Categorical and gradient aspects of wordlikeness judgements (workshop)

Background Halle (1962) illustrates speakers' ability to distinguish between possible and impossible words in their language with the nonce words *blick* and *bnick*. While *blick* is unattested, it is phonologically well-formed and a "possible" word of English, but speakers immediately recognize *bnick* to be ill-formed and "impossible". A simple explanation for this contrast is that /bn/ is not a possible onset in English and that *bnick* is ill-formed because it cannot be syllabified (e.g., Kiparsky 1982). However, others have argued that this implies a firm boundary between the well-formed and impossible which does not match the granularity in speakers' responses in wordlikeness tasks (e.g., Albright 2009, Chomsky and Halle 1968, Hayes and Wilson 2008, Shademan 2006).

Linking hypothesis However, there is some evidence against a naïve hypothesis connecting speakers' competence to their use of intermediate ratings in wordlikeness tasks (which permit intermediate ratings). For this hypothesis to be falsifiable, it must be the case that ratings of so-called "definitional" concepts do not produce intermediate values. However, Armstrong et al. find that subjects use of a wide range of values when asked to rate the extent to which certain odd counting numbers represent the concept "odd number" on a 7-point scale, despite the fact that "no odd number seems odder than any other odd number." (Armstrong et al. 1983:274). Thus the mere use of intermediate ratings is not compelling evidence that the internal system of wordlikeness judgements produces gradient responses. This has led some (e.g., Schütze 2011) to suggest that intermediate ratings are primarily artifacts of the wordlikeness task itself. We approach this question through an evaluation of computational models of wordlikeness. This is the first study to evaluate categorical and gradient models of wordlikeness (GMWs) on an equal footing.

Data sources Data is taken from two studies which have been widely used to evaluate wordlikeness models. Scholes (1966) presents 63 monosyllabic nonce words to 33 native-speaker subjects, who give each item a "yes"/"no" rating to the question of whether the item "is likely to be usable as a word of English". The second source is a norming study by Albright and Hayes (2003, henceforth A&H) in which 87 monosyllabic nonce words are presented to 20 native speakers, who rate each item on a 7-point Likert scale according to how "natural, or English-like" they sound. Following Hayes and Wilson (2008) and Albright (2009), responses are averaged by item before analysis.

Models GMWs are represented by the MAXENT system of Hayes and Wilson (2008), a segmental BIGRAM model (Albright 2009), and a neighborhood DENSITY model (Bailey and Hahn 2001). We also consider a number of variants of these three models, and for each, we report the most performant variant. These are compared against a primitive BASELINE. This baseline decomposes nonce words into onsets and rimes—as these are known to be important domains for phonotactic patterns (e.g., Treiman 1986)—and distinguishes only between those words which consists of onsets and rimes which are both attested in a database of English monomorphs, and those which do not.

Evaluation Wordlikeness ratings and model scores are compared using non-parametric rank correlation coefficients. Unlike linear models and the familiar Pearson correlation, rank correlation statistics make only one assumption about the relationship between wordlikeness model scores and speakers' judgements, namely monotonicity. As shown in Table 1, there is no consistent advantage of GMWs over the baseline. We also compute the residual contribution of GMWs (Table 2). The results of this latter test indicates that many (and in some cases, most) of the distinctions that GMWs draw within the sets of well-formed and ill-formed clusters (as identified by the baseline) are not reflected in speaker' acceptability ratings.

Conclusions A simple categorical baseline closely models wordlikeness judgements, but beyond this, state-of-the-art GMWs fail to produce consistent improvements, suggesting that the strong performance of current GMWs derives primarily from their ability to mimic categorical distinctions (and the categorical baseline). This result provides support for recent findings that speakers asked to perform gradient syntactic judgements produce responses closely corresponding to a widely recognized categorical grammatical/ungrammatical distinction (Sprouse 2007).

	Spearman ρ				Kendall τ_b			
	BASELINE	MAXENT	BIGRAM	DENSITY	BASELINE	MAXENT	BIGRAM	DENSITY
Scholes	0.791	0.762	**0.827**	0.827	**0.664**	0.597	0.652	0.565
A&H	0.725	0.429	0.708	**0.742**	**0.599**	0.343	0.506	0.556

Table 1: Rank correlations between human wordlikeness judgements and computational model scores reveal no consistent advantage for gradient computational models, and are in some cases outperformed by a primitive binary baseline. All correlations are significant at $p = 0.05$.

	Spearman ρ			Kendall τ_b		
	Δ MAXENT	Δ BIGRAM	Δ DENSITY	Δ MAXENT	Δ BIGRAM	Δ DENSITY
Scholes	-0.029	**0.047**	-0.035	-0.067	**0.003**	-0.061
A&H	-0.008	-0.015	**0.018**	$-$**0.038**	-0.092	-0.049

Table 2: To estimate the residual contribution of the three GMWs beyond the baseline, stimuli are ranked first according to the baseline contrast, and then according to the GMW scores, then the baseline correlation coefficient is subtracted out to produce a difference score. The residual contribution of GMWs is large in no case, and it is negative in many cases.

References

Albright, A. 2009. Feature-based generalisation as a source of gradient acceptability. *Phonology* 26:9–41. Albright, A., and B. Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90:119–161. Armstrong, S., L. Gleitman, and H. Gleitman. 1983. What some concepts might not be. *Cognition* 13:263–308. Bailey, T., and U. Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods. *JML* 44:586–591. Chomsky, N., and M. Halle. 1968. *The sound pattern of English*. Cambridge: MIT Press. Halle, M. 1962. Phonology in generative grammar. *Word* 18:54–72. Hayes, B., and C. Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *LI* 39:379–440. Kiparsky, P. 1982. Lexical morphology and phonology. In *Linguistics in the Morning Calm: Selected Papers from SICOL 1981*, 3–91. Seoul: Hanshin. Scholes, R. 1966. *Phonotactic grammaticality*. Berlin: Mouton. Schütze, C. 2011. Linguistic evidence and grammatical theory. *WIREs Cognitive Science* 2:206–221. Shademan, S. 2006. Is phonotactic knowledge grammatical knowledge? *WCCFL* 25:371–379. Sprouse, J. 2007. Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics* 1:118–129. Treiman, R. 1986. The division between onsets and rimes in English syllables. *JML* 25:476–491.