
	Z . B	R	0	18	6	3.193	
	Z . P	L	0	18	3	3.193	VAssim
	Z . S	T	0	18	10	5.407	Degem, VAssim
	Z . SH		0	18	7	5.407	VAssim
	Z . S	K	0	18	1	5.407	Degem, VAssim
	Z . N		0	18	28	2.182	
	Z . TH		0	18	5	8.238	VAssim
	Z . F	L	0	18	1	5.407	VAssim
	Z . TH	R	0	18	1	8.238	VAssim
	Z . D		0	18	88	3.193	
	Z . F		0	18	20	5.407	VAssim
	Z . S		0	18	92	5.407	Degem, VAssim
	Z . T	R	0	18	19	3.193	VAssim
N	G . S	W	0	5	1	3.193	VAssim
N	G . W		0	5	8	0.000	
N	G . CH		0	5	3	3.193	VAssim
N	G . S	L	0	5	1	3.193	VAssim
N	G . D	R	0	5	10	3.193	
N	G . K		0	5	63	5.121	Degem, VAssim
N	G . M		0	5	36	0.000	
N	G . L		0	5	28	0.000	
N	G . Z		0	5	13	3.193	
N	G . F	R	0	5	5	3.193	VAssim
N	G . S	T R	0	5	3	3.193	VAssim
N	G . G		0	5	48	5.121	Degem
N	G . P		0	5	73	3.193	VAssim
N	G . G	W	0	5	2	5.121	Degem
N	G . K	W	0	5	5	5.121	Degem, VAssim
N	G . JH		0	5	16	3.193	
N	G . G	R	0	5	3	5.121	Degem
N	G . P	R	0	5	5	3.193	VAssim
N	G . B	L	0	5	1	3.193	
N	G . V		0	5	18	3.193	
N	G . K	L	0	5	2	5.121	Degem, VAssim

N	G . K	R	0	5	4	5.121	Degem, VAssim
N	G . B	R	0	5	6	3.193	
N	G . P	L	0	5	3	3.193	VAssim
N	G . Y		0	5	98	0.000	
N	G . SH		0	5	7	3.193	VAssim
N	G . S	K	0	5	1	3.193	VAssim
N	G . N		0	5	28	0.000	
N	G . TH		0	5	5	6.024	VAssim
N	G . F	L	0	5	1	3.193	VAssim
N	G . B		0	5	65	3.193	
N	G . TH	R	0	5	1	6.024	VAssim
N	G . D		0	5	88	3.193	
N	G . F		0	5	20	3.193	VAssim
N	G . T	R	0	5	19	3.193	VAssim
M	P . S	W	0	4	1	0.000	
M	P . W		0	4	8	3.112	
M	P . CH		0	4	3	4.255	
M	P . S	L	0	4	1	0.000	
M	P . D	R	0	4	10	3.007	VAssim
M	P . M		0	4	36	3.112	
M	P . L		0	4	28	0.000	
M	P . Z		0	4	13	3.007	VAssim
M	P . F	R	0	4	5	5.410	
M	P . S	T R	0	4	3	0.000	
M	P . G		0	4	48	3.007	VAssim
M	P . P		0	4	73	5.410	Degem
M	P . G	W	0	4	2	3.007	VAssim
M	P . HH		0	4	3	2.368	
M	P . K	W	0	4	5	0.000	
M	P . R		0	4	32	0.000	
M	P . JH		0	4	16	5.011	VAssim
M	P . G	R	0	4	3	3.007	VAssim
M	P . P	R	0	4	5	5.410	Degem
M	P . B	L	0	4	1	6.118	Degem, VAssim
M	P . V		0	4	18	7.896	VAssim
M	P . K	L	0	4	2	0.000	
M	P . K	R	0	4	4	0.000	
M	P . B	R	0	4	6	6.118	Degem, VAssim
M	P . P	L	0	4	3	5.410	Degem
M	P . S	T	0	4	10	0.000	
M	P . Y		0	4	98	0.000	
M	P . SH		0	4	7	2.250	
M	P . S	K	0	4	1	0.000	
M	P . N		0	4	28	0.000	
M	P . TH		0	4	5	2.831	
M	P . F	L	0	4	1	5.410	
M	P . B		0	4	65	6.118	Degem, VAssim
M	P . TH	R	0	4	1	2.831	
M	P . D		0	4	88	3.007	VAssim
M	P . F		0	4	20	5.410	
M	P . T	R	0	4	19	0.000	
	G . S	W	0	29	1	3.193	VAssim
	G . CH		0	29	3	3.193	VAssim
	G . S	L	0	29	1	3.193	VAssim
	G . D	R	0	29	10	3.193	
	G . K		0	29	63	5.121	Degem, VAssim
	G . F	R	0	29	5	3.193	VAssim
	G . S	T R	0	29	3	3.193	VAssim
	G . G		0	29	48	5.121	Degem
	G . P		0	29	73	3.193	VAssim
	G . G	W	0	29	2	5.121	Degem
	G . HH		0	29	3	5.561	VAssim
	G . K	W	0	29	5	5.121	Degem, VAssim
	G . JH		0	29	16	3.193	
	G . G	R	0	29	3	5.121	Degem
	G . T		0	29	150	3.193	VAssim
	G . P	R	0	29	5	3.193	VAssim
	G . B	L	0	29	1	3.193	
	G . V		0	29	18	3.193	
	G . K	L	0	29	2	5.121	Degem, VAssim
	G . K	R	0	29	4	5.121	Degem, VAssim

G . B	R	0	29	6	3.193	
G . P	L	0	29	3	3.193	VAssim
G . S	T	0	29	10	3.193	VAssim
G . SH		0	29	7	3.193	VAssim
G . S	K	0	29	1	3.193	VAssim
G . TH		0	29	5	6.024	VAssim
G . F	L	0	29	1	3.193	VAssim
G . TH	R	0	29	1	6.024	VAssim
G . D		0	29	88	3.193	
G . F		0	29	20	3.193	VAssim
G . S		0	29	92	3.193	VAssim
G . T	R	0	29	19	3.193	VAssim
P . S	W	0	21	1	0.000	
P . W		0	21	8	3.112	
P . CH		0	21	3	4.255	
P . S	L	0	21	1	0.000	
P . D	R	0	21	10	3.007	VAssim
P . Z		0	21	13	3.007	VAssim
P . F	R	0	21	5	5.410	
P . S	T R	0	21	3	0.000	
P . G		0	21	48	3.007	VAssim
P . P		0	21	73	5.410	Degem
P . G	W	0	21	2	3.007	VAssim
P . HH		0	21	3	2.368	
P . K	W	0	21	5	0.000	
P . JH		0	21	16	5.011	VAssim
P . G	R	0	21	3	3.007	VAssim
P . P	R	0	21	5	5.410	Degem
P . B	L	0	21	1	6.118	Degem, VAssim
P . V		0	21	18	7.896	VAssim
P . K	L	0	21	2	0.000	
P . K	R	0	21	4	0.000	
P . B	R	0	21	6	6.118	Degem, VAssim
P . P	L	0	21	3	5.410	Degem
P . SH		0	21	7	2.250	
P . S	K	0	21	1	0.000	
P . TH		0	21	5	2.831	
P . F	L	0	21	1	5.410	
P . B		0	21	65	6.118	Degem, VAssim
P . TH	R	0	21	1	2.831	
P . D		0	21	88	3.007	VAssim
P . F		0	21	20	5.410	
P . T	R	0	21	19	0.000	
T . S	W	0	34	1	0.000	
T . CH		0	34	3	5.818	
T . S	L	0	34	1	0.000	
T . D	R	0	34	10	4.913	Degem, VAssim
T . Z		0	34	13	3.007	VAssim
T . F	R	0	34	5	2.298	
T . S	T R	0	34	3	0.000	
T . G		0	34	48	3.007	VAssim
T . P		0	34	73	2.298	
T . G	W	0	34	2	3.007	VAssim
T . HH		0	34	3	2.368	
T . K	W	0	34	5	0.000	
T . JH		0	34	16	8.825	VAssim
T . G	R	0	34	3	3.007	VAssim
T . T		0	34	150	1.906	Degem
T . P	R	0	34	5	2.298	
T . B	L	0	34	1	3.007	VAssim
T . V		0	34	18	3.007	VAssim
T . K	L	0	34	2	0.000	
T . K	R	0	34	4	0.000	
T . B	R	0	34	6	3.007	VAssim
T . P	L	0	34	3	2.298	
T . S	T	0	34	10	0.000	
T . SH		0	34	7	1.907	
T . S	K	0	34	1	0.000	
T . TH		0	34	5	2.831	
T . F	L	0	34	1	2.298	
T . B		0	34	65	3.007	VAssim

	T . TH R	0	34	1	2.831		
	T . D	0	34	88	4.913	Degem, VAssim	
	T . T R	0	34	19	1.906	Degem	
N	S . S W	0	1	1	2.214	Degem	
N	S . W	0	1	8	0.000		
N	S . CH	0	1	3	3.912		
N	S . S L	0	1	1	2.214	Degem	
N	S . D R	0	1	10	3.007	VAssim	
N	S . K	0	1	63	0.000		
N	S . M	0	1	36	0.000		
N	S . L	0	1	28	0.000		
N	S . Z	0	1	13	5.221	Degem, VAssim	
N	S . F R	0	1	5	2.214		
N	S . S T R	0	1	3	2.214	Degem	
N	S . G	0	1	48	3.007	VAssim	
N	S . P	0	1	73	0.000		
N	S . G W	0	1	2	3.007	VAssim	
N	S . HH	0	1	3	4.583		
N	S . K W	0	1	5	0.000		
N	S . R	0	1	32	2.142		
N	S . JH	0	1	16	6.919	VAssim	
N	S . G R	0	1	3	3.007	VAssim	
N	S . T	0	1	150	0.000		
N	S . P R	0	1	5	0.000		
N	S . B L	0	1	1	3.007	VAssim	
N	S . V	0	1	18	5.221	VAssim	
N	S . K L	0	1	2	0.000		
N	S . B R	0	1	6	3.007	VAssim	
N	S . P L	0	1	3	0.000		
N	S . S T	0	1	10	2.214	Degem	
N	S . Y	0	1	98	0.000		
N	S . SH	0	1	7	4.121		
N	S . S K	0	1	1	2.214	Degem	
N	S . N	0	1	28	2.182		
N	S . TH	0	1	5	5.046		
N	S . F L	0	1	1	2.214		
N	S . B	0	1	65	3.007	VAssim	
N	S . TH R	0	1	1	5.046		
N	S . D	0	1	88	3.007	VAssim	
N	S . F	0	1	20	2.214		
N	S . S	0	1	92	2.214	Degem	
N	S . T R	0	1	19	0.000		
	V . S W	0	6	1	5.407	VAssim	
	V . W	0	6	8	3.112		
	V . CH	0	6	3	5.443	VAssim	
	V . S L	0	6	1	5.407	VAssim	
	V . D R	0	6	10	3.193		
	V . K	0	6	63	5.076	VAssim	
	V . M	0	6	36	3.112		
	V . Z	0	6	13	5.407		
	V . F R	0	6	5	8.519	Degem, VAssim	
	V . S T R	0	6	3	5.407	VAssim	
	V . G	0	6	48	3.193		
	V . P	0	6	73	6.304	VAssim	
	V . G W	0	6	2	3.193		
	V . HH	0	6	3	7.775	VAssim	
	V . K W	0	6	5	5.076	VAssim	
	V . JH	0	6	16	3.193		
	V . G R	0	6	3	3.193		
	V . T	0	6	150	3.193	VAssim	
	V . P R	0	6	5	6.304	VAssim	
	V . B L	0	6	1	6.304		
	V . V	0	6	18	10.296	Degem	
	V . K L	0	6	2	5.076	VAssim	
	V . K R	0	6	4	5.076	VAssim	
	V . B R	0	6	6	6.304		
	V . P L	0	6	3	6.304	VAssim	
	V . S T	0	6	10	5.407	VAssim	
	V . SH	0	6	7	7.657	VAssim	
	V . S K	0	6	1	5.407	VAssim	
	V . N	0	6	28	2.182		

V . TH	0	6	5	8.238	VAssim	
V . F L	0	6	1	8.519	Degem, VAssim	
V . B	0	6	65	6.304		
V . TH R	0	6	1	8.238	VAssim	
V . D	0	6	88	3.193		
V . F	0	6	20	8.519	Degem, VAssim	
V . S	0	6	92	5.407	VAssim	
V . T R	0	6	19	3.193	VAssim	
N CH . S W	0	1	1	2.835		
N CH . W	0	1	8	4.832		
N CH . CH	0	1	3	8.652	Degem	
N CH . S L	0	1	1	2.835		
N CH . D R	0	1	10	7.747	VAssim	
N CH . K	0	1	63	2.835		
N CH . M	0	1	36	2.835		
N CH . L	0	1	28	2.835		
N CH . Z	0	1	13	5.841	VAssim	
N CH . F R	0	1	5	5.133		
N CH . S T R	0	1	3	2.835		
N CH . G	0	1	48	5.841	VAssim	
N CH . P	0	1	73	5.133		
N CH . G W	0	1	2	5.841	VAssim	
N CH . HH	0	1	3	5.203		
N CH . K W	0	1	5	2.835		
N CH . R	0	1	32	2.835		
N CH . JH	0	1	16	11.659	Degem, VAssim	
N CH . G R	0	1	3	5.841	VAssim	
N CH . T	0	1	150	4.741		
N CH . P R	0	1	5	5.133		
N CH . B L	0	1	1	5.841	VAssim	
N CH . V	0	1	18	5.841	VAssim	
N CH . K L	0	1	2	2.835		
N CH . K R	0	1	4	2.835		
N CH . B R	0	1	6	5.841	VAssim	
N CH . P L	0	1	3	5.133		
N CH . S T	0	1	10	2.835		
N CH . Y	0	1	98	2.835		
N CH . SH	0	1	7	4.742		
N CH . S K	0	1	1	2.835		
N CH . N	0	1	28	2.835		
N CH . TH	0	1	5	5.666		
N CH . F L	0	1	1	5.133		
N CH . TH R	0	1	1	5.666		
N CH . D	0	1	88	7.747	VAssim	
N CH . F	0	1	20	5.133		
N CH . S	0	1	92	2.835		
N CH . T R	0	1	19	4.741		
N K . S W	0	3	1	0.000		
N K . W	0	3	8	0.000		
N K . CH	0	3	3	2.005		
N K . S L	0	3	1	0.000		
N K . D R	0	3	10	3.007	VAssim	
N K . K	0	3	63	1.928	Degem	
N K . M	0	3	36	0.000		
N K . L	0	3	28	0.000		
N K . Z	0	3	13	3.007	VAssim	
N K . F R	0	3	5	2.298		
N K . S T R	0	3	3	0.000		
N K . G	0	3	48	4.935	Degem, VAssim	
N K . P	0	3	73	2.298		
N K . G W	0	3	2	4.935	Degem, VAssim	
N K . HH	0	3	3	2.368		
N K . K W	0	3	5	1.928	Degem	
N K . R	0	3	32	0.000		
N K . JH	0	3	16	5.011	VAssim	
N K . G R	0	3	3	4.935	Degem, VAssim	
N K . P R	0	3	5	2.298		
N K . B L	0	3	1	3.007	VAssim	
N K . V	0	3	18	3.007	VAssim	
N K . K L	0	3	2	1.928	Degem	
N K . K R	0	3	4	1.928	Degem	

N	K . B	R	0	3	6	3.007	VAssim
N	K . P	L	0	3	3	2.298	
N	K . S	T	0	3	10	0.000	
N	K . Y		0	3	98	0.000	
N	K . SH		0	3	7	0.000	
N	K . S	K	0	3	1	0.000	
N	K . N		0	3	28	0.000	
N	K . TH		0	3	5	2.831	
N	K . F	L	0	3	1	2.298	
N	K . B		0	3	65	3.007	VAssim
N	K . TH	R	0	3	1	2.831	
N	K . D		0	3	88	3.007	VAssim
N	K . F		0	3	20	2.298	
N	K . S		0	3	92	0.000	
N	K . T	R	0	3	19	0.000	
	L . S	W	0	96	1	0.000	
	L . S	L	0	96	1	0.000	
	L . L		0	96	28	2.113	Degem
	L . Z		0	96	13	0.000	
	L . S	T R	0	96	3	0.000	
	L . G	W	0	96	2	0.000	
	L . HH		0	96	3	2.368	
	L . K	W	0	96	5	0.000	
	L . B	L	0	96	1	0.000	
	L . K	L	0	96	2	0.000	
	L . K	R	0	96	4	0.000	
	L . B	R	0	96	6	0.000	
	L . P	L	0	96	3	0.000	
	L . S	K	0	96	1	0.000	
	L . TH		0	96	5	2.831	
	L . F	L	0	96	1	0.000	
	L . TH	R	0	96	1	2.831	
	SH . S	W	0	3	1	5.049	
	SH . W		0	3	8	2.835	
	SH . CH		0	3	3	6.746	
	SH . S	L	0	3	1	5.049	
	SH . D	R	0	3	10	5.841	VAssim
	SH . K		0	3	63	2.835	
	SH . L		0	3	28	2.835	
	SH . Z		0	3	13	8.056	VAssim
	SH . F	R	0	3	5	5.049	
	SH . S	T R	0	3	3	5.049	
	SH . G		0	3	48	5.841	VAssim
	SH . P		0	3	73	2.835	
	SH . G	W	0	3	2	5.841	VAssim
	SH . HH		0	3	3	7.417	
	SH . K	W	0	3	5	2.835	
	SH . JH		0	3	16	9.753	VAssim
	SH . G	R	0	3	3	5.841	VAssim
	SH . P	R	0	3	5	2.835	
	SH . B	L	0	3	1	5.841	VAssim
	SH . V		0	3	18	8.056	VAssim
	SH . K	L	0	3	2	2.835	
	SH . K	R	0	3	4	2.835	
	SH . B	R	0	3	6	5.841	VAssim
	SH . P	L	0	3	3	2.835	
	SH . S	T	0	3	10	5.049	
	SH . Y		0	3	98	2.835	
	SH . SH		0	3	7	6.956	Degem
	SH . S	K	0	3	1	5.049	
	SH . N		0	3	28	5.017	
	SH . TH		0	3	5	7.880	
	SH . F	L	0	3	1	5.049	
	SH . B		0	3	65	5.841	VAssim
	SH . TH	R	0	3	1	7.880	
	SH . D		0	3	88	5.841	VAssim
	SH . F		0	3	20	5.049	
	SH . S		0	3	92	5.049	
	SH . T	R	0	3	19	2.835	
	N . S	W	0	365	1	0.000	
	N . W		0	365	8	2.570	
	N . M		0	365	36	2.570	

N	P		0	365	73	2.178	NPA
N	R		0	365	32	2.113	
N	P	R	0	365	5	2.178	NPA
N	B	L	0	365	1	2.178	NPA
N	B	R	0	365	6	2.178	NPA
N	P	L	0	365	3	2.178	NPA
N	N		0	365	28	2.113	Degem
N	F	L	0	365	1	0.000	NPA
N	B		0	365	65	2.178	NPA
TH	S	W	0	3	1	5.046	
TH	W		0	3	8	2.831	
TH	CH		0	3	3	6.743	
TH	S	L	0	3	1	5.046	
TH	D	R	0	3	10	5.838	VAssim
TH	K		0	3	63	2.831	
TH	Z		0	3	13	8.052	VAssim
TH	F	R	0	3	5	5.046	
TH	S	T R	0	3	3	5.046	
TH	G		0	3	48	5.838	VAssim
TH	P		0	3	73	2.831	
TH	G	W	0	3	2	5.838	VAssim
TH	HH		0	3	3	7.414	
TH	K	W	0	3	5	2.831	
TH	R		0	3	32	4.973	
TH	JH		0	3	16	9.750	VAssim
TH	G	R	0	3	3	5.838	VAssim
TH	T		0	3	150	2.831	
TH	P	R	0	3	5	2.831	
TH	B	L	0	3	1	5.838	VAssim
TH	V		0	3	18	8.052	VAssim
TH	K	L	0	3	2	2.831	
TH	K	R	0	3	4	2.831	
TH	B	R	0	3	6	5.838	VAssim
TH	P	L	0	3	3	2.831	
TH	S	T	0	3	10	5.046	
TH	SH		0	3	7	6.953	
TH	S	K	0	3	1	5.046	
TH	N		0	3	28	5.014	
TH	TH		0	3	5	7.877	Degem
TH	F	L	0	3	1	5.046	
TH	B		0	3	65	5.838	VAssim
TH	TH	R	0	3	1	7.877	Degem
TH	D		0	3	88	5.838	VAssim
TH	F		0	3	20	5.046	
TH	S		0	3	92	5.046	
TH	T	R	0	3	19	2.831	
N	T	S W	0	1	1	0.000	
N	T	W	0	1	8	1.998	
N	T	CH	0	1	3	5.818	
N	T	S L	0	1	1	0.000	
N	T	D R	0	1	10	4.913	Degem, VAssim
N	T	K	0	1	63	0.000	
N	T	L	0	1	28	0.000	
N	T	Z	0	1	13	3.007	VAssim
N	T	F R	0	1	5	2.298	
N	T	S T R	0	1	3	0.000	
N	T	G	0	1	48	3.007	VAssim
N	T	P	0	1	73	2.298	
N	T	G W	0	1	2	3.007	VAssim
N	T	HH	0	1	3	2.368	
N	T	K W	0	1	5	0.000	
N	T	R	0	1	32	0.000	
N	T	JH	0	1	16	8.825	VAssim
N	T	G R	0	1	3	3.007	VAssim
N	T	T	0	1	150	1.906	Degem
N	T	P R	0	1	5	2.298	
N	T	B L	0	1	1	3.007	VAssim
N	T	V	0	1	18	3.007	VAssim
N	T	K L	0	1	2	0.000	
N	T	K R	0	1	4	0.000	
N	T	B R	0	1	6	3.007	VAssim
N	T	P L	0	1	3	2.298	

N	T . S T	0	1	10	0.000	
N	T . Y	0	1	98	0.000	
N	T . SH	0	1	7	1.907	
N	T . S K	0	1	1	0.000	
N	T . N	0	1	28	0.000	
N	T . TH	0	1	5	2.831	
N	T . F L	0	1	1	2.298	
N	T . B	0	1	65	3.007	VAssim
N	T . TH R	0	1	1	2.831	
N	T . D	0	1	88	4.913	Degem, VAssim
N	T . F	0	1	20	2.298	
N	T . S	0	1	92	0.000	
N	T . T R	0	1	19	1.906	Degem
	B . S W	0	14	1	3.193	VAssim
	B . W	0	14	8	3.112	
	B . CH	0	14	3	5.443	VAssim
	B . S L	0	14	1	3.193	VAssim
	B . D R	0	14	10	3.193	
	B . K	0	14	63	3.193	VAssim
	B . M	0	14	36	3.112	
	B . Z	0	14	13	3.193	
	B . F R	0	14	5	6.304	VAssim
	B . S T R	0	14	3	3.193	VAssim
	B . G	0	14	48	3.193	
	B . P	0	14	73	6.304	Degem, VAssim
	B . G W	0	14	2	3.193	
	B . HH	0	14	3	5.561	VAssim
	B . K W	0	14	5	3.193	VAssim
	B . G R	0	14	3	3.193	
	B . T	0	14	150	3.193	VAssim
	B . P R	0	14	5	6.304	Degem, VAssim
	B . B L	0	14	1	6.304	Degem
	B . V	0	14	18	8.082	
	B . K L	0	14	2	3.193	VAssim
	B . K R	0	14	4	3.193	VAssim
	B . B R	0	14	6	6.304	Degem
	B . P L	0	14	3	6.304	Degem, VAssim
	B . S T	0	14	10	3.193	VAssim
	B . SH	0	14	7	5.443	VAssim
	B . S K	0	14	1	3.193	VAssim
	B . TH	0	14	5	6.024	VAssim
	B . F L	0	14	1	6.304	VAssim
	B . B	0	14	65	6.304	Degem
	B . TH R	0	14	1	6.024	VAssim
	B . D	0	14	88	3.193	
	B . F	0	14	20	6.304	VAssim
	B . T R	0	14	19	3.193	VAssim
	D . S W	0	14	1	3.193	VAssim
	D . W	0	14	8	1.998	
	D . CH	0	14	3	5.099	VAssim
	D . S L	0	14	1	3.193	VAssim
	D . D R	0	14	10	5.099	Degem
	D . K	0	14	63	3.193	VAssim
	D . Z	0	14	13	3.193	
	D . F R	0	14	5	3.193	VAssim
	D . S T R	0	14	3	3.193	VAssim
	D . G	0	14	48	3.193	
	D . G W	0	14	2	3.193	
	D . HH	0	14	3	5.561	VAssim
	D . K W	0	14	5	3.193	VAssim
	D . JH	0	14	16	5.099	
	D . G R	0	14	3	3.193	
	D . T	0	14	150	5.099	Degem, VAssim
	D . P R	0	14	5	3.193	VAssim
	D . B L	0	14	1	3.193	
	D . K L	0	14	2	3.193	VAssim
	D . K R	0	14	4	3.193	VAssim
	D . B R	0	14	6	3.193	
	D . P L	0	14	3	3.193	VAssim
	D . S T	0	14	10	3.193	VAssim
	D . SH	0	14	7	3.193	VAssim

	D . S K	0	14	1	3.193	VAssim	
	D . TH	0	14	5	6.024	VAssim	
	D . F L	0	14	1	3.193	VAssim	
	D . B	0	14	65	3.193		
	D . TH R	0	14	1	6.024	VAssim	
	D . D	0	14	88	5.099	Degem	
	D . F	0	14	20	3.193	VAssim	
	D . S	0	14	92	3.193	VAssim	
	D . T R	0	14	19	5.099	Degem, VAssim	
	DH . S W	0	1	1	8.238	VAssim	
	DH . W	0	1	8	2.831		
	DH . CH	0	1	3	6.024	VAssim	
	DH . S L	0	1	1	8.238	VAssim	
	DH . D R	0	1	10	6.024		
	DH . K	0	1	63	6.024	VAssim	
	DH . L	0	1	28	2.831		
	DH . Z	0	1	13	8.238		
	DH . F R	0	1	5	8.238	VAssim	
	DH . S T R	0	1	3	8.238	VAssim	
	DH . G	0	1	48	6.024		
	DH . P	0	1	73	6.024	VAssim	
	DH . G W	0	1	2	6.024		
	DH . HH	0	1	3	10.606	VAssim	
	DH . K W	0	1	5	6.024	VAssim	
	DH . R	0	1	32	4.973		
	DH . JH	0	1	16	6.024		
	DH . G R	0	1	3	6.024		
	DH . T	0	1	150	6.024	VAssim	
	DH . P R	0	1	5	6.024	VAssim	
	DH . B L	0	1	1	6.024		
	DH . V	0	1	18	8.238		
	DH . K L	0	1	2	6.024	VAssim	
	DH . K R	0	1	4	6.024	VAssim	
	DH . B R	0	1	6	6.024		
	DH . P L	0	1	3	6.024	VAssim	
	DH . S T	0	1	10	8.238	VAssim	
	DH . Y	0	1	98	2.831		
	DH . SH	0	1	7	8.238	VAssim	
	DH . S K	0	1	1	8.238	VAssim	
	DH . N	0	1	28	5.014		
	DH . TH	0	1	5	11.069	Degem, VAssim	
	DH . F L	0	1	1	8.238	VAssim	
	DH . B	0	1	65	6.024		
	DH . TH R	0	1	1	11.069	Degem, VAssim	
	DH . D	0	1	88	6.024		
	DH . F	0	1	20	8.238	VAssim	
	DH . S	0	1	92	8.238	VAssim	
	DH . T R	0	1	19	6.024	VAssim	
D	Z . S W	0	1	1	8.600	Degem, VAssim	
D	Z . CH	0	1	3	6.385	VAssim	
D	Z . S L	0	1	1	8.600	Degem, VAssim	
D	Z . D R	0	1	10	6.385		
D	Z . K	0	1	63	6.385	VAssim	
D	Z . M	0	1	36	3.193		
D	Z . L	0	1	28	3.193		
D	Z . Z	0	1	13	8.600	Degem	
D	Z . F R	0	1	5	8.600	VAssim	
D	Z . S T R	0	1	3	8.600	Degem, VAssim	
D	Z . G	0	1	48	6.385		
D	Z . P	0	1	73	6.385	VAssim	
D	Z . G W	0	1	2	6.385		
D	Z . HH	0	1	3	10.968	VAssim	
D	Z . K W	0	1	5	6.385	VAssim	
D	Z . R	0	1	32	5.334		
D	Z . JH	0	1	16	6.385		
D	Z . G R	0	1	3	6.385		
D	Z . T	0	1	150	6.385	VAssim	
D	Z . P R	0	1	5	6.385	VAssim	
D	Z . B L	0	1	1	6.385		
D	Z . V	0	1	18	8.600		
D	Z . K L	0	1	2	6.385	VAssim	

D	Z . K R	0	1	4	6.385	VAssim	
D	Z . B R	0	1	6	6.385		
D	Z . P L	0	1	3	6.385	VAssim	
D	Z . S T	0	1	10	8.600	Degem, VAssim	
D	Z . Y	0	1	98	3.193		
D	Z . SH	0	1	7	8.600	VAssim	
D	Z . S K	0	1	1	8.600	Degem, VAssim	
D	Z . N	0	1	28	5.375		
D	Z . TH	0	1	5	11.431	VAssim	
D	Z . F L	0	1	1	8.600	VAssim	
D	Z . B	0	1	65	6.385		
D	Z . TH R	0	1	1	11.431	VAssim	
D	Z . D	0	1	88	6.385		
D	Z . F	0	1	20	8.600	VAssim	
D	Z . S	0	1	92	8.600	Degem, VAssim	
D	Z . T R	0	1	19	6.385	VAssim	
	F . S W	0	15	1	2.214		
	F . W	0	15	8	3.112		
	F . CH	0	15	3	4.255		
	F . S L	0	15	1	2.214		
	F . D R	0	15	10	3.007	VAssim	
	F . K	0	15	63	1.884		
	F . M	0	15	36	3.112		
	F . L	0	15	28	0.000		
	F . Z	0	15	13	5.221	VAssim	
	F . F R	0	15	5	5.326	Degem	
	F . S T R	0	15	3	2.214		
	F . P	0	15	73	3.112		
	F . G W	0	15	2	3.007	VAssim	
	F . HH	0	15	3	4.583		
	F . K W	0	15	5	1.884		
	F . JH	0	15	16	5.011	VAssim	
	F . G R	0	15	3	3.007	VAssim	
	F . P R	0	15	5	3.112		
	F . B L	0	15	1	6.118	VAssim	
	F . V	0	15	18	10.11	Degem, VAssim	
	F . K L	0	15	2	1.884		
	F . K R	0	15	4	1.884		
	F . B R	0	15	6	6.118	VAssim	
	F . P L	0	15	3	3.112		
	F . S T	0	15	10	2.214		
	F . SH	0	15	7	4.464		
	F . S K	0	15	1	2.214		
	F . F L	0	15	1	5.326	Degem	
	F . B	0	15	65	6.118	VAssim	
	F . TH R	0	15	1	5.046		
	F . D	0	15	88	3.007	VAssim	
	F . F	0	15	20	5.326	Degem	
	F . S	0	15	92	2.214		
	F . T R	0	15	19	0.000		
	S . S W	0	87	1	2.214	Degem	
	S . CH	0	87	3	3.912		
	S . S L	0	87	1	2.214	Degem	
	S . D R	0	87	10	3.007	VAssim	
	S . Z	0	87	13	5.221	Degem, VAssim	
	S . F R	0	87	5	2.214		
	S . S T R	0	87	3	2.214	Degem	
	S . G	0	87	48	3.007	VAssim	
	S . G W	0	87	2	3.007	VAssim	
	S . HH	0	87	3	4.583		
	S . K W	0	87	5	0.000		
	S . R	0	87	32	2.142		
	S . JH	0	87	16	6.919	VAssim	
	S . G R	0	87	3	3.007	VAssim	
	S . B L	0	87	1	3.007	VAssim	
	S . V	0	87	18	5.221	VAssim	
	S . K L	0	87	2	0.000		
	S . K R	0	87	4	0.000		
	S . B R	0	87	6	3.007	VAssim	
	S . P L	0	87	3	0.000		
	S . S T	0	87	10	2.214	Degem	

S . SH	0	87	7	4.121	
S . S K	0	87	1	2.214	Degem
S . F L	0	87	1	2.214	
S . TH R	0	87	1	5.046	
S . D	0	87	88	3.007	VAssim
S . S	0	87	92	2.214	Degem

B English syllabification

For every entry in the CELEX database, there is a corresponding broad syllabified transcription of the word in a Received Pronunciation accent. This appendix describes an automated procedure used to process these transcripts and to separate medial clusters from their flanking nuclei, parsing the resulting sequences into coda and onset, and reversing allophonic processes targeting medial clusters.

While the segmental content of these transcriptions is precise, the CELEX syllabifications are unsystematic. Given the absence of contrastive syllabification in English (if not all languages: see Blevins 1995:221, Elfner 2006), any sequence of a medial consonant cluster and its flanking nuclei should receive the same syllabification in all words in which it occurs. This is not always the case with the CELEX transcriptions, however. For instance, the sequence [ɪstɹɪ] receives a different parse in *chemistry* ['kɛ.mɪ.stɹɪ] and *ministry* ['mɪ.nɪs.tɹɪ].[8] Consequently, these syllabifications are not used here.

B.1 Ambiguous segments

The syllabification procedure begins by separating sequences of vocalic and consonantal segments. In English, *r* and onglides pattern with consonants or with vowels depending on the context in which they occur. The heuristic adopted here is that ambiguous segments which impose restrictions on adjacent vowels are themselves vocalic, and those which impose restrictions on adjacent consonants are consonantal.

Initially, between two vowels, or finally, *r* is consonantal. Before another consonant, however, *r* has been lost in Received Pronunciation. Even in *r*-ful dialects, though, post-vocalic non-onset *r* patterns with vowels, not coda consonants. Before non-onset *r* many vowel contrasts are suspended (e.g., Fudge 1969:269f., Harris 1994:255): compare American English *fern*/*fir*/*fur* to *pet*/*pit*/*putt*. In this position, *r* is the only consonant which permits variable glottalization of a following /t/ in *r*-ful British dialects (Harris 1994:258), and the only consonant which does not trigger variable deletion of a following word-final /t, d/ in American dialects (Guy 1980:8). This is shown in (9–10) below.

(9) /t/-Glottalization in *r*-ful British dialects:

a.	des[ɚt]	∼	des[ɚʔ]
	c[ɚt]ain	∼	c[ɚʔ]ain
b.	fi[st]	∼	*fi[sʔ]
	mi[st]er	∼	*mi[sʔ]er

[8] Note that word-final *y* is usually lax in Received Pronunciation (Wells 1982:II.294).

(10) /t, d/-Deletion in American English:

a.	be[lt]	∼	be[l]
	me[nd]	∼	me[n]
b.	sk[ɚt]	∼	*sk[ɚ]
	th[ɚd]	∼	*th[ɚ]

Following Pierrehumbert (1994), pre-consonantal *r* is assigned to the preceding nucleus.

The front onglide is assigned to onset position when initial or preceded by a single consonant, as in [j]*arn* or *ju*[n.j]*or*. When the glide is preceded by two or more consonants, it is assigned to the nucleus. There is considerable evidence in support of this assumption. When [j] is assigned to the onset, it may be followed by any vowel (Borowsky 1986:276), but when it is nuclear, the following vowel is always [u], suggesting a nuclear affiliation (Harris 1994:61f., Hayes 1980:232). Clements and Keyser (1983:42) note that [j] is the only consonant which can follow onset /m/ and /v/: [mj]*use*, [vj]*iew*. Finally, [ju] sequences in words such as *spew* behave as a unit in language games (Davis and Hammond 1995, Nevins and Vaux 2003) and speech errors (Shattuck-Hufnagel 1986:130).[9]

The phonotactic properties of the back onglide [w] are quite different than those of the front onglide, and it is consequently assigned to the onset portion of medial clusters. Whereas [j] shows only limited selectivity for preceding tautosyllabic consonants (Kaye 1996), [w] only rarely occurs after onset consonants other than [k] (e.g., *tran*[kw]*il*), and never after tautosyllabic labials in the native vocabulary. Whereas [kj] is always followed by [u], [kw] may precede nearly any vowel (Davis and Hammond 1995:161).

B.2 Parsing medial consonant clusters

Medial consonant clusters are segmented into coda and onset using a heuristic version of the principle of onset maximization (e.g., Kahn 1976:42f., Kuryłowicz 1948, Pulgram 1970:75, Selkirk 1982:358f.) which favors parses of word-medial clusters in which as much of the cluster as possible is assigned to the onset. A medial onset is defined to be "possible" simply if it occurs word-initially (according to the rules defined above). As an example, the medial clusters in words such as *neu*[.tɹ]*on* or *bi*[.stɹ]*o* also occur in word-initial position (e.g., [tɹ]*ain*, [stɹ]*ike*), so the entire cluster is assigned ot the onset. In contrast, the cluster in *mi*[n.stɹ]*el* is not found word-initially; the maximal onset here is [stɹ] and the remaining [n] is assigned to the preceding coda.

In English, when a medial consonant cluster is preceded by a stressed lax vowel, as *wh*[ɪs.p]*er*, *v*[ɛs.t]*ige*, or *m*[ʌs.k]*et*, the first consonant of the cluster checks the lax vowel (Hammond 1997:3, Treiman and Zukowski 1990). As Harris (1994:55) notes, however, when the medial cluster is also a valid onset, as in *whi*[s.p]*er*, *ve*[s.ti]*ge*, and *mu*[s.k]*et*, onset maximization will incorrectly assign the entire cluster to the onset and leave the lax vowel unchecked. For this reason, onset maximization parses are modified to assign the first consonant of a complex medial consonant cluster to the coda before a stressed lax vowel (Pulgram 1970:48).

[9]The glide is also assumed to be present in underlying representation (e.g., Anderson 1988, Borowsky 1986:278) rather than inserted by rule (e.g., *SPE*:196, Halle and Mohanan 1985:89, McMahon 1990:217) since presence or absense of the glide is contrastive (e.g., *booty*/*beauty*, *coot*/*cute*).

B.3 Phonologization

Following Pierrehumbert (1994), the traditional analysis of affricates as single segment (e.g., *SPE*:321f., Jakobson et al. 1961:24) rather than sequences of a stop and fricative (e.g., Hualde 1988, Lombardi 1990) is adopted here. In many languages, affricates pattern with simple onsets; for instance, Classical Nahua bans true onset clusters but permits the affricate series [ts, tʃ, tɬ] (Launey 2011:9). Other languages, such as Polish, distinguish affricates and stop-fricative sequences (Brooks 1965), providing further evidence that "true" affricates are represented as single segments (or single timing units), and in contrast with stop-fricative clusters (Clements and Keyser 1983:34f.).

In English, [ŋ] has been analyzed as a pure allophone of /n/ before underlying /k, g/ (with later deletion of /g/ in some contexts; Borowsky 1986:65f., *SPE*:85, Halle and Mohanan 1985:62), or as a phoneme in its own right (e.g., Jusczyk et al. 2002, Sapir 1925). Onset [ŋ] is totally absent in onset position, where it cannot be followed by a /k, g/ needed to derive the velar allophone, a fact predicted only by the former account, and English speakers have considerable difficulty producing initial [ŋ] (Rusaw and Cole 2009). Following Pierrehumbert (1994), the allophonic analysis is assumed here. When followed by /k, g/, [ŋ] is mapped to /n/. When not followed by a velar stop (i.e., finally), [ŋ] is analyzed as underlying /ng/.

C Parameterizing Zipf's Law

Zipf (1949) famously observes the linear relationship between log rank r and log word frequency $f(r)$ in several linguistic samples. This is generalized below:

(11) $$f(C, \alpha) = \frac{C}{r^{\alpha}}$$

C is a constant, sensitive to sample size. Zipf assumes a 1-to-1 relationship, implying $\alpha = -1$, but it is possible to compute an optimal estimate for this parameter by taking the logarithm of both sides of the equation and solving for values of C and α that minimize the error term ε; this can be done efficiently with linear regression:

(12) $$\log f(r) \sim \log C + \alpha \log r + \varepsilon$$

Good (1953) notes that sparse distributions exhibit quantization at low frequencies, resulting in an artificially long flat right tail imposing an upward bias on estimates of α. Church and Gale (1991:29) propose a transform which eliminates this quantization. The vectors r, n are defined so such that n_i is the number of types which occur at frequency r_i (that is, n is a vector of frequencies of individual type frequencies). Z contains the elements of n by normalized by the points to the left and right.

(13) $$Z_i = \frac{2n_i}{r_{i+1} - r_{i-1}}$$

Church and Gale do not define this transform for the lowest and highest points (i.e., when $i = 1$

or N), but a natural extension of their definition is to scale the endpoints according to the next intermost point, as defined below.

$$Z_1 = \frac{n_1}{r_2 - r_1} \tag{14}$$

$$Z_N = \frac{n_N}{r_N - r_{N-1}} \tag{15}$$

The effect of applying this transform to sparse frequency data is shown in Figure 3.

References

Anderson, John M. 1988. More on slips and syllable structure. *Phonology* 5:157–159.

Anderson, Stephen R. 1973. Remarks on the phonology of English inflection. *Language and Literature* 1:33–52.

Anderson, Stephen R. 1974. *The organization of phonology*. New York: Academic Press.

Aronoff, Mark. 1976. *Word formation in generative grammar*. Cambridge: MIT Press.

Baayen, R. Harald, Richard Piepenbrock, and Léon Gulikers. 1996. CELEX2. Linguistic Data Consortium: LDC96L14.

Baković, Eric. 2005. Antigemination, assimilation and the determination of identity. *Phonology* 22:279–315.

Basbøll, Hans. 1972. Remarks on the regular plural formation of English nouns. *Language and Literature* 3:39–42.

Blevins, Juliette. 1995. The syllable in phonological theory. In *The handbook of phonological theory*, ed. John Goldsmith, 206–244. Oxford: Blackwell.

Bloomfield, Leonard. 1933. *Language*. New York: Henry Holt & Co.

Borowsky, Toni. 1986. Topics in the lexical phonology of English. Doctoral dissertation, University of Massachusetts, Amherst. Published by Garland, New York, 1991.

Brooks, Maria Zagorska. 1965. On Polish affricates. *Word* 20:207–210.

Brysbaert, Marc, and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41:977–990.

Chomsky, Noam, and Morris Halle. 1968. *The sound pattern of English*. Cambridge: MIT Press.

Church, Kenneth W., and William A. Gale. 1991. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language* 5:19–54.

Clements, George N., and Samuel Jay Keyser. 1983. *CV phonology: A generative theory of the syllable*. Cambridge: MIT Press.

Cortes, Corinna, and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* 20:273–297.

Côté, Marie-Hélène. 2000. Consonant cluster phonotactics: A perceptual approach. Doctoral dissertation, MIT.

Davidsen-Nielsen, Niels. 1974. Syllabification of English words with medial *sp*, *st*, *sk*. *Journal of Phonetics* 2:15–45.

Davis, Stuart, and Michael Hammond. 1995. On the status of onglides in American English. *Phonology* 12:159–182.

Duanmu, San. 2009. *Syllable structure: The limits of variation*. Oxford: Oxford University Press.

Elfner, Emily. 2006. Contrastive syllabification in Blackfoot. In *Proceedings of the 25th West Coast Conference on Formal Linguistics*, 141–149. Somerville, MA: Cascadilla.

Emmorey, Karen. 1989. Auditory morphological priming in the lexicon. *Language and Cognitive Processes* 4:73–92.

Ettlinger, Marc. 2008. Input-driven opacity. Doctoral dissertation, University of California, Berkeley.

Forster, Kenneth I., and Tamiko Azuma. 2000. Masked priming for prefixed words with bound stems: Does *submit* prime *permit*? *Language and Cognitive Processes* 15:539–561.

Fudge, Erik C. 1969. Syllables. *Journal of Linguistics* 5:253–286.

Good, Irving J. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika* 40:237–264.

Gropen, Jess, Steven Pinker, Michelle Hollander, Richard Goldberg, and Ronald Wilson. 1989. The learnability and acquisition of the dative alternation in English. *Language* 65:203–257.

Guy, Gregory R. 1980. Variation in the group and the individual: The case of final stop deletion. In *Locating language in time and space*, ed. William Labov, 1–35. New York: Academic Press.

Halle, Morris, and K. P. Mohanan. 1985. Segmental phonology of Modern English. *Linguistic Inquiry* 16:57–116.

Hammond, Michael. 1997. Vowel quantity and syllabification in English. *Language* 73:1–17.

Hammond, Michael. 1999. *The phonology of English*. Oxford: Oxford University Press.

Harley, Heidi. 2009. A morphosyntactic account of the 'Latinate' ban on dative shift in English. Paper presented at the University of California, Santa Cruz.

Harris, John. 1994. *English sound structure*. Cambridge: Blackwell.

Haugen, Einer. 1956. Syllabification in Kutenai. *International Journal of American Linguistics* 22:196–201.

Hayes, Bruce. 1980. A metrical theory of stress rules. Doctoral dissertation, MIT.

Hayes, Bruce, and James White. In press. Phonological naturalness and phonotactic learning. *Linguistic Inquiry* to appear.

Hayes, Bruce, and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39:379–440.

Hill, Archibald A. 1954. Juncture and syllable division in Latin. *Language* 30:439–447.

Hoard, James E., and Clarence Sloat. 1971. The inflectional morphology of English. *Glossa* 5:47–56.

Hockett, Charles F. 1958. *A course in modern linguistics*. London: Macmillan.

Hooper, Joan. 1973. Aspects of natural generative phonology. Doctoral dissertation, University of California, Los Angeles.

Hualde, Jose Ignacio. 1988. Affricates are not contour segments. In *Proceedings of the 7th West Coast Conference on Formal Linguistics*, 143–157. Stanford: Stanford Linguistics Association.

Isac, Daniela, and Charles Reiss. 2008. *I-language: An introduction to linguistics as cognitive science*. Oxford: Oxford University Press.

Jakobson, Roman, Gunnar Fant, and Morris Halle. 1961. *Preliminaries to speech analysis: The distinctive features and their correlates*. Cambridge: MIT Press.

Jusczyk, Peter W., Paul Smolensky, and Theresa Allocco. 2002. How English-learning infants respond to markedness and faithfulness constraints. *Language Acquisition* 10:31–37.

Kahn, Daniel. 1976. Syllable-based generalizations in English phonology. Doctoral dissertation, MIT. Published by Garland, New York, 1980.

Kaye, Jonathan. 1996. Do you believe in magic? The story of s+C sequences. In *A festschrift for Edmund Gussmann*, ed. Henryk Kardela and Bogdan Szymanek, 155–176. Lublin: Lublin University Press.

Kiparsky, Paul. 1985. Some consequences of lexical phonology. *Phonology Yearbook* 2:85–138.

Kuryłowicz, Jerzy. 1948. Contribution à la théorie de la syllabe. *Bulletin de la Société Polonaise de Linguistique* 8:80–114.

Labov, William. 1975. *What is a linguistic fact?* Lisse: Peter de Ridder.

Lass, Roger. 1971. Boundaries as obstruents. *Journal of Linguistics* 7:15–30.

Launey, Michel. 2011. *An introduction to Classical Nahuatl*. New York: Cambridge University Press.

Lightner, Theodore M. 1970. On regular inflectional endings in English. *Papers in Linguistics* 5:503–519.

Lombardi, Linda. 1990. The nonlinear organization of the affricate. *Natural Language and Linguistic Theory* 8:375–425.

Luelsdorff, Philip A. 1969. On the phonology of English inflection. *Glossa* 3:39–48.

McGowan, Kevin B. In press. Gradient lexical reflexes of the syllable contact law. Paper presented at CLS 45 to appear in the proceedings.

McMahon, April. 1990. Vowel shifts, free rides and strict cyclicity. *Lingua* 80:197–225.

Miner, Kenneth L. 1975. English inflectional endings and unordered rules. *Foundations of Language* 12:339–365.

Moulton, William G. 1947. Juncture in Modern Standard German. *Language* 23:212–226.

Nevins, Andrew, and Bert Vaux. 2003. Metalinguistic, shmetalinguistic: The phonology of shm-reduplication. In *Papers from the 39th meeting of the Chicago Linguistic Society*, 702–721. Chicago: Chicago Linguistic Society.

Newman, Stanley. 1944. *Yokuts language of California*. New York: Viking Fund Publications in Anthropology.

Nida, Eugene A. 1948. The identification of morphemes. *Language* 24:414–441.

Pande, Hemlata, and Hoshiyar S. Dhami. 2010. Mathematical modeling of occurrence of letters and word initials in texts in Hindi. *SKASE Journal of Theoretical Linguistics* 7:19–38.

Pierrehumbert, Janet. 1994. Syllable structure and word structure: A study of triconsonantal clusters in English. In *Phonological structure and phonetic form: Papers in Laboratory Phonology III*, ed. Patricia A. Keating, 168–188. Cambridge: Cambridge University Press.

Pinker, Steven, and Alan Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. In *Connections and Symbols*, ed. Steven Pinker and Jacques Mehler, 73–193. Cambridge: MIT Press.

Pulgram, Ernst. 1970. *Syllable, word, nexus, cursus*. The Hague: Mouton.

Rice, Keren. 2009. Athapaskan: Slave. In *The Oxford handbook of compounding*, ed. Rochelle Lieber and Pavol Stekauer, 542–573. Oxford: Oxford University Press.

Rusaw, Erin, and Jennifer Cole. 2009. Learning constraints that oppose native phonotactics from brief experience. Paper presented at the Mid-Continental Workshop on Phonology.

Sapir, Edward. 1925. Sound patterns in language. *Language* 1:37–51.

Schütze, Carson T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.

Selkirk, Elisabeth O. 1982. The syllable. In *The structure of phonological representations*, ed. Harry van der Hulst and Norval Smith, 337–385. Dordrecht: Foris.

Shattuck-Hufnagel, Stefanie. 1986. The representation of phonological information during speech production planning: Evidence from vowel errors in spontaneous speech. *Phonology Yearbook* 3:117–149.

Shibatani, Masayoshi. 1972. The phonological representations of English inflectional endings. *Glossa* 6:117–127.

Sigurd, Bengt. 1968. Rank-frequency distributions for phonemes. *Phonetica* 18:1–15.

Taft, Marcus. 1979. Recognition of affixed words and the word frequency effect. *Memory and Cognition* 7:263–272.

Taft, Marcus, and Sam Ardasinski. 2006. Obligatory decomposition in reading prefixed words. *The Mental Lexicon* 1:183–199.

Taft, Marcus, and Kenneth I. Forster. 1975. Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior* 14:638–647.

Taft, Marcus, and Kenneth I. Forster. 1976. Lexical storage and retrieval of polymorphemic and polysyllabic words. *Journal of Verbal Learning and Verbal Behavior* 15:607–620.

Taft, Marcus, Gail Hambly, and Sachiko Kinoshita. 1986. Visual and auditory recognition of prefixed words. *Quarterly Journal of Experimental Psychology* 38A:351–366.

Tambovtsev, Yuri, and Colin Martindale. 2007. Phoneme frequencies follow a Yule distribution. *SKASE Journal of Theoretical Linguistics* 4:1–11.

Treiman, Rebecca, and Andrea Zukowski. 1990. Towards an understanding of English syllabification. *Journal of Memory and Language* 29:66–85.

Trubetzkoy, Nikolai S. 1958. *Grundzüge der Phonologie*. Göttigen, Germany: Vandenhoeck & Ruprecht, 2nd edition.

Vaux, Bert. 2003. Syllabification in Armenian, Universal Grammar, and the lexicon. *Linguistic Inquiry* 34:91–125.

Weiss, Mary. 1961. Über die relative Häufigkeit der Phoneme des Schwedischen. *Statistical Methods in Linguistics* 1:41–55.

Wells, John C. 1982. *Accents of English*. Cambridge: Cambridge University Press. 3 volumes.

Zipf, George K. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge: Addison-Wesley.

Zuraw, Kie. 2003. Optimality Theory in linguistics. In *Handbook of brain theory and neural networks*, ed. Michael Arbib, 819–822. Cambridge: MIT Press, 2nd edition.

Zwicky, Arnold M. 1975. Settling on an underlying form: The English inflectional endings. In *Testing linguistic hypotheses*, ed. David Cohen and Jessica R. Wirth, 129–185. New York: Wiley.

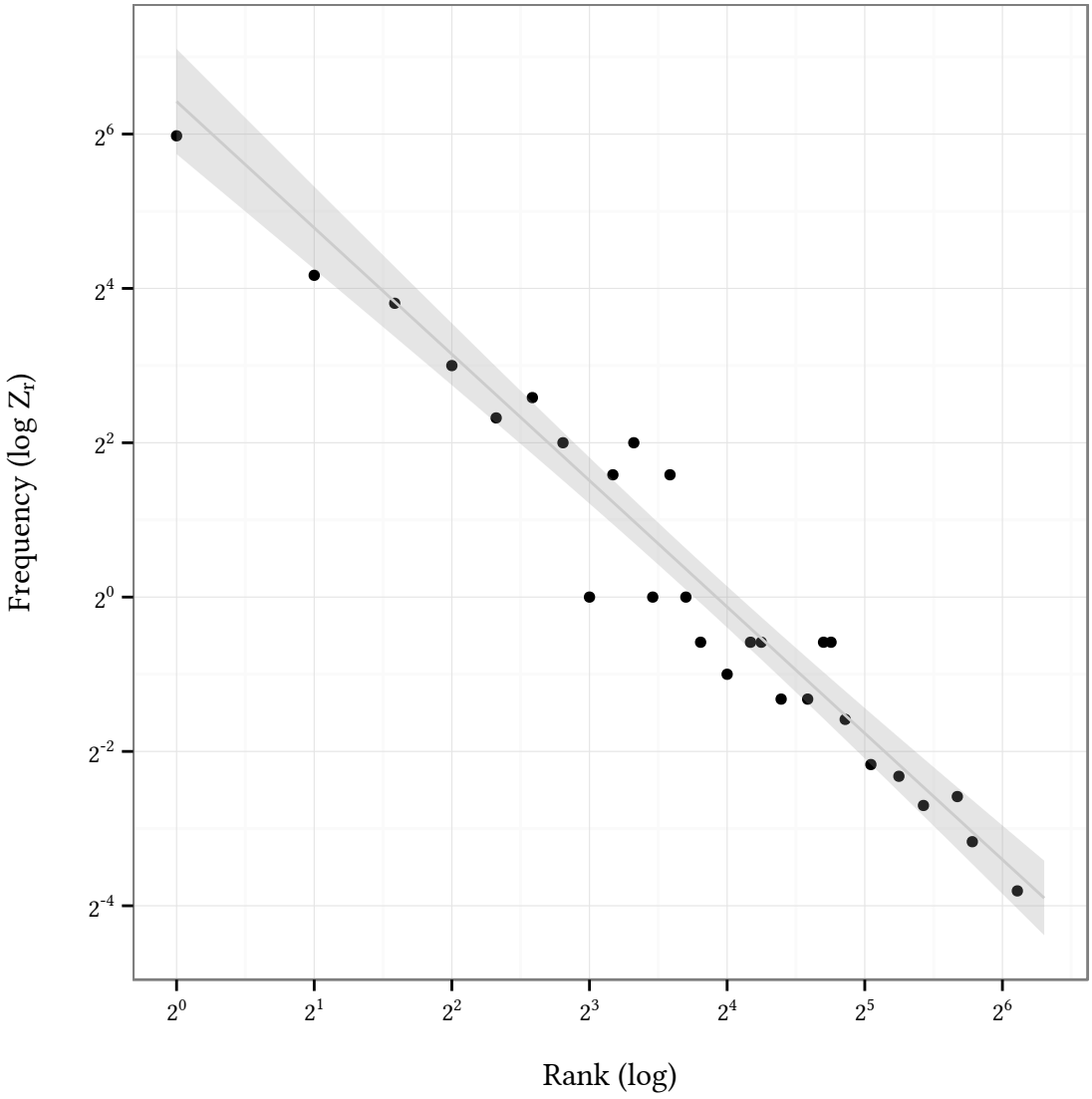

Figure 2: A simulated cluster inventory closely matches the observed cluster frequencies (represented by the smoothing line); frequencies have been smoothed using the Z_r transform (see Appendix C)

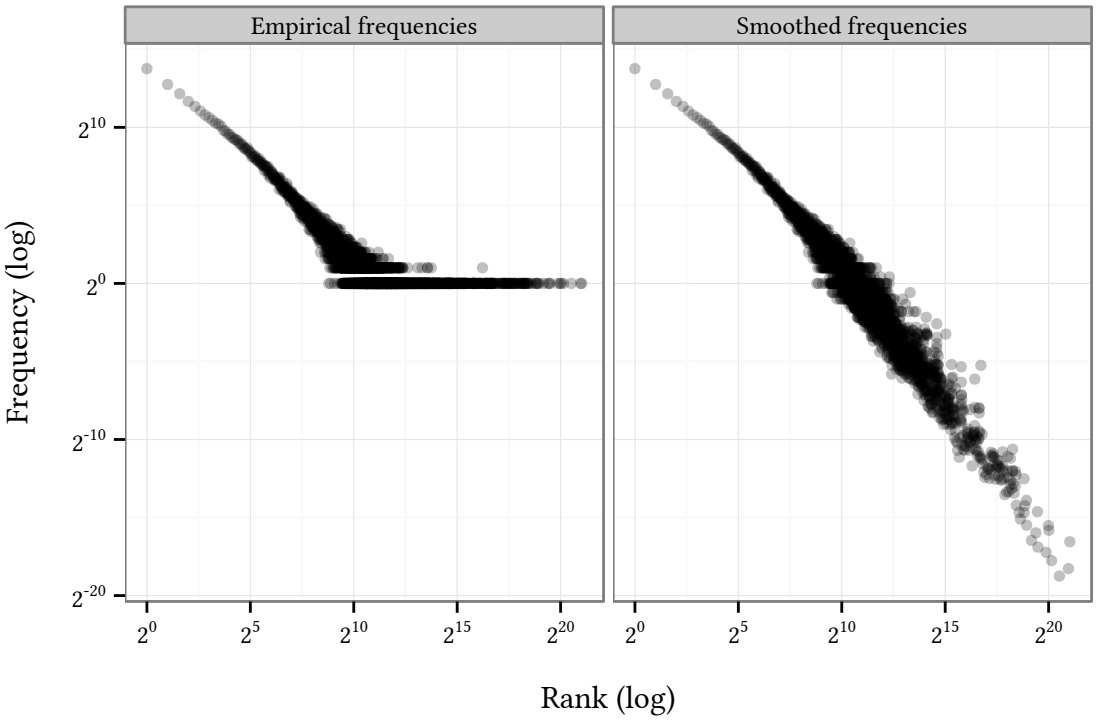

Figure 3: Left panel: word frequencies from the SUBTLEX-US frequency norms (Brysbaert and New 2009); Right panel: the same data smoothed with the Z_r transform