I. Phonotactics in speech production

(1) Phonotactics—patterns of sound sequences that characterize a language.
   A. Phonotactic legality—what are the possible forms of words?
      • e.g., No English word begins with /dl/.
   B. Phonotactic probability—more vs. less likely sequences.
      • e.g., English words are more likely to begin with /k/ as compared to /g/.
   C. How are phonotactics reflected by processing mechanisms?

(2) Evidence for processing mechanisms encoding phonotactic legality
   A. Spontaneous speech errors rarely result in phonotactically illegal forms.
      i. e.g., English speakers rarely produce forms beginning with /dl/.
         • “dorm friend” → “drorm fend”; much more likely than
         • “dorm floor” → “*dlorm four”
      iii. Several studies have noted significant numbers of phonotactically illegal errors, (Butterworth & Whittaker, 1980; Laver, 1980), but such errors are vastly outnumbered by legal ones (but see Goldstein et al., in prep.; Mowrey & MacKay, 1990).

B. Suggests the presence of phonotactic processing constraints (PPCs) that encode phonotactic legality.
   • PPCs are biased against phonotactically illegal forms; block such forms from being produced in errors.

II. PPCs and phonotactically legal forms

(3) How do PPCs distinguish among phonotactically legal forms?

(4) Are PPCs sensitive to markedness or frequency?
   A. Phonotactically legal forms (e.g., English word-initial /k/, /g/) differ in terms of:
      i. Frequency: Word-initially, /k/ is more frequent than /g/.
         Token frequency in CELEX (Baayen et al. 1995): /k/ .09 /g/ .03
      ii. Cross-linguistic markedness: If a language has /k/ in its inventory, it will also have /g/.
         True of 86% of languages in UPSID (2000)
B. Some results suggest PPCs are sensitive to frequency. Errors are more likely to result in high frequency structures than low frequency structures (e.g., Levitt & Healy, 1985; Motley & Baars, 1975).

- /g/ \rightarrow /k/ is more likely than /k/ \rightarrow /g/

C. Other results suggest PPCs are sensitive to markedness. Compared to marked structure, unmarked structures are more likely to be produced as an error outcome (e.g., Béland & Paradis, 1997; Romani & Calabrese, 1998).

D. However, markedness and frequency are correlated; these studies fail to de-confound two variables.

- Results attributed to frequency may be due to markedness difference (or vice versa).
- Motley & Baars (1975) attempt to contrast effects of two variables, finding only an effect of frequency. However, fails to match different groups for other variables (e.g., similarity of consonants).

(5) Do PPCs prefer to eliminate or insert marked/infrequent material?
A. Studies above—marked/infrequent material is eliminated.

- /g/ \rightarrow /k/ is more likely; Marked/infrequent /g/ is eliminated, replaced by unmarked/frequent /k/.

B. Stemberger (1991a,b, 1992); Stemberger & Treiman (1986): Marked/infrequent material is inserted.

- /k/ \rightarrow /g/ is more likely; Marked/infrequent /g/ is inserted, replacing unmarked/frequent /k/.
- Goldstein et al. (in prep.)—insertion bias in errors; gestures are more likely to be inserted than deleted.

C. However, these studies fail to control for properties of stimuli other than markedness/frequency—properties that may influence speech errors.

- e.g., Neighborhood density (the number of lexical items similar to a target; shown by Vitevitch, 2002, to influence speech errors)

- Conflicting results may be due to failure to control for this factor.

(6) Current study attempts to resolve these questions by:

- De-confounding markedness and frequency.
- Equating structures contrasting in frequency/markedness in terms of other factors.

III. Experimental investigation of PPCs: Method

(7) Are PPCs sensitive to markedness or frequency?
A. Previous studies suggest likelihood of an error outcome is related to markedness/frequency.

- Positively (/g/ \rightarrow /k/ is more likely) or negatively (/k/ \rightarrow /g/ is more likely)
B. Examine bias in error outcomes on structures where frequency and markedness are de-confounded.
   • Which theory better predicts pattern of bias?

C. Specific theoretical positions
i. Frequency
   a. **Instance-Based** theory. PPCs reflect token frequency of segments in particular linear positions. (based on Luce et al., 2000)
      • e.g., Token frequency of /k/ word-initial position
   b. **Lexical Distribution** theory. PPCs reflect type frequency of syllable constituents in particular contexts/positions (based on Coleman & Pierrehumbert, 1997; Frisch et al., 2000).
      • e.g., Type frequency of onset /k/ in stressed, word-initial syllables.

ii. **Markedness** theory
   • PPCs reflect cross-linguistic markedness of phonological structures. (based on generative markedness theories, e.g., Prince & Smolensky, 1993)

D. **Test** pairs: First member of each pair is unmarked and less frequent than second

<table>
<thead>
<tr>
<th>Consonant pair with markedness contrast</th>
<th>Relative token frequency: Instance-Based theory</th>
<th>Relative type frequency: Lexical Distribution theory</th>
</tr>
</thead>
<tbody>
<tr>
<td>t – d</td>
<td>4.4% - 5.7%</td>
<td>4.0% - 5.2%</td>
</tr>
<tr>
<td>t – s</td>
<td>4.4% - 11.2%</td>
<td>4.0% - 6.9%</td>
</tr>
<tr>
<td>p – k</td>
<td>7.6% - 8.9%</td>
<td>6.2% - 6.9%</td>
</tr>
</tbody>
</table>

⇒ First member of each pair is well-formed with respect to markedness, ill-formed with respect to frequency.
(see Appendix A1 for frequency calculations and markedness references)

(8) Do PPCs prefer to eliminate or insert marked/infrequent material?
A. Examine bias on **control pairs**: matched consonant pairs where all theories agree on well-formedness.

B. Direction of bias will show whether marked/infrequent structure is eliminated or inserted.

C. **Control** pairs: First member of each pair is unmarked and more frequent than second.

<table>
<thead>
<tr>
<th>Consonant pair with markedness contrast</th>
<th>Relative token frequency: Instance-Based theory</th>
<th>Relative type frequency: Lexical Distribution theory</th>
</tr>
</thead>
<tbody>
<tr>
<td>k – g</td>
<td>8.9% - 2.6%</td>
<td>6.9% - 2.1%</td>
</tr>
<tr>
<td>d – z</td>
<td>5.7% - 0.1%</td>
<td>5.2% - 0.2%</td>
</tr>
<tr>
<td>b – g</td>
<td>5.3% - 2.6%</td>
<td>5.1-2.1%</td>
</tr>
</tbody>
</table>

⇒ First member of each pair is well-formed with respect to markedness and frequency.
Data collection.
A. Participants: 45 Johns Hopkins University undergraduates, native English speakers. Compensated with either $7 or extra-credit in introductory courses.
B. Materials
i. Errors generated using tongue twisters: Each control and test pair embedded in an alliterating sequence of four CVC nonwords.
   • kev geeb keeb gev
ii. Nonwords within a pair were matched in terms of lexical neighborhood density and transitional probability of initial consonants (see Appendix A2).
iii. Participants read sequences aloud at rate of 2.5 syllables/second to induce speech errors.
C. Responses were tape recorded and transcribed for analysis.

IV. Experimental investigation of PPCs: Results
A. For two control pairs, significantly more errors on marked/infrequent consonants.
   • /g/ → /k/ is more likely than /k/ → /g/

(See Appendix A3 for details of analysis)
B. Suggests that PPCs prefer to eliminate (rather than insert) marked/infrequent material.

C. Null results on /b/-/g/ could be due to lack of a significant difference in well-formedness.
   i. Markedness relationship between labials and velars is controversial.
      • Typologically, /b/ appears less marked than /g/ (Gamkrelidze, 1978);
        however, Rice (1996) claims that velars can be unmarked (i.e., default place).
   ii. Frequency difference is quite small

<table>
<thead>
<tr>
<th>Consonant pair</th>
<th>Token frequency difference</th>
<th>Type frequency difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>k – g</td>
<td>6.3%</td>
<td>4.8%</td>
</tr>
<tr>
<td>d – z</td>
<td>5.6%</td>
<td>5.0%</td>
</tr>
<tr>
<td>b – g</td>
<td><strong>2.7%</strong></td>
<td><strong>3.0%</strong></td>
</tr>
</tbody>
</table>

(11) Results: Test pairs
A. For one test pair, significantly more errors on marked/frequent structure.
   • /s/ \(\rightarrow\) /t/ is more likely than /t/ \(\rightarrow\) /s/
B. Suggests that PPCs prefer to eliminate unmarked (as opposed to infrequent) material.

C. Null results
i. /p/-/k/: labial-velar relationship unclear.
   • Gamkrelidze (1978; see also Ohala, 1983) suggests that /p/ is in fact more marked than /k/.
ii. /t/-/d/: unclear.

V. Implications
(12) PPCs prefer to eliminate marked structure in favor of unmarked structure.
A. For two control pairs and one test pair, errors are more likely to result in unmarked consonants than marked consonants.

B. How can we account for conflicting results in the literature?
   i. Lexical neighborhood effects?
      a. This study: Tendency to eliminate marked structure is not due to neighborhood effects.
         • Marked-unmarked matched in terms of neighborhood density.
      b. Stemberger (2002): Tendency to insert marked structure not due to density.
         • Results of Stemberger (1991a) and Stemberger & Treiman (1986).

   ii. Another possible variable: type of stimuli used as targets.
      a. All studies showing tendency to insert marked structure have used real words as targets.
      b. Studies using nonwords have found tendency to eliminate marked structure.
         • Frisch (1996): Found tendency to insert marked structure with word targets, no such bias with very similar nonword targets.

   iii. Different types of stimuli may tap into different levels of processing.
      a. Word targets tap into “lexical retrieval processes.” Here, errors arise due to competition between underspecified lexical representations. Marked structure overwhelms underspecified unmarked structure (Stemberger, 1991a).
      b. Nonword targets, lacking lexical representations, tap into post-retrieval phonological processes. PPCs—part of these post-retrieval processes—bias errors towards unmarked structure.

   iv. Future work: directly contrasting word/nonword targets matched for lexical neighborhood properties.
PPCs prefer to eliminate unmarked (as opposed to infrequent) material.
A. For test pair /t/-/s/, errors are more likely to result in unmarked consonant than frequent consonant.

B. Alternative frequency account?
   i. Instance-Based and Lexical Distribution theories ((7)C) assume that segments are most basic level of representation.
   
   ii. Alternative versions of each theory: Frequency counted over subsegmental units.

<table>
<thead>
<tr>
<th>Pairs</th>
<th>Sub-segmental contrast</th>
<th>Sum frequency of segment class: Instance-Based</th>
<th>Sum frequency of segment class: Lexical Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/d/-/z/;</td>
<td>/k, g, p, b, t, d/</td>
<td>stop: 29.2%</td>
<td>stop: 29.5%</td>
</tr>
<tr>
<td>/t/-/s/</td>
<td>Fricative</td>
<td>fricative: 25.7%</td>
<td>fricative: 15.1%</td>
</tr>
<tr>
<td>/θ, θ, z, s, f, v, ŋ, ʒ/</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Segment class frequency makes same prediction as markedness for other pairs as well)

iii. Frequency account: /t/ is preferred to /s/ because it belongs to a more frequent segment class.

iv. Markedness or frequency can account for these data so long as PPCs encode sub-segmental information.
   • Goldrick (2002a,b): Learning of PPCs influenced by sub-segmental similarity.

See Appendix A4 for implications for formal theories of phonotactic constraints

VI. Conclusions

Phonotactic patterns are reflected by processing mechanisms—PPCs—which are biased against marked/sub-segmentally infrequent structures.

Future work
A. Testing contrasting predictions of sub-segmental markedness, token frequency and type frequency theories.
   • Sub-segmental Instance-Based theory: Dorsal stops are more frequent than coronals.
   • Sub-segmental Lexical Distribution, Markedness theories: Dorsals stops are less frequent/more marked than coronal stops.

B. Characterization of the representations over which PPCs operate.
   • Do PPCs bias sub-categorical/ “phonetic” errors?


(18) Frequency: Based on CELEX lexical database (Baayen et al., 1995)
A. Frequency based on “words.”
   i. Exclude entries with spaces and those consisting of all capital letters.
   ii. Type counts: Collapsed across all word forms with same spelling and pronunciation.
   iii. Token counts: Minimum log token frequency = 1.
B. Instance-Based theory: Following Luce et al. (2000), calculated relative log (base 10) token frequency of segments in absolute word initial position.
C. Lexical Distribution theory: Following Frisch et al. (2000), calculated relative type frequency of onsets of stressed initial syllables (excluding “null” onsets from counts).
D. Coleman & Pierrehumbert (1997) differentiate among classes of initial syllables (initial non-final vs. initial and final). Could this alternative account for the results?
   i. No— monosyllabic statistics make the wrong prediction for the pair /t/-/s/.
      • Relative frequency of /t/ in monosyllables: 3.8%; /s/: 4.3%.
   ii. However, segment-class statistics for monosyllabic words do make the correct prediction.
      • Relative frequency of stops, 23.9%; fricatives, 12.1%.

A2. Appendix: Control for transitional probabilities and lexical neighborhood density

(19) Calculation of probabilities and densities.
A. Forward and backward transitional probabilities calculated using token frequencies in CELEX (following Luce et al., 2000).
B. Neighborhood density measures: monosyllabic CVC words in CELEX sharing either:
   i. The initial consonant-vowel sequence of the nonword target.
   ii. The initial and final consonants of the nonword target.
C. Mean values for control pairs (note: *p < .05, paired t-test)

<table>
<thead>
<tr>
<th>Consonant type</th>
<th>Forward transitional probability</th>
<th>Backward transitional probability</th>
<th>Number of words sharing both initial consonant and vowel*</th>
<th>Number of words sharing both initial and final consonants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unmarked/Frequent</td>
<td>2.9%</td>
<td>3.8%</td>
<td>8.5</td>
<td>2.7</td>
</tr>
<tr>
<td>Marked/Infrequent</td>
<td>2.9%</td>
<td>1.6%</td>
<td>3.8</td>
<td>1.8</td>
</tr>
</tbody>
</table>
D. Mean values for test pairs

<table>
<thead>
<tr>
<th>Consonant type</th>
<th>Forward transitional probability*</th>
<th>Backward transitional probability</th>
<th>Number of words sharing both initial consonant and vowel</th>
<th>Number of words sharing both initial and final consonants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unmarked/Infrequent</td>
<td>3.0%</td>
<td>6.1%</td>
<td>7.3</td>
<td>2.3</td>
</tr>
<tr>
<td>Marked/Frequent</td>
<td>2.2%</td>
<td>5.6%</td>
<td>6</td>
<td>2.8</td>
</tr>
</tbody>
</table>

E. Frequency and transitional probability within experiment controlled by introducing two non-critical pairs (/p/-/b/ and /s/-/z/).
- Other consonants (/k/, /g/, /d/, /t/) occur in two control/test pairs; introduction of one extra pair for each balances frequencies.

A3. Appendix: Raw numbers and analysis-by-participants

(20) Notes on analysis
A. Statistical determination of bias: $\chi^2$ test comparing error rate on two consonants of each pair.
B. “Cutoff” errors such as “kev”→“g…kev” excluded.
C. Only contextual errors analyzed; e.g., k→g errors only in the context of “g” targets.
- Note: same pattern found when these other errors are included.

(21) Raw numbers of consonantal errors (out of 2160 targets)

<table>
<thead>
<tr>
<th>Pair</th>
<th>Unmarked/Frequent $\rightarrow$ Marked/Infrequent</th>
<th>Marked/Infrequent $\rightarrow$ Unmarked/Frequent</th>
<th>$\chi^2$ (N = 4320)</th>
</tr>
</thead>
<tbody>
<tr>
<td>/k/-/g/</td>
<td>221</td>
<td>302</td>
<td>13.9 *</td>
</tr>
<tr>
<td>/d/-/z/</td>
<td>145</td>
<td>287</td>
<td>51.1 *</td>
</tr>
<tr>
<td>/b/-/g/</td>
<td>147</td>
<td>123</td>
<td>2.1</td>
</tr>
</tbody>
</table>

(22) Analysis-by-participants
A. Statistical test: Two-tailed paired t-test using number of errors produced by each participant.
- No continuity correction due to small number of errors per participant.
A4. Appendix: Implications for formal theories of phonotactic constraints

Phonotactic probabilities (patterns among phonotactically legal forms) are encoded by the language processing system.

• Such patterns are not expressed by theories making use of only inviolable constraints (e.g., Calabrese, 1995; Paradis, 1988).

Phonotactic constraints have a probabilistic effects.

A. Easily interpreted within theories invoking stochastic constraint evaluation mechanisms.
   i. Stochastic constraint ranking (e.g., Antilla, 1997; Boersma, 1998).
   ii. Stochastic constraints (e.g., Frisch, 1996, 2000; Frisch et al., 1997).
   iii. Stochastic context-free grammars (e.g., Coleman & Pierrehumbert, 1997).

B. Theories without such a mechanism (e.g., Prince & Smolensky, 1993/2002) must attribute effects to non-grammatical “performance” factors.

Segmental phonotactic constraints are insufficient; sub-segmental information must be included
• Theories which exclude such representations (e.g., Coleman & Pierrehumbert, 1997) cannot express these constraints.
References


University of California, Los Angeles Phonetics Laboratory (2000). UCLA Phonetic Segment Inventory Database (UPSID).


Contact

Matt Goldrick
Department of Cognitive and Linguistic Sciences
Box 1978, Brown University
Providence, RI 02912 USA

Matt_Goldrick@brown.edu
http://cog.brown.edu/~goldrick