Chapter 1

Gradient and categorical aspects of wordlikeness judgements

Halle (1962) and Chomsky and Halle (1965) observe that speakers internalize generalizations about possible and impossible words in their language. Chomsky and Halle illustrate this knowledge with two nonce words *blick* [blɪk] and *bnick* [bnɪk]; whereas *blick* is unattested, but possible word of English, speakers immediately recognize *bnick* to be ill-formed, an impossible word of the language.

The classic account of this contrast is as follows. Segments must be assigned to prosodic structures like the syllable, or undergo phonological repair (e.g., Hooper 1973:10f., Itô 1989, Noske 1992). In English, [bl] is a permissible onset, but unlike some languages (e.g., Moroccan Arabic *bni* 'building', *bnat* 'daughters', *bniqa* 'closet'), [bn] is not. Wolf and McCarthy (2009:19f.) suggest, for instance, that an underlying /bnɪk/ would be realized as [nɪk]; it is also possible to imagine resolution by prothesis—[əbnɪk]—or anapytxis—[bənɪk], and indeed Davidson (2006) finds that English speakers produce these repairs when attempting to mimic foreign pronunciations of obstruent-nasal onsets.

Many linguists view this account as overly simpliest, for, as Shademan argues, it draws

a categorical boundary between the possible and impossible which fails to match the granularity at which speakers may report wordlikeness intuitions.

> A defect of current grammatical acounts of phonotactics is that they render simple up-or-down decisions concerning well-formedness and cannot account for gradient judgements. But when judgements are elicited in a controlled fashion from speakers, they always emerge as gradient, including all intermediate values. (Shademan 2006:371)

This chapter focuses on the hypothesis that speakers' knowledge of wordlikeness is gradient. §1.1 places this debate in a historical perspective, and argues that the naïve but widely accepted form of this hypothesis fails the test of falsifiability. In §1.2, a revision of this hypothesis is subjected to an quantitative evaluation and is found wanting.

[Explain that there's not going to be any issue of, e.g., heterovoiced obstruents, that stuff is all in chap. 4]

1.1 The gradience hypothesis

1.1.1 A brief history of the gradience hypothesis

Chomsky and Halle (1965) were not the first to consider contrasts between possible and impossible words. Their primary contribution to this question, is, essentially, mentalism: they recognize that naïve speakers effortlessly acquire language-specific wordlikeness generalizations and can report them with ease. But the nature of these contrasts were also considered by structuralists. In an early study, Fischer-Jørgensen (1952) proposes that possible and impossible words represent endpoints on a continuum, and that no non-arbitrary line can be drawn between them. While it would anachronistic to assign this a mentalistic

interpretation, this appears to be the earliest proposal that wordlikeness contrasts are in some sense gradient.

Early generative grammar

Gradience is also discussed in the earliest work in generative syntax. Chomsky (1955:132) writes that "there is little doubt that speakers can fairly consistently order new utterances, never previously heard, with respect to their degree of 'belongingness' to the language". Chomsky (1965) proposes to link different degrees of ungrammaticality to different kinds of syntactic violations. Unlike Fischer-Jørgensen, there is for Chomsky a hard-and-fast boundary between the grammatical and ungrammatical, only the latter showing any gradience. For Chomsky, grammatical utterances are all equally grammatical, but ungrammatical utterances may be deviant in their their own ways (Schütze 1996:61).

However, most of the early literature on wordlikeness concerns apparently categorical contrasts the possible and impossible. Vogt (1954:31) recognizes that the taxonomic phoneme is insufficient to account for many wordlikeness contrasts. Generalizing somewhat from Vogt's discussion, allophony may account for the absence of certain phone sequences, but it does not provide a suitable explanation for the absence of initial [bn] in English, nor does it make correct predictions about the surface realization of an underlying initial /bn/. Vogt concludes that additional grammatical machinery will be needed to account for possible and impossible words. Halle (1962) and Stanley (1967) propose to derive wordlikeness effects from morphophonemic redundancy rules acting on underlying representations. The discussion of *bnick* by Chomsky and Halle (1965:101) is a good example of this type of analysis. They correctly observe that before a word-initial stop, a English consonant must be liquid. A side effect of a redundancy rule which specifies a consonant in this position as [+Liquid] precludes the derivation of *bnick*.[1]

[1]This is not the only redundancy that might be used to rule out *bnick*; for instance, the only word-initial

The sound pattern of English

In *The sound pattern of English* (SPE), Chomsky and Halle (1968) extend this model of wordlikeness to account for

[Explain the SPE model]

Autosegmental phonology and beyond

The *SPE* model was not widely adopted, and research on wordlikeness turned to other principles. Syllabification plays no role in *SPE* theory, though the syllable is found in many earlier studies (see Goldsmith 2011). Hooper (1973) and Kahn (1976) argue that wordlikeness generalizations argue for the need for syllabification as part of the phonological computation. This classic syllabification-based model is further enriched by the autosegmental theory of the syllable (McCarthy 1979), which envisions the syllable as an articulated tree structure (as envisioned by Pike and Pike 1947), and the theory of prosodic licensing (Itô 1989), in which syllabification triggers phonological repairs.

1.1.2 A naïve gradience axiom

Recent critiques of this syllabification-based model focus on the existence of intermediate wordlikeness ratings (see also Coleman and Pierrehumbert 1997, Anttila 2008).

> In the particular domain of phonotactics gradient intuitions are pervasive: they have been found in every experiment that allowed participants to rate forms on a scale. (Hayes and Wilson 2008:382)

> When native speakers are asked to judge made-up (nonce) words, their intuitions are rarely all-or-nothing. In the usual case, novel items fall along a gradient cline of acceptability. (Albright 2009:9)

obstruent that may follow a nasal in English is /s/.

The observation itself cannot be questioned, but Hayes and Wilson and Albright do not explicitly state why this data is relevant to the construction of models of wordlikeness. The following seems the most likely: Hayes and Wilson and Albright believe this proves that wordlikeness, as an internal state, is gradient simply because subjects make use of intermediate degrees of wordlikeness in judgement tasks. This proposition, generalized below, is "naïve" not because it lacks sophistication, but because it is rooted in a belief in naïve realism, a philosophy which holds that perception provides a relatively direct picture of the nature of the world, and an influential view in the cognitive sciences (see Fodor and Pylyshyn 1981 for a critique).

(1) Naïve Gradience Proposition: If graded judgements of a concept or mental state use intermediate ratings, that concept or mental state is gradient

This proposition is an indicative conditional. While the antecedent of this condition, the presence or absence of intermediate ratings in a graded judgement task, is readily observable, it has no bearing on the truth or falsity of the conditional proposition. However, whereas a false consequent (e.g., an all-or-nothing concept) would potentially falsify the proposition, the consequent, a mental state, cannot be directly observed. If this is the case, then both the proposition and the consequent (the proposition that wordlikeness is itself gradient) fail the test of falsification, placing it beyond the reach of scientific inquiry.

Armstrong et al. (1983) cut the Gordian knot by asserting the possibility of observing a false consequent; they find the existence of all-or-nothing concepts like "odd number" or (more controversially) "female" to be apparent.

> Are there definitional concepts? Of course. For example. consider the concept *odd number*. This seems to have a clear definition, a precise description…No integer seems to sit on the fence, undecided as to whether it is quite even, or perhaps a bit odd. No odd number seems odder than any other odd number.

(Armstrong et al. 1983:274)

Armstrong et al. note that if subjects produce intermediate ratings for these categories, then the naïve proposition is false, and this is what Armstrong et al. find in their experiments. They ask subjects to rate, on a seven-point scale, the extent to which, e.g., certain odd counting numbers represent the concept "odd number", and find that speakers do make use of intermediate values. For instance, *7* is rated 1.4 out of 7 on average, whereas the average ratings of *447* and *501* are 3.5 and 3.7, respectively.

In summary, the proposition that intermediate ratings indicate mental gradience is either untestable, or it is false, and is thus rejected.

1.1.3 A falsifiable gradience hypothesis

Armstrong et al. do not view their results as evidence that subjects have an internal model of odd numbers which is at odds with the formal, all-or-nothing definition, and in fact, several additional experiments show that subjects deploy the extension of the formal definition in other tasks. Reviewing this study, Schütze (2011:215) writes that the experiments show that judgements are "sensitive to factors other than our underlying competence". In the case of wordlikeness, a case can be made that one such factor is similarity to existing words (e.g., Chomsky 1955:151, fn. 27, Greenberg and Jenkins 1964, Ohala and Ohala 1986, Bailey and Hahn 2001). However, the mere fact that subjects use intermediate ratings does not show that they do so with any systematicity, or that, e.g., the contrast between *447* and *7* with respect could be given any satisfying explanation. Schütze further suggests that gradience responses is a property of tasks, not concepts.

> …when asked for gradient responses, participants will find some way to oblige the experimenter; if doing so is incompatible with the experimenter's actual

> question, they apparently infer that she must have really intended to ask something slightly different. (Schütze 2011:215)

This introduces a troubling possibility, that subjects "oblige" experimenters by introducing random noise to their categorical judgements when presented with a Likert task. Were this the case, the observed intermediate judgements would scarcely be worthy of modeling. The alternative considered here is that there exists some model in which the gradience in ratings is systematic.

(2) Gradience Hypothesis: If graded judgements of a concept or mental state use intermediate ratings, there is some model which can predict this behavior

There are now many computational models of wordlikeness which proport to do just this. An unfortunate defect of prior evaluations of these models, however, is that they adopt the naïve gradience hypothesis. Hayes and Wilson (2008:382) are just some of the authors who consider the ability to model gradient intuitions to be so important that they do not include any categorical models in their evaluation. As a result, the literature contains no serious attempt to evaluate the gradience hypothesis.

1.2 Evaluation

The remainder of this chapter is devoted to evaluting the gradience hypothesis, using state-of-the-art computational wordlikeness models and data from English.

1.2.1 Data sources

The evaluation makes exclusive use of previously published English wordlikeness data. For inclusion in the evaluation, a study must conform to all three of the following conditions. First, the stimuli must consist only of monosyllabic words presented auditorily. Secondly,

a significant portion of the stimuli must contain gross phonotactic violations (e.g., *bnick*). Finally, the ratings, averaged across subjects, must be publicly available. The large-scale studies by Bailey and Hahn (2001) and Shademan (2006, 2007) are ineligible because the former lack stimuli with gross phonotactic violations and the latter data are not available to the public in any form.

Greenberg and Jenkins 1964

Greenberg and Jenkins (1964) investigated wordlikeness using the technique of free magnitude estimation, a mechanism which has become increasingly popular among syntactians (e.g., Bard et al. 1996). At the beginning of the experiment, the subject heard a recording of the word *stick*. In subsequent trials, the subjects heard a nonce word and were asked to report "how far would you say that is from English?", with *stick* at "1"; subjects are told that a word that is "twice as far from English" as *stick* should be scored "2". The data used here are from Greenberg and Jenkins's Experiment B, in which 17 undergraduates were presented 17 stimuli in all. In addition to *stick*, the stimuli include three other English words; these four items were excluded from further analyses, leaving 13 stimuli. As is standard practice in psychophysics (e.g., Butler et al. 2007), these magnitude ratings were log-transformed before analysis.

Scholes 1966

Scholes (1966) conducted a number of English wordlikeness judgement with middle school children. The data used here come from his "experiment 5", in which 63 monosyllabic items were presented to 33 seventh-grade students. For each stimulus, the subjects produced forced choice "yes"/"no" answers to the question of whether the item "is likely to be usable as a word of English". Hayes and Wilson (2008) and Albright (2009) analyse this data as gradient by performing an item averaging using the fraction of "yes" answers for each

stimulus (see also Pierrehumbert 1994, Coleman and Pierrehumbert 1997, Frisch et al. 2000). For instance, 22 of the 33 students answered "yes" for *shlerk* [ʃlɚk], so it is assigned a score of 0.666.[2] These stimuli are all ostensibly non-words, but include *clung* [klʌŋ], the preterite and past participle of the verb *cling*, and *brung* [brʌŋ], a dialectical past participle for *bring*. These two words were excluded and the remaining 61 stimuli were submitted to analysis.

Albright and Hayes 2003

Albright and Hayes (2003) gathered wordlikeness judgements to serve as norms for a *wug*-test. 87 items were presented to 20 undergraduate subjects, who rated each word on a seven-point Likert scale. The lowest point on the scale was labeled "completely bizarre, impossible as an English word", and that the highest point was labeled "complete normal, would make a fine English word".

1.2.2 Model comparison

The four models used here consist of a binary baseline and three computationally implemented gradient models. These three models are chosen because prior studies have shown they are correlated with wordlikeness judgements. Non-parametric rank correlation statistics are used evaluate the correlation between model predictions and wordlikeness ratings. Rank correlation are used rather than the parametric Pearson correlation statistic that is used by Hayes and Wilson (2008), for instance, because they make none of the potentially troublesome assumptions of the latter method (see Albright 2009:23, fn. 12). The Spearman ρ is most widely known rank correlation statistic, but it is difficult to give a natural interpretation to this quantity. On the other hand, the Kendall τ_b and Goodman-Kruskal γ can be interpreted as fractions of the number of *concordant* and *discordant* pairs (Noether

[2] This procedure conflates intraspeaker variation (which may be considerable; see Shademan 2007) and speaker-internal gradience, and its adoption here should not be construed as an endorsement. It is provided solely for comparison with prior studies using the Scholes data.

Spearman ρ	baseline	maxent	bigram	density
Greenberg and Jenkins 1964	0.845	0.765	**0.863**	0.648
Scholes 1966	0.791	0.762	**0.827**	0.827
Albright and Hayes 2003	0.725	0.429	0.708	**0.742**
Kendall τ_b	baseline	maxent	bigram	density
Greenberg and Jenkins 1964	**0.716**	0.585	0.692	0.462
Scholes 1966	**0.664**	0.597	0.652	0.565
Albright and Hayes 2003	**0.599**	0.343	0.506	0.556
Goodman-Kruskal γ	baseline	maxent	bigram	density
Greenberg and Jenkins 1964	**1.000**	0.684	0.692	0.462
Scholes 1966	**0.995**	0.634	0.667	0.614
Albright and Hayes 2003	**0.953**	0.656	0.509	0.575

Table 1.1: Rank correlations between wordlikeness ratings and phonotactic models surprisingly reveal that the binary baseline meeds or exceeds the coverage of three state-of-the-art phonotactic models. All correlations are significant at $p = 0.05$.

1981). Consider the case here, in which model scores are compared with wordlikeness ratings. If a model rates *dresp* [dɹɛsp] more wordlike than **srest* [sɹɛst], this pair is concordant if speakers also rate *dresp* more English-like than *srest*, and it is discordant if **srest* is rated more English-like. The τ_b and γ differ only in their treatment of "ties" (i.e., if e.g., **dresp* and **srest* are scored the same, or rated the same); τ_b includes a correction for ties, whereas γ ignores tied pairs. Much like the familiar Pearson correlation, ρ, τ_b and γ are all in the range [-1, 1]. Correlations of the four models are given in Table 1.2.2.

Binary baseline

The binary baseline used here is a crude implementation of the null hypothesis that there are no gradient effects in wordlikeness judgements. To create such a baseline, it is necessary to distinguish between nonce words which contain a gross phonotactic violation and those which do not. As all stimuli here are monosyllables, this task can be further simplified by separately considering the two major subcomponents of the syllable, the onset

and rime. Speakers are particularly adept as separating onsets from rimes (Treiman 1986, Treiman et al. 1995, Fowler et al. 1993), and a large portion of phonotactic violations can be localized to one or the other unit (e.g., Fudge 1969, Treiman et al. 2000).

The baseline considers a nonce syllable to be well-formed if it consists of both an attested onset and an attested rime; the free combination of these two components is the only mechanism by which this model can generalize beyond attested words. It is surely the case that this is an ultimately insufficient model: Albright (2009) observes, for instance, that [ɛsp] is a well-formed rime, even though it is found in no word of English. This problem is more acute in languages with more permissive syllable structures than those of English, and thus a sparser lexicon with regard to phonotactic constraints. For instance, Fischer-Jørgensen (1952) and Vogt (1954) assert that there are many accidental gaps (i.e., possible but unattested structures) in the inventory of consonant clusters in Georgian, a language which admits as many as five adjacent consonants. A similar result can be found in Chapter 4, which considers the inventory of syllable contact clusters in English. Despite these defects, the baseline outperforms the gradient models in most contexts.

In Figure 1.2.2, the densities of ratings from the three studies, linearly transformed to the interval [0, 1] and split according to this binary baseline, are shown in Figure 1.2.2. In all three studies, it is possible to discern a relatively sharp separation between valid and invalid clusters, but also the presence of intermediate values.

Maximum entropy phonotactics

Hayes and Wilson (2008; henceforth H&W) develop a sophisticated model of phonotactic grammaticality which estimates a probability distribution over phoneme sequences by weighing constraints according to the principle of maximum entropy. H&W find that the predictions of their model are closely correlated with the Scholes (1966) wordlikeness ratings. A direct replication of their predictions was attempted by using the software, model

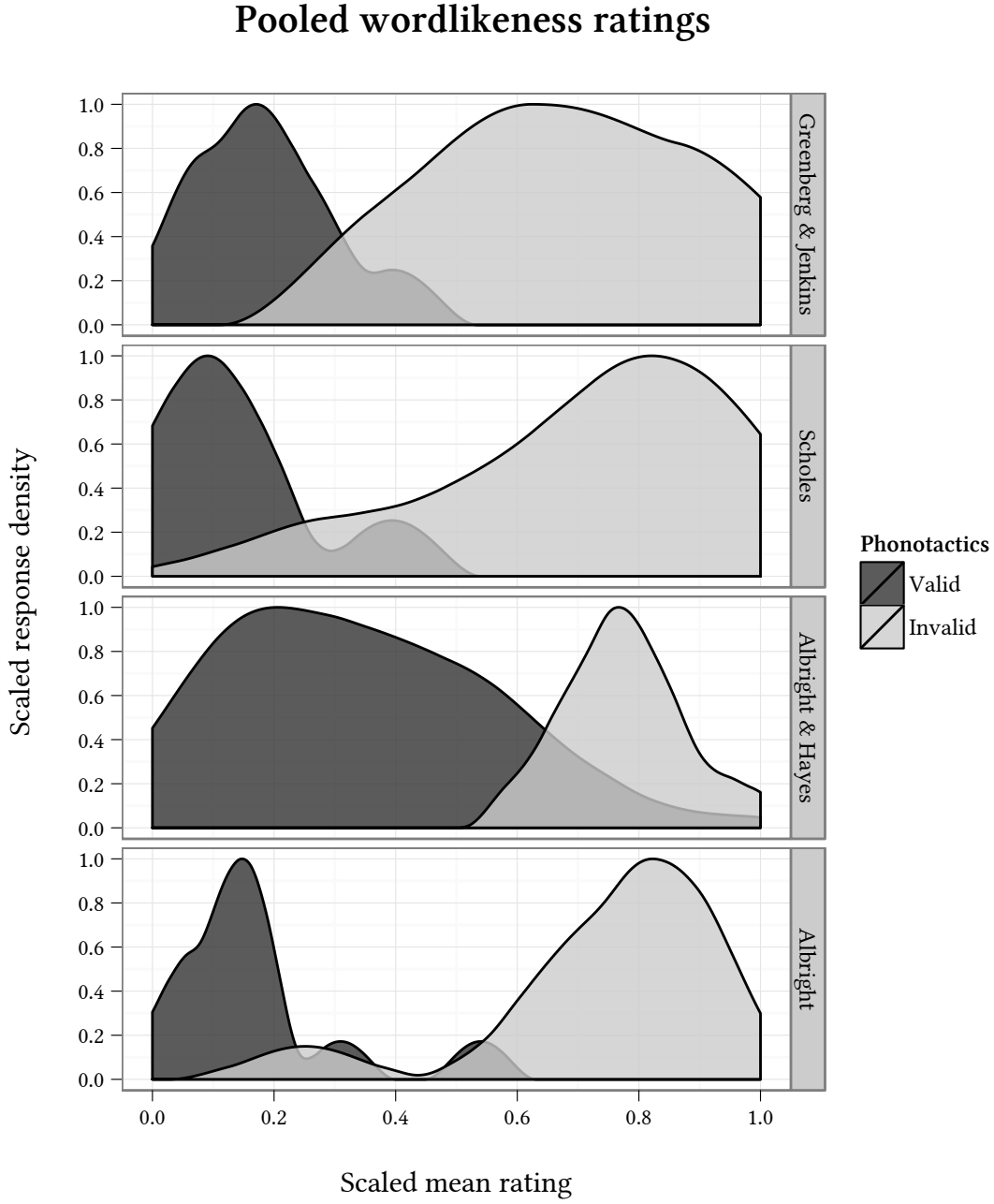

Figure 1.1: Average ratings of individual nonce words, linearly transformed to the interval [0, 1], tend to clump into two groups with little overlap; words which consist of attested onsets and rimes receive ratings near ceiling, whereas ratings of phonotactically invalid words are spread across the lower half of the spectrum.

parameters, and training data as described in that study. Since the training of the maximum entropy model is inherently stochastic, producing slightly different outcomes on each run, the lowest scoring of ten runs is reported (H&W:396), though in general there is not a great deal of variation between individual runs. One limitation of this model is that it is not feasible to score whole words, as the number of constraints which must be inspected grows expontentially as their scope increases. Following H&W and of Albright (2009), who also applies the maximum entropy model to the Albright and Hayes (2003) norms, the model is trained and scored only on stimulus onsets. However, as a consequence, the maximum entropy model performs particularly poorly on this data set, as many stimuli contain phonotactic violations in other positions.

Segment bigram probability

The bigram probability of a sequence *ijk* is the product of the probability of an sequence-initial *i*, the probability that *j* follows *i*, and the probability that *k* follows *j*, and the product of sequence-final *k*.

(3) $\hat{p}(ijk) = p(i|\text{start}) \cdot p(j|i) \cdot p(k|j) \cdot p(\text{stop}|k)$

Bigram models are widely used in natural language processing, and Albright (2009) considers their relevance to modeling wordlikeness judgements. While the focus of Albright's study is on developing a model which uses bigrams over phonological features rather than segments themselves, Albright's evaluation, which includes both the Scholes and Albright and Hayes data sets, finds an advantage for segmental bigrams.

Albright estimates bigram probabilities using the method of maximum likelihood over types in the lexicon. The variant of segmental bigrams used here computes probabilities with a simple type of smoothing in which the count of all possible bigrams (including those never observed) are incremented by one. This technique is known as Laplacian, or

	maximum likelihood	Laplacian smoothing
Spearman ρ	0.660	**0.708**
Kendall τ_b	0.467	**0.506**
Goodman-Kruskal γ	0.473	**0.509**

Table 1.2: Laplacian smoothing increases the correlation between the segmental bigram model proposed by Albright (2009), which uses maximum likelihood estimation, and the Albright and Hayes (2003) wordlikeness norms. All correlations are significant at $p = 0.05$.

"add one" smoothing. This has the desirable effect that no nonce word is ever assigned a zero probability, and produces in a small increase in the correlation between the Albright and Hayes wordlikeness norms compared with the maximum likelihood estimate (Table 1.2.2). For all three data sets, this model also consistenly outperforms "positional" probability models implemented by Vitevitch and Luce (2004) and Vaden et al. (2009), and thus these are not considered further. The bigram model consistently performs well in all the evaluations, and has the highest Spearman correlation with the Greenberg and Jenkins and Scholes, and is frequently second place model to the binary baseline elsewhere.

Neighborhood density

There are now many methods for computing similarity between nonce words and existing words, long thought to be reflected in wordlikeness judgements. For this study, a number of such methods were evaluated, including the Generalized Neighborhood Model (Bailey and Hahn 2001), PLD20 (Suárez et al. 2011), and a number of variations on neighborhood density (Coltheart et al. 1977) provided by Vaden et al. (2009). The best performance was obtained with the simplest version of neighborhood density, which is defined as the number of real monomorphemic words which can be changed into the target nonce word by a single insertion, deletion, or substitution of a phone. For instance, the neighbors of *blick* include *blink* (insertion), *lick* (deletion) and *black* (substitution). While many studies (e.g., Bailey and Hahn 2001) report robust lexical similarity effects, it may be that the relatively

weak performance of neighborhood density is the result of the presence of gross phonotactic violations.

1.2.3 Discussion

The gradience hypothesis

The primary result is that no gradient model reliably exceeds the accuracy of the binary baseline. Despite this, there are relatively strong correlations between the binary baseline and these gradient models (see Table 1.2.3). However, the fact that the gradient models are generally outperformed by the binary baseline suggest that they do not reliably predict intermediate ratings. To quantify this, the following method was used to estimate the residual contribution of the three gradient models once gross phonotactic violations are taken into account. Instead of calculating rank correlations directly on the model scores as in Table 1.2.2, the model scores are mapped to ranks with the additional constraint that all "valid" stimlui be ranked above all "invalid" stimuli. The resulting ranks are used to compute new correlation statistics. Finally, the binary baseline correlation is subtracted from this number, so that the resulting value is the amount of improvement derived from augmenting the binary model with gradience. These difference numbers are shown in Table 1.2.3. In most cases, including the gradient models on top of the binary baseline produces a worse correlation than is obtained with the binary baseline alone. The intepretation of this is direct, at least in the case of τ_b and γ. Each gradient model draws contrasts within the sets of phonotactically valid and invalid clusters, respectively. For instance, the bigram model favors *troog* [tɹuːg] over *swach* [swætʃ], though neither contains any gross phonotactic violation. Among these contrasts, however, the majority are not reflected in relative wordlikeness judgements. For instance, *troog* is rated less English-like than *swach*. This is stark evidence against the gradience hypothesis.

Kendall τ_b	maxent	bigram	density
Greenberg and Jenkins 1964	0.670	**0.680**	0.501
Scholes 1966	**0.685**	0.632	0.639
Albright and Hayes 2003	0.542	0.603	**0.623**

Table 1.3: The binary baseline is strongly correlated with the three gradient model scores; all correlations are significant at $p = 0.05$.

Δ Spearman ρ	maxent	bigram	density
Greenberg and Jenkins 1964	−0.060	**0.038**	−0.017
Scholes 1966	−0.029	**0.047**	−0.035
Albright and Hayes 2003	−0.008	−0.015	**0.018**
Δ Kendall τ_b	maxent	bigram	density
Greenberg and Jenkins 1964	−0.114	**−0.007**	−0.084
Scholes 1966	−0.067	**0.003**	−0.061
Albright and Hayes 2003	**−0.038**	−0.092	−0.049
Δ Goodman-Kruskal γ	maxent	bigram	density
Greenberg and Jenkins 1964	−0.268	**−0.260**	−0.337
Scholes 1966	−0.361	**−0.313**	−0.345
Albright and Hayes 2003	**−0.137**	−0.443	−0.386

Table 1.4: The change in rank correlation generated by augmenting the purely binary model with gradient predictions is small and in most cases it is negative.

Possible extensions to the binary baseline

The strong performance of the binary baseline should not be taken as evidence either that wordlikenesss judgements are binary, or that the binary baseline is a plausible model. The most serious limitation of this evaluation is the primitive nature of the binary baseline. The inability to generalization within onsets and rimes is a serious flaw, as is the assumption of independence of onset and rime. A cognitively plausible version of this model must entertain phonotactic generalizations that are larger than these units.

A possible further extension to the binary baseline would be the introduction of additional levels of wellformedness. While the evaluation has shown that current gradient models do not reliably identify intermediate wellformedness, it does seem possible to identify at least three levels of grammaticality: for instance, one might share the intuition that *zhlick* [ʒlɪk] is more similar to English than *bnick*, though both have unattested onsets.

There are precedents for labeling certain attested words as phonotactically "peripheral" (see, e.g., the appendices in Myers 1987 and Borowsky 1989); such words are regarded as lexical exceptions to language-general principles of syllabification. If this extends to nonce words, then an intermediate level of grammaticality could be assigned to "possible" but formally marked words. Another likely source of additional levels of grammaticality is the cumulative effect of multiple phonotactic violations. While, as Coleman and Pierrehumbert (1997) note, classic model predict taht a nonce word is as ill-formed as its worst deviation from syllable structure, it is possible to imagine that multiple phonotactic violations would result in greater degrees of ill-formedness. Many competing models make this prediction (e.g., Legendre et al. 1990, Levelt et al. 2000, Goldwater and Johnson 2003, Jäger 2007, Albright et al. 2008, Pater 2009) but despite this, there is currently no data bearing on whether cumulative effects are found in wordlikeness tasks.

1.3 Conclusion

This chapter has evaluated the axiom of gradience as a falsifiable alternative hypothesis. The surprising result is that virtually all of the apparent coverage of state-of-the-art gradient phonotactic models is simply a reflection of their ability to distinguish between the possible and the totally impossible; beyond this, they are unreliable. A trivial baseline, endowed with few abilities to project beyond the observed data, generally outperforms the state of the art. It follows that the projections made by the state-of-the-art gradient models are not like those made by speakers.

Appendix A

Data from Chapter 1

A.1 Greenberg and Jenkins (1964) experiment B

					lexical density	−log p (bigram)	−log p (MaxEnt)	phonotactic category	rating (MagE)
S	W	IH	T		19	15.975	0.000	valid	−25.10
K	L	AE	B		12	15.641	0.000	valid	−28.15
S	L	AH	K		14	13.905	0.000	valid	−29.16
S	W	AE	CH		3	20.051	0.000	valid	−29.25
B	R	AH	D		16	14.144	0.000	valid	−33.40
K	N	AE	P		5	20.884	10.877	invalid	−33.90
K	L	EH	B		4	17.325	0,000	valid	−32.92
T	R	UW	G		10	17.383	0.000	valid	−41.16
S	R	AH	M		9	20.200	5.010	invalid	−46.12
TH	Y	AH	NG		1	23.470	10.121	invalid	−46.49
TH	W	AE	ZH		0	27.819	3.876	invalid	−63.19
ZH	R	IH	K		8	29.197	13.640	invalid	−67.59
CH	W	UW	P		2	29.071	7.467	invalid	−87.97

A.2 Scholes (1966) experiment 5

					lexical density	−log p (bigram)	−log p (MaxEnt)	phonotactic category	rating (sum)
G	R	AH	N		17	11.915	0.000	valid	33
K	R	AH	N		13	11.467	0.000	valid	33
S	T	IH	N		17	11.124	0.000	valid	33
S	M	AE	T		10	16.197	0.000	valid	32
P	R	AH	N		8	11.660	0.000	valid	32
S	L	ER	K		6	16.821	0.000	valid	31

F	L	ER	K	10	16.428	0.000	valid	31
B	L	AH	NG	7	15.885	0.000	valid	31
D	R	AH	NG	5	16.207	0.000	valid	31
T	R	AH	N	7	11.767	0.000	valid	31
F	R	AH	N	10	12.330	0.000	valid	29
S	P	EY	L	15	14.271	0.000	valid	29
S	N	EH	T	6	17.026	0.000	valid	28
P	L	AH	NG	10	15.504	0.000	valid	28
SH	R	AH	K	4	16.073	2.204	valid	27
G	L	AH	NG	6	16.955	0.000	valid	27
M	R	AH	NG	1	21.038	11.466	invalid	27
SH	L	ER	K	2	26.724	4.989	invalid	22
S	K	IY	P	13	15.895	0.000	valid	20
V	R	AH	N	3	16.934	4.864	invalid	19
S	R	AH	N	8	19.164	5.042	invalid	14
V	L	ER	K	2	22.779	4.864	invalid	14
M	L	AH	NG	4	20.821	11.466	invalid	13
SH	T	IH	N	2	18.812	8.654	invalid	13
F	P	EY	L	5	23.502	6.968	invalid	13
ZH	R	AH	N	3	27.692	13.084	invalid	11
F	SH	IH	P	2	23.060	8.572	invalid	11
SH	N	EH	T	1	26.113	7.081	invalid	10
F	T	IH	N	4	14.455	6.968	invalid	10
Z	R	AH	N	4	24.289	12.637	invalid	8
N	R	AH	N	5	21.876	19.631	invalid	8
SH	M	AE	T	4	24.561	7.081	invalid	7
S	F	IY	D	7	20.262	6.704	invalid	7
Z	L	ER	K	2	23.286	8.675	invalid	6
Z	T	IH	N	1	25.677	9.637	invalid	6
F	S	EH	T	6	24.317	8.572	invalid	6
V	Z	IH	P	1	20.814	21.736	invalid	6
V	Z	AH	T	1	18.642	21.736	invalid	6
ZH	L	ER	K	2	32.314	16.949	invalid	5
SH	F	IY	D	2	26.008	13.266	invalid	5
Z	N	AE	T	1	28.400	9.222	invalid	4
F	N	EH	T	3	21.470	5.099	invalid	3
F	K	IY	P	1	25.609	6.968	invalid	3
V	T	IH	N	1	23.530	12.794	invalid	3
Z	V	IY	L	2	26.856	20.947	invalid	3
Z	M	AE	T	2	19.052	9.222	invalid	2
ZH	M	AE	T	2	22.476	19.589	invalid	2
F	M	AE	T	5	19.918	5.099	invalid	2
SH	P	EY	L	2	25.061	8.654	invalid	2
V	M	AE	T	2	24.601	10.511	invalid	1
V	N	EH	T	3	26.153	10.511	invalid	1
SH	K	IY	P	2	27.168	5.869	invalid	1
Z	P	EY	L	2	27.247	9.637	invalid	1
ZH	P	EY	L	2	30.650	21.576	invalid	1
ZH	T	IH	N	1	29.079	21.576	invalid	1
ZH	K	IY	P	1	32.757	17.712	invalid	1
ZH	N	EH	T	1	31.703	19.589	invalid	0
Z	K	IY	P	1	29.354	8.557	invalid	0
V	P	EY	L	6	25.100	12.794	invalid	0
V	K	IY	P	1	27.207	12.794	invalid	0
ZH	V	IY	L	1	30.259	30.795	invalid	0

A.3 Albright and Hayes (2003) norming study

					lexical density	−log p (bigram)	−log p (MaxEnt)	phonotactic category	rating (Likert)
S	L	EY	M		10	16.570	0.000	valid	5.84
W	IH	S			37	13.297	0.000	valid	5.84
P	IH	N	T		20	12.577	0.000	valid	5.67
P	AE	NG	K		17	11.868	0.000	valid	5.63
S	T	IH	P		15	12.998	0.000	valid	5.53
M	IH	P			34	12.581	0.000	valid	5.47
S	T	AY	R		9	15.698	0.000	valid	5.47
M	ER	N			40	12.140	0.000	valid	5.42
P	L	EY	K		15	16.127	0.000	valid	5.39
S	N	EH	L		9	16.220	0.000	valid	5.32
S	T	IH	N		17	11.124	0.000	valid	5.28
R	AE	S	K		12	14.624	0.000	valid	5.21
T	R	IH	S	K	3	16.760	0.000	valid	5.21
S	P	A	E	K	16	13.254	0.000	valid	5.16
D	EY	P			19	14.047	0.000	valid	5.11
G	EH	R			36	12.084	0.000	valid	5.11
G	L	IH	T		6	15.497	0.000	valid	5.11
S	K	EH	L		13	14.049	0.000	valid	5.11
SH	ER	N			26	14.441	0.000	valid	5.11
T	AA	R	K		14	16.148	0.000	valid	5.11
CH	EY	K			22	15.137	0.000	valid	5.05
G	L	IY	D		13	16.794	0.000	valid	5.05
G	R	AY	N	T	3	17.164	0.000	valid	5.00
P	R	IY	K		14	14.722	0.000	valid	5.00
SH	IH	L	K		3	17.983	0.000	valid	4.89
D	AY	Z			38	13.332	0.000	valid	4.84
N	EY	S			17	15.495	0.000	valid	4.84
T	AH	NG	K		18	13.400	0.000	valid	4.84
S	K	W	IH	L	5	17.684	0.000	valid	4.83
L	AH	M			35	10.329	0.000	valid	4.79
P	AH	M			28	10.006	0.000	valid	4.79
S	P	L	IH	NG	13	17.093	0.000	valid	4.72
G	R	EH	L		6	13.686	0.000	valid	4.63
T	EH	SH			8	15.421	0.000	valid	4.63
T	IY	P			29	12.914	0.000	valid	4.63
B	AY	Z			47	12.688	0.000	valid	4.58
G	L	IH	P		6	16.393	0.000	valid	4.53
CH	AY	N	D		15	16.768	0.000	valid	4.37
P	L	IH	M		6	14.320	0.000	valid	4.37
G	UW	D			21	14.519	0.000	valid	4.32
B	L	EY	F		6	17.997	0.000	valid	4.21
G	EH	Z			9	16.619	0.000	valid	4.21
D	R	IH	T		7	14.369	0.000	valid	4.16
F	L	IY	P		9	15.224	0.000	valid	4.16
Z	EY				35	13.153	2.769	valid	4.16
S	K	R	AY	D	4	17.976	0.000	valid	4.11
K	IH	V			12	14.008	0.000	valid	4.05
F	L	EH	T		15	14.449	0.000	valid	4.00
N	OW	L	D		23	19.131	0.000	valid	4.00
S	K	IH	K		13	14.146	0.000	valid	4.00
B	R	EH	JH		5	16.438	0.000	valid	3.95
K	W	IY	D		7	16.694	0.000	valid	3.95
S	K	OY	L		7	16.937	0.000	valid	3.89
D	R	AY	S		16	16.479	0.000	valid	3.84

F	L	IH	JH			5	16.594	0.000	valid	3.79
B	L	IH	G			4	15.487	0.000	valid	3.53
Z	EY	P	S			5	24.147	2.769	valid	3.47
CH	UW	L				13	14.085	0.000	valid	3.42
SH	AY	N	T			5	16.438	0.000	valid	3.42
SH	R	UH	K	S		4	22.453	2.127	invalid	3.32
G	W	EH	N	JH		1	23.165	2.929	invalid	3.32
N	AH	NG				15	14.633	0.000	valid	3.28
S	K	W	AA	L	K	0	24.319	0.000	invalid	3.26
T	W	UW				8	16.779	0.000	valid	3.17
S	M	AH	M			6	14.638	0.000	valid	3.05
S	N	OY	K	S		3	25.335	0.000	invalid	3.00
S	F	UW	N	D		0	23.906	6.703	invalid	2.94
P	W	IH	P			2	24.135	4.818	invalid	2.89
R	AY	N	T			8	14.815	0.000	valid	2.89
S	K	L	UW	N	D	0	20.721	3.046	invalid	2.83
S	M	IY	R	G		0	27.414	0.000	invalid	2.79
F	R	IH	L	G		2	22.117	0.000	invalid	2.68
SH	W	UW	JH			0	30.868	4.878	invalid	2.68
TH	R	OY	K	S		0	25.623	1.907	invalid	2.68
T	R	IH	L	B		1	20.903	0.000	invalid	2.63
K	R	IH	L	G		0	21.254	0.000	invalid	2.58
S	M	EH	R	G		0	21.873	0.000	invalid	2.58
TH	W	IY	K	S		2	25.867	3.879	invalid	2.53
S	M	EH	R	F		0	23.030	0.000	invalid	2.47
S	M	IY	L	TH		0	25.712	0.000	invalid	2.47
P	L	OW	M	F		0	21.857	0.000	invalid	2.42
D	W	OW	JH			0	24.150	2.929	invalid	2.29
P	L	OW	N	TH		0	20.893	0.000	invalid	2.26
TH	EY	P	T			3	23.473	1.907	invalid	2.26
S	M	IY	N	TH		0	23.098	0.000	invalid	2.06
S	P	R	AA	R	F	0	23.675	0.000	valid	2.05
P	W	AH	JH			0	25.463	4.818	invalid	1.74

Bibliography

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

Albright, Adam, and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90:119–161.

Albright, Adam, Giorgio Magri, and Jennifer Michaels. 2008. Modeling doubly marked lags with a split additive model. In *Proceedings of 32nd annual Boston University Conference on Lanugage Development*, ed. Harvey Chan, Heather Jacob, and Enkeleida Kapia, volume 1, 36–47. Somerville, MA: Cascadilla.

Anttila, Arto. 2008. Gradient phonotactics and the complexity hypothesis. *Natural Language and Linguistic Theory* 26:695–729.

Armstrong, Sharon L., Lila R. Gleitman, and Henry Gleitman. 1983. What some concepts might not be. *Cognition* 13:263–308.

Bailey, Todd M., and Ulrike Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods. *Journal of Memory and Language* 44:586–591.

Bard, Ellen G., Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72:32–68.

Borowsky, Toni. 1989. Structure preservation and the syllable coda in English. *Linguistic Inquiry* 7:145–166.

Butler, Gail, Linda M. Poste, Mark S. Wolynetz, Vivian E. Agar, and Elizabeth Larmond. 2007. Alternative analyses of magnitude estimation data. *Journal of Sensory Studies* 2:243–257.

Chomsky, Noam. 1955. The logical structure of linguistic theory. Ms., Harvard University and MIT. Revised version published by Plenum, New York, 1975.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge: MIT Press.

Chomsky, Noam, and Morris Halle. 1965. Some controversial questions in phonological theory. *Journal of Linguistics* 1:97–138.

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

Coleman, John, and Janet Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. In *3rd meeting of the ACL Special Interest Group in Computational Phonology: Proceedings of the workshop, 12 July 1997*, ed. John Coleman, 49–56. Somerset, NJ: Association for Computational Linguistics.

Coltheart, Max, Eddy J. Davelaar, Jon T. Jonasson, and Derek Besner. 1977. Access to the internal lexicon. In *Attention and performance VI*, ed. Stanislav Dornic, 535–555. Hillsdale, NJ: Lawrence Erlbaum.

Davidson, Lisa. 2006. Phonotactics and articulatory coordination interact in phonology: Evidence from non-native production. *Cognitive Science* 30:837–862.

Fischer-Jørgensen, Eli. 1952. On the definition of phoneme categories on a distributional basis. *Acta Linguistica* 7:8–39.

Fodor, Jerry A., and Zenon Pylyshyn. 1981. How direct is visual perception? Some reflections on Gibson's "ecological approach". *Cognition* 9:139–196.

Fowler, Carol A., Rebecca Treiman, and Jennifer Gross. 1993. The structure of English syllables and polysyllables. *Journal of Memory and Language* 32.

Frisch, Stefan A., Nathan R. Large, and David B. Pisoni. 2000. Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language* 42:481–496.

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

Goldsmith, John. 2011. The syllable. In *The handbook of phonological theory*, ed. John Goldsmith, Jason Riggle, and Alan C.L. Yu, 164–196. Malden, MA: Blackwell, 2nd edition.

Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the waorkshop on variation within Optimality Theory, Stockholm University*, ed. Jennifer Spenader, Anders Eriksson, and Osten Dahl, 111–120. Stockholm: Stockholm University.

Greenberg, Joseph H., and James J. Jenkins. 1964. Studies in the psychological correlates of the sound system of American English, I and II. *Word* 20:157–177.

Halle, Morris. 1962. Phonology in generative grammar. *Word* 18:54–72.

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

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

Itô, Junko. 1989. A prosodic theory of epenthesis. *Natural Language and Linguistic Theory* 7:217–259.

Jäger, Gerhard. 2007. Maximum entropy models and Stochastic Optimality Theory. In *Architectures, rules, and preferences: Variations on themes by Joan W. Bresnan*, ed. Annie Zaenen, Jane Simpson, Tracy H. King, Jane Grimshaw, Joan Maling, and Chris Manning, 467–479. Stanford, CA: CSLI.

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

Legendre, Geraldine, Yoshiro Miyata, and Paul Smolensky. 1990. Harmonic Grammar: A formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. Technical Report 90-5, Institute of Cognitive Science, University of Colorado, Boulder.

Levelt, Clara, Niels O. Schiller, and Willem J.M. Levelt. 2000. The acquisition of syllable types. *Language Acquisition* 8:237–264.

McCarthy, John J. 1979. On stress and syllabification. *Linguistic Inquiry* 10:443–465.

Myers, Scott. 1987. Vowel shortening in English. *Natural Language and Linguistic Theory* 5:485–518.

Noether, Gottfried E. 1981. Why Kendall tau? *Teaching Statistics* 3:41–41.

Noske, Roland. 1992. A theory of syllabification and segmental alternation, with studies on the phonology of French, German, Tonkawa and Yawelmani. Doctoral dissertation, Tilburg University. Revised version published by Niemeyer, Tübingen, 1993.

Ohala, John J., and Manjari Ohala. 1986. Testing hypotheses regarding the psychological manifestation of morpheme structure constraints. In *Experimental phonology*, ed. John J. Ohala and Jeri J. Jaeger, 239–252. Orlando: Academic Press.

Pater, Joe. 2009. Weighted constraints in generative linguistics. *Cognitive Science* 33:999–1035.

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.

Pike, Kenneth L., and Eunice V. Pike. 1947. Immediate constituents of Mazateco syllables. *International Journal of American Linguistics* 13:78–91.

Scholes, Robert J. 1966. *Phonotactic grammaticality*. Berlin: Mouton.

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

Schütze, Carson T. 2011. Linguistic evidence and grammatical theory. *Wiley Interdisciplinary Reviews: Cognitive Science* 2:206–221.

Shademan, Shabnam. 2006. Is phonotactic knowledge grammatical knowledge? In *Proceedings of the 25th West Coast Conference on Formal Linguistics*, ed. Donald Baumer, David Montero, and Michael Scanlon, 371–379. Somerville, MA: Cascadilla.

Shademan, Shabnam. 2007. Grammar and analogy in phonotactic well-formedness. Doctoral dissertation, University of California, Los Angeles.

Stanley, Richard. 1967. Redundancy rules in phonology. *Language* 43:393–436.

Suárez, Lidia, Seok Hui Tan, Melvin J. Yap, and Winston D. Goh. 2011. Observing neighborhood effects without neighbors. *Psychonomic Bulletin and Review* 18:605–611.

Treiman, Rebecca. 1986. The division between onsets and rimes in English syllables. *Journal of Memory and Language* 25:476–491.

Treiman, Rebecca, Carol A. Fowler, Jennifer Gross, Denise Berch, and Sarah Weatherston. 1995. Syllable structure or word structure? Evidence for onset and rime units with disyllabic and trisyllabic stimuli. *Journal of Memory and Language* 34:132–155.

Treiman, Rebecca, Brett Kessler, Stephanie Knewasser, Ruth Tincoff, and Margo Bowman. 2000. English speakers' sensitivity to phonotactic patterns. In *Papers in Laboratory Phonology V: Acquisition and the lexicon*, ed. Michael Broe and Janet Pierrehumbert, 269–282. Cambridge: Cambridge University Press.

Vaden, Kenneth, Harry R. Halpin, and Gregory S. Hickok. 2009. Irvine phonotactic online dictionary. URL `http://www.iphod.com`.

Vitevitch, Michael S., and Paul A. Luce. 2004. A web-based interface to calculate phonotactic probability for words and nonwords in english. *Behavior Research Methods, Instruments, and Computers* 36:481–487.

Vogt, Hans. 1954. Phoneme classes and phoneme classification. *Word* 10:28–34.

Wolf, Matthew, and John J. McCarthy. 2009. Less than zero: Correspondence and the null output. In *Modeling ungrammaticality in Optimality Theory*, ed. Curt Rice and Sylvia Blaho, 17–66. London: Equinox.