The Role of Bigram and Unigram Frequencies in Phonotactic Acceptability Judgments

Jane Chandlee

7 March 2012

DGfS

Department of Linguistics and
Cognitive Science

Phonotactic Acceptability

- Categorical:
  - blick
  - *bnick
- Gradient (Halle 1962, Chomsky and Halle 1965):
  - blick > bnick > bzick

Phonotactic Acceptability

- Hypothesis: words that are statistically more probable will be rated higher than words that are less probable.

Findings of the current study

The most obvious means of determining word likelihood actually do not matter when we,

1. control for neighborhood density

2. assess probability based on a model that does not incorporate phonology-specific concepts into the grammar

Modeling gradient judgments

1. Model selection: evaluation metric projected from the lexicon, based on

   - syllable constituents (Coleman 1996, Coleman & Pierrehumbert 1997, Frisch et al. 2000, Treiman et al. 2000, Shademan 2007)
   - feature-based bigrams (Albright 2009, Heinz & Koirala 2010)
   - bigrams and unigrams combined (Vitevitch et al. 1997)
   - whole word (= neighborhood density) (Bailey and Hahn 2001)

Modeling gradient judgments

2. Parameter tuning:

   - optimization of weighted constraints (Hammond 2004, Antilla 2008, Coetzee 2008, Hayes and Wilson 2008)
   - maximum likelihood estimation of n-gram models

Modeling gradient judgments

- These models often have built-in phonological components:
  - Syllable structure (Coleman 1996, Coleman & Pierrehumbert 1997, Frisch et al. 2000, Treiman et al. 2000, Shademan 2007)
  - Features (Albright 2009, Heinz & Koirala 2010)
- Experimental stimuli often have gradient constraint violations:
  - bnick < bzick

Objectives of the current study

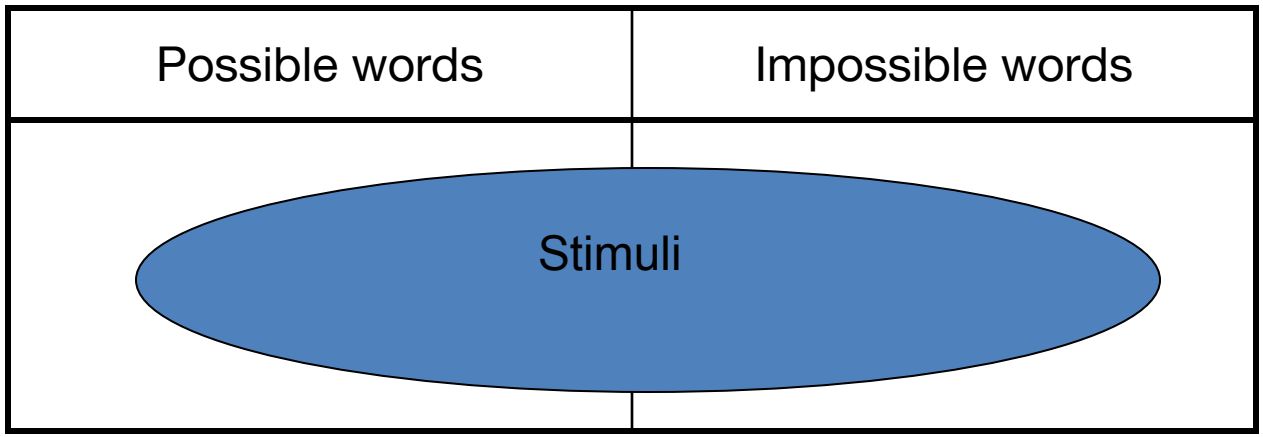
1. Isolate the role of frequency with a phonotactic acceptability experiment using fully legal CVC nonce words that violate no known phonotactic constraints.

    - no clusters

    - no words beginning with /ŋ/, /ʒ/, /ð/, ending with /h/, etc.

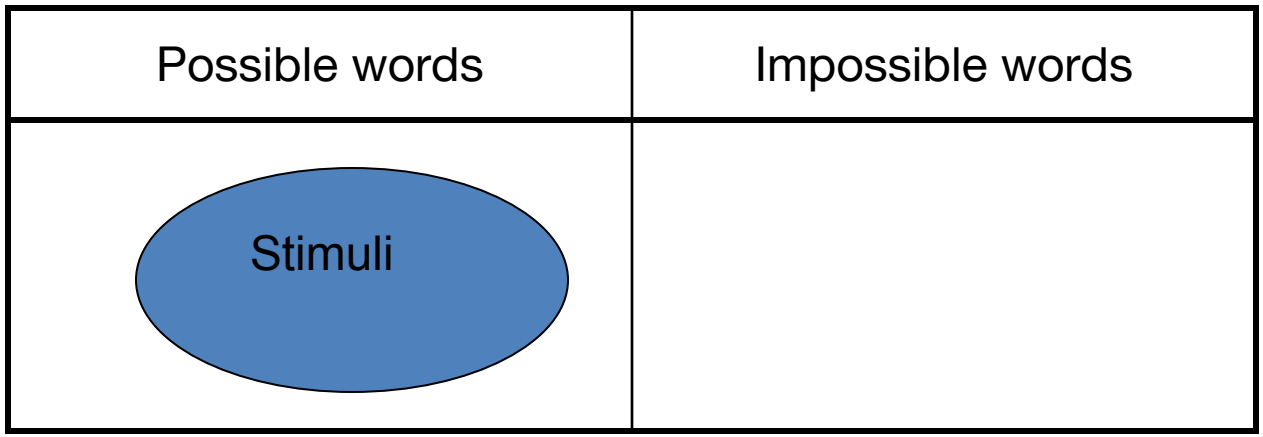
Objectives of the current study

- Previous studies



Objectives of the current study

- Current study



Objectives of the current study

2. Test whether speakers are sensitive to bigram and/or unigram frequency alone by controlling for neighborhood density.

Bigrams vs. unigrams

- Bigram model: distribution of phonemes in terms of previous sound

Bigrams of ‘blick’ [blɪk]:

#b, bl, lɪ, ɪk, k#

Bigram probability =

P(b|#) x P(l|b) x P(ɪ|l) x P(k|ɪ) x P(#|k)

Bigrams vs. unigrams

- Unigram model: distribution of individual phonemes

Unigrams of ‘blick’:

b, l, ɪ, k

Unigram probability =

P(b) x P(l) x P(ɪ) x P(k)

Training Data

- CELEX English lemma corpus with transcriptions from the CMU Pronouncing Dictionary
- Stress information removed
- Bigram and unigram models trained on resulting corpus of 23,911 words

Experiment

- Forced choice between pairs of nonce words systematically selected to test the individual roles of bigram and unigram frequency.

Stimuli

- All possible CVC words were generated with the English phoneme inventory.
- Real words were removed.
- Bigram and unigram probabilities were determined for each word; zero probability words were removed.
- Each word was given a neighborhood density score (number of real words with a string edit distance of 1) (Jurafsky and Martin 2000)

Stimuli

- Pairs of words selected in five categories - in all pairs, the two words were matched for neighborhood density.

Type 1

- The two words are equivalent* in both bigram and unigram probability.

  – [tʃɔɪb] vs. [θaʊg]

  BG: 1.15E-08 1.16E-08

  UG: 1.77E-07 1.76E-07

- Prediction: subjects will perform at chance on pairs of this type.

Type 2

- The two words are equivalent in bigram probability, but one word has a higher unigram probability than the other.

	[nʌs]	vs.	[bʌʃ]
BG:	2.82E-05		2.82E-05
UG:	0.0004		3.39E-05

- Prediction: subjects will prefer the word with higher unigram probability.

Type 3

- The two words are equivalent in unigram probability, but one word has a higher bigram probability than the other.

–	[sejʃ]	vs.	[nɪθ]
BG:	3.95E-05		4.39E-07
UG:	1.70E-05		1.70E-05

- Prediction: subjects will prefer the word with higher bigram probability.

Type 4

- One word has a *higher* unigram probability but a *lower* bigram probability than the other.

  – [zajp] vs. [dʃejʃ]

  BG: 2.75E-08 4.66E-06

  UG: 5.13E-06 2.37E-06

- Prediction: ??

Type 5

- One word has a both a higher unigram and a higher bigram probability than the other.

	[pʌd]	vs.	[hijð]
BG:	4.41E-05		2.92E-07
UG:	.0001		3.34E-07

- Prediction: subjects will prefer the word with a higher probability in both models.

Stimuli

- ‘Equivalent probability’ defined as the probability of one word fell into a range of plus/minus 2% of the probability of the other.

- When the words differed in probability, the difference was maximized: one word’s probability fell outside of a range of plus/minus 99.3% of the probability of the other.

Stimuli

- 100 pairs = 200 words total, recorded by a female native speaker of American English

Subjects

- 37 University of Delaware undergraduates, recruited from linguistics courses
- 34 female, 36 right-handed
- ages 18-21
- monolingual speakers of American English

Task

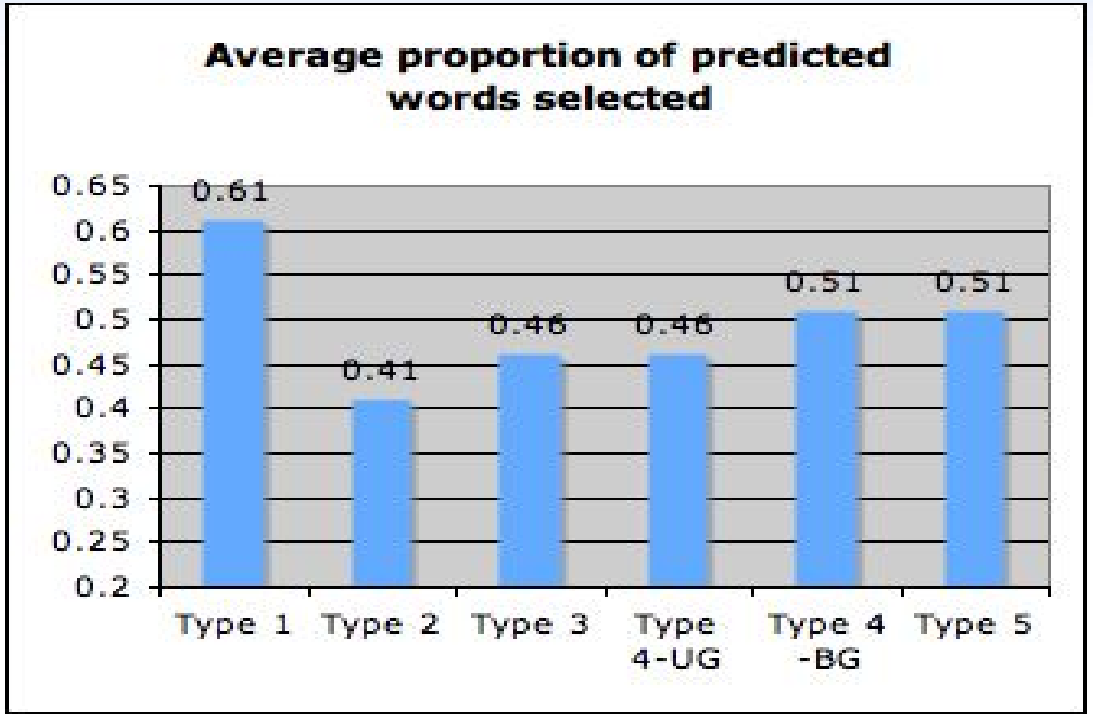
- 100 word pairs were presented aurally in a randomized order.
- Within the pairs, order of presentation for the predicted/not predicted word was counterbalanced.
- “Choose the word you believe is more likely to enter the English language.”

Analysis

- For each stimuli pair type (1-5), calculated the average proportion of trials (out of 20) for which the subjects selected the word predicted by the relevant model.

- Linear mixed-effects models (subject and trial as random effects) were fitted to the data in order to compare the average to chance.

Results


Average proportion of predicted words selected
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.61
0.41
0.46
0.46
0.51
0.51
Type 1
Type 2
Type 3
Type 4-UG
Type 4 -BG
Type 5

Results

p values for comparison to chance (α = .05, n = 37)

Pair type	p
1	0.02*
2	0.04*
3	0.5
4-UG	0.5
4-BG	0.9
5	0.7

Interim summary

- For the pairs in which the models made no prediction, subjects demonstrated a bias for the ‘first word heard’.
- Subjects *disfavored* the words selected by only the unigram model.
- In all other cases, subjects performed at chance.

Model comparison

- Do the previous models correlate better with these results?

1. Vitevitch & Luce (2004) unigram model
2. Vitevitch & Luce (2004) bigram model
3. Hayes & Wilson (2008) Maximum Entropy model
4. Heinz & Koirala (2010) featural bigram model

Vitevitch & Luce (2004)

- *Positional* unigram probability:
  - CVC
    ⇑
    P(C) = probability of C appearing in onset position.
- Log frequencies of words with that segment in that position summed and divided by summed log frequencies of all words in the corpus (Merriam-Webster Dictionary)

Vitevitch & Luce (2004)

- Positional bigram probability
  - CVC
- Log frequencies of words with those two segments in adjacent positions summed and divided by summed log frequencies of all words in the corpus.

Hayes & Wilson (2008)

- Grammar of phonotactic constraints derived from the training data.
- Nonce word probability calculated based on a weighted sum of its constraint violations.

Heinz & Koirala (2010)

- Distribution of individual feature values (as opposed to phonemes) given the previous value determined from the training data.
- Nonce word acceptability assessed by combining individual feature models into a single model.

Model comparison

- For each stimuli pair, determined the difference in the proportions in which each word was selected:

Word 1	% chosen	Word 2	% chosen	Diff.
θuwb	68	θowf	32	36
nɔs	27	huwb	73	-46

Model comparison

- Likewise, the difference between the scores assigned to the words by each model was calculated:

Model	Word 1	Score	Word 2	Score	Diff.
VLUG	nɔs	.1192	huwb	.0875	.0317
VLBG	nɔs	.0013	huwb	.0022	-.0009
etc.					

Model comparison

- Spearman rank correlations were run to determine which model(s) best predicted the experimental results.

Model comparison

Model	ρ	p
Unigram	0.14	0.1
Bigram	0.07	0.4
VL-UG	0.06	0.5
VL-BG	0.02	0.9
Hayes&Wilson08	n/a	
Heinz&Koirala10	0.19	0.04*

Implications

- If frequency alone can account for gradience (Bybee 1995), then we would expect the subjects to meet the predictions of the bigram and unigram models.

Conclusion and future work

- The current findings suggest that frequency alone is not sufficient to model phonotactic acceptability.
- Further investigation is needed to determine whether this null result is a reflection of the experimental task.

Conclusion and future work

- Additional models can also be compared to better understand what additional information (neighborhood density, syllable structure, featural representations, etc.) should be incorporated.

References

Albright, A. (2009). Feature-based generalization as a source of gradient acceptability. *Phonology 26*, 9-41.

Antilla, A. (2008). Gradient phonotactics and the complexity hypothesis. *Natural Language and Linguistic Theory 26*, 695-729.

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language, 44*(4), 568-591.

Bybee, J. (1995). Regular Morphology and the Lexicon. *Language and Cognitive Processes 10,* 425-455.

Chomsky, N., & Halle, M. (1965). Some controversial questions in phonological theory. *Journal of Linguistics, 1*(2), pp. 97-138.

Coetzee, A.W. (2008). Grammaticality and ungrammaticality in phonology. *Language 84*(2). 218-57.

Coleman, J., & Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. *Computational Phonology: 3rd Meeting of the ACL Special Interest Group in Computational Phonology,* 49-56.

Coleman, J. S. (1996). The psychological reality of language-specific constraints. *Paper Presented at the Fourth Phonology Meeting, 16-18 1996,* University of Manchester.

Frisch, S. A., Large, N. R., & Pisoni, D. B. (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords*, Journal of Memory and Language, 4*2(4), 481-496.

References

Halle, M. (1962). Phonology in generative grammar. *Word, 18*. 54-72.

Hammond, M. (2004). Gradience, phonotactics, and the lexicon in English phonology. *International Journal of English Studies 4*(2), 1-24.

Hayes, B. and Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry 39*, 379-440.

Heinz, J. and Koirala, C. (2010). Maximum likelihood estimation of feature-based distribution. *Proceedings of the 11th Meeting of the ACL-SIGMORPHON, ACL 2010*, 28-37.

Jurafsky, D. and Martin, J.H. (2000). *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Upper Saddle River, NJ: Prentice Hall.

Shademan, S. (2007). *Grammar and analogy in phonotactic well-formedness judgments.* (Unpublished PhD). UCLA.

Treiman, R., Kessler, B., Knewasser, S., Tincoff, R., & Bowman, M. (2000). English speakers' sensitivity to phonotactic patterns. In M. B. Broe, & J. Pierrehumbert (Eds.), *Papers in laboratory phonology V: Acquisition and the lexico*n (pp. 269-282). Cambridge, England: Cambridge University Press.

Vitevitch, M.S., Luce, P.A., Charles-Luce, J. and Kemmerer, D. (1997). Phonotactics and syllable stress: implications for the processing of spoken nonsense words. *Language and Speech 40*(1), 47-62.

Vitevitch, M.S. and Luce, P.A. (2004) A web-based interface to calculate phonotactic probability for words and nonwords in English. Behavior Research Methods, Instruments, and Computers, 36, 481-487.

Acknowledgments

- Thanks to Jeffrey Heinz, Irene Vogel, Regine Lai, Arild Hestvik, Anne Peng, members of the University of Delaware Phonetics and Phonology Lab, the UD Department of Linguistics and Cognitive Science, and the Office of Graduate and Professional Education.