Efficient Communication II
Communicatively efficient and activation-based models of incremental production

LSA Summer Institute 2011, 
*Computational Psycholinguistics*, Lecture 4

Florian Jaeger, Human Language Processing Lab
http://www.hlp.rochester.edu/
Readings

• **Required:** Jaeger, 2010 (20+ pp); Aylett & Turk 2006 (11pp)

• **Suggested:** Genzel & Charniak, 2002; Levy & Jaeger, 2007; Moscoso del Prado Martin, submitted; Qian and Jaeger, 2010, 2011

• **Technical reading (optional):** Shannon (1948)
Plan

• Uniform information density

• Constant Entropy Rate: Evidence and Critique

• **Today:** Focus on results; next time: interpretation
  – Beginning to link computational level considerations about efficient communication to mechanisms:
Part 1

Beyond the lexicon –
the distribution of information across
the linguistic signal
‘Choice points’

Inter-clausal: Move the triangle to the left.
Select the triangle. Move it to the left.

Phrasal level: She gave {him the key/the key to him}

She already ate (dinner)
She stabbed him (with a knife).
She met him/Tom/my friend yesterday.

Word level: I read a book (that) she wrote.

Morphology: I’ve\have gone there.

Phonology: t/d-deletion; final cluster reduction; vowel weakening

Phonetics: formant energies, F1/F2 ratio, speech rate
Preferences at choice points

If language production system is organized to facilitate efficient and robust communication ...

→ *Given a choice point, speakers prefer to keep the amount of information transmitted per unit signal uniform (= uniform information density, UID).*

Inter-clausal:  Move the triangle to the left.
               Select the triangle. Move it to the left.

Phrasal level:  She gave {him the key/the key to him}

She already ate (dinner)
She met him/Tom/my friend yesterday.
She stabbed him (with a knife).

Word level:  I read a book (that) she wrote.

Morphology:  I’ve\have gone there.

Phonology:  t/d-deletion; final cluster reduction; vowel weakening

Phonetics:  formant energies, F1/F2 ratio, speech rate
Questions

• What is known?

• What are the open question?
  – Empirically
  – Theoretically
Word Duration in Spontaneous Speech

Information(\textit{mind}) = -\log p(\textit{mind}) ≈ -\log p(\textit{mind} | \textit{degrade the})

Information(\textit{mind}) ≈ -\log p(\textit{mind} | \textit{on my})

Paid jobs degrade the \textit{mind} 409 ms

Mama you've been on my \textit{mind} 305 ms
The Smooth Signal Redundancy Hypothesis

A There is an inverse relationship between language redundancy and acoustic redundancy (as manifested by syllable duration).

B Prosodic prominence smooths signal redundancy by controlling syllabic duration.
Some results

[Figure 5 from Aylett and Turk, 2004]
Prosody as a mechanism to modulate word duration and articulatory detail

[Figure 7 from Aylett and Turk, 2004]
Discounting prosodic boundaries

[Figure 7 from Aylett and Turk, 2004]
Phonetic reduction in non-conversational speech

Fig. 4. Syllabic duration means (log ms), by redundancy group and by vowel. All differences in means are significant. As a reference: 2.00 log ms = 100 ms, 2.20 log ms = 158.5 ms, 2.40 log ms = 251.2 ms, 2.60 log ms = 398.1 ms.

Fig. 5. Unique contributions of prosodic prominence and redundancy factors to linear regression models for each vowel.

[Figures 4 and 5 from Aylett and Turk, 2006]
Some open questions

- Why are the effects of predictability on phonetic reduction so variable?
  [cf. Bell et al., 2003, 2009; Table 10 from Bell et al 2009 shown to the right]

<table>
<thead>
<tr>
<th></th>
<th>High-frequency function words</th>
<th>High-frequency function words</th>
<th>Content words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word frequency</td>
<td>NS</td>
<td>NS</td>
<td>Highly significant</td>
</tr>
<tr>
<td>Previous conditional</td>
<td>Highly significant</td>
<td>Marginally significant</td>
<td>NS</td>
</tr>
<tr>
<td>Following conditional</td>
<td>NS</td>
<td>Highly significant</td>
<td>Highly significant</td>
</tr>
<tr>
<td>Repetition</td>
<td>NS</td>
<td>NS</td>
<td>Significant</td>
</tr>
</tbody>
</table>
Beyond duration: Centralization based on syllable predictability

[Figures 6 and 7 from Aylett and Turk, 2006]

FIG. 7. Female spectral results by vowel (n=5). 1=high language redundancy, 2=medium language redundancy, 3=low language redundancy. All differences between 1 & 2, 2 & 3, 1 & 3 are significant (p<0.005) for both F1 and F2 (unless noted) on the basis of a posthoc t-test with Bonferroni correction. Nonsignificant group differences: æ/ F1 1 & 2, /æ/ F2 1 & 2, /ɪ/ F1 2 & 3, /u/ F2 1 & 2.
Beyond duration: Center of gravity based on segment information

[van Son and Pols, 2003; also van Son and van Santen, 2005]

Fig. 4. Variance (first and last column) and reduction in variance of the segmental information $I_s$ (vertical scale, % of total on the right hand scale) due to accounting for the indicated factor and all the factors to the left of it (horizontal scale). The order of the factors is the same as in table 1. The data concern phonemes, excluding phonemes from syllables containing a schwa. N=26,411, maximal number of subsets: 6,428. Not Explained variance: 10.2% of Total variance. All factors p<0.001 (F-test). White columns are plain variances, not differences.

[Figures 4 and 6 from van Son and Pols, 2003]
The information carried by a segment

• Pols et al. (1998):

\[ I_L = -\log_2 \left( \frac{\text{Frequency}([\text{word onset}]+s)}{\text{Frequency}([\text{word onset}]+\text{any segment})} \right) \]

• To account for the predictability of the word due to its distributional (contextual) properties, van Son and Pols (2003) extend this measure (following MacDonald and Shillcock, 2001):

\[ I_s = -\log_2 \left( \frac{\text{Frequency}([\text{word onset}]+s)+D(w)}{\text{Frequency}([\text{word onset}]+\text{any segment})+D(w)} \right) \]

Adjustment for average predictability of word in context (ideally, we’d like to use the actual predictability in context)
[Illustrating variability in phonetic reduction across words for the most frequent function words in Switchboard, Post and Jaeger, 2010]
Some open questions

• How can these effects be related to mechanistic models of language production? How do they relate to questions about audience design?

[cf. Arnold, 2008; Bard et al., 1999; Bard and Aylett, 2005; Gahl et al., submitted; Galati and Brennan, 2010; Gregory, 2000; Jaeger and Buz, in progress; Yao, 2010]
Efficient Morpho-Syntactic Production

[Frank and Jaeger, 2008]

Pres. Clinton did *n’t/not* have ...

Information content of NOT
Estimating the information carried by a contractible element

\[
\text{Information theoretic definition of Shannon information content}
\]

\[
I(\text{NOT} \mid \text{context}) = -\log p(\text{NOT} \mid \text{context}) = -\log p(\text{NOT} \mid "\text{Clinton did }") = -\log \left[ p("\text{not}" \mid "\text{Clinton did }") + p("\text{\text{'}t}" \mid "\text{Clinton did }") \right]
\]

Use trigram model to estimate probability (backoff)
Efficient Morpho-Syntactic Production

[Frank and Jaeger, 2008]

Pres. Clinton did n’t/not have ...
• Extracted from a corpus of spontaneous American English speech (Switchboard, 800k sentences in 650 dialogues)
  – Only cases that are contractible in American English are included (e.g. *not “I have/*’ve a car”*).
  
  – **HAVE:** e.g. ’d vs. had (>2,400 contractible cases)
  – **NOT:** n’t vs. not (> 5,000 contractible cases)
  – **BE:** e.g. ’s vs. is (> 9,000 contractible cases)
Analysis

[Frank and Jaeger, 2008]

• Mixed logit model to analyze when speakers’ choose **full** over **contracted** forms depending on the information carried by it.

\[
\text{logit}[p(\text{full})] = \ln \frac{p(\text{full})}{p(\text{reduced})} = -\beta \log p(NOT | w_{i-1}) - \beta \log p(NOT | w_{i+1}) + X_{\text{Controls}} \beta_{\text{Controls}} + Zb
\]

• Simultaneously controlling for:
  – Position in intonational phrase
  – Complexity of *upcoming* material
  – Complexity of *host* (e.g. pronominality, number of words)
  – Speech rate and fluency (e.g. presence of filled pauses)
  – Social effects (gender, education)
  – Random effects for individual differences
Replicated for
[Frank and Jaeger, 2008]

\{\text{WAS, WERE, AM, ARE, IS, WILL}\} \quad \{\text{HAD, HAS, HAVE}\}

Information content of \text{BE}:
\(I(\text{BE} \mid \text{preceding context})\)

Counts
- 537, 504, 470, 436, 403, 370, 336, 302, 269, 236, 202, 168, 135, 102, 68, 34

Information content of \text{HAVE}:
\(I(\text{HAVE} \mid \text{preceding context})\)

Counts
- 169, 158, 148, 138, 127, 116, 106, 96, 85, 74, 64, 54, 43, 32, 22, 12, 1
Summary

• On the one hand:
  – Pretty solid evidence that reduction is correlated with redundancy, frequency of use, collocation, information density, and/or alike

• On the other hand:
  – A variety of theories that make similar predictions but differ conceptually and these theories have not been pitched against each other:
    • Spreading activation [cf. Arnold, 2008; Bard and Aylett, 2005]
    • Strategic coordination of planning and articulation [cf. Bell et al., 2009]
    • Automatization/chunking [Bybee, 2001, 2007; Bybee and Thompson, 1997; Bybee and Scheibman, 1999; Thompson and Mulac, 1991; cf. also exemplar-based models, Pierrehumbert, 2001]
Activation-based models

[Figure 2 from Dell 1986]
Activation and Reduction

• “[...] when reference is predictable it facilitates production processes, making the expressions shorter, more fluent, and lower in pitch” (Arnold, 2008)
Incremental activation-based accounts

*I saw three buses.*

- Revised availability account:
  - $p(r \mid I\ saw\ th)$
  - $p(ee \mid I\ saw\ thr)$
  - ...

- **Q:** Could an account like this explain the effects of information density on morphological contraction?
‘Storage’/‘automatization’/ ‘entrenchment’ accounts
[Bybee and Scheibman, 1999]

• “The more often two elements are used together, the more tightly they will be fused or bonded phonologically and semantically, and thus the tighter their constituency.” [p. 576]

• “Among the effects of repetition on grammar (see Bybee and Thompson 1997), two are of special importance to our understanding of the change of constituent structure in grammaticization: chunking (automatization) and lexical autonomy. In chunking (Haiman 1994), a frequently repeated stretch of speech becomes automated as a processing unit. [...] At the same time, high levels of usage also lead to autonomy. High-frequency collocations weaken their association with related items, as when, for example, be supposed to [sposte] (showing phonological assimilation) takes on functions increasingly less related to the meaning of suppose.” [p. 577]

• “When sequences of neuromotor routines are repeated, their execution becomes more fluent. This increased fluency is the result of the establishment of a new routine, as when a group of words comes to be processed as a single unit (Anderson 1993, Boyland 1996). In the new
‘“Storage’/‘automatization’/‘entrenchment’ accounts
[Bybee, 2007]

• “When sequences of neuromotor routines are repeated, their execution becomes more fluent. This increased fluency is the result of the establishment of a new routine, as when a group of words comes to be processed as a single unit (Anderson 1993, Boyland 1996). In the new routine articulatory gestures reduce and overlap as the routine is repeated.” [p. 715]

• **Pro:** Could explain why we reduce even when it’s bad for communication (cf. announcements by flight attendants; *can’t* vs. *can*).

• **Q:** What does this view miss? (Hint: think of professional sports, where people repeat an action over and over again or when you learn something and it becomes automatic)
Conventionalization

• Are reduction effects partially due to phonologization or lexicalization (e.g. storage of reduced forms)?
  – Interestingly, there is some evidence that the average predictability of words rather than the contextualized predictability matters for phonological deletion. [cf. Table 3 and 6 from Cohen Priva, 2008]

### Table 3

<table>
<thead>
<tr>
<th>Place</th>
<th>Voiceless Stops</th>
<th>Voiced Stops</th>
<th>Nasal Stops</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Del Pr</td>
<td>Phone Pr</td>
<td>Del Pr</td>
</tr>
<tr>
<td>Labial</td>
<td>0.010</td>
<td>0.020</td>
<td>0.099</td>
</tr>
<tr>
<td>Dorsal</td>
<td>0.022</td>
<td>0.031</td>
<td>0.038</td>
</tr>
<tr>
<td>Coronal</td>
<td>0.169</td>
<td>0.048</td>
<td>0.230</td>
</tr>
</tbody>
</table>

### Table 6

<table>
<thead>
<tr>
<th>Place</th>
<th>Voiceless Stops</th>
<th>Voiced Stops</th>
<th>Nasal Stops</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labial</td>
<td>0.010</td>
<td>3.93</td>
<td>0.099</td>
</tr>
<tr>
<td>Dorsal</td>
<td>0.022</td>
<td>2.75</td>
<td>0.038</td>
</tr>
<tr>
<td>Coronal</td>
<td>0.169</td>
<td>1.61</td>
<td>0.230</td>
</tr>
</tbody>
</table>
Different types of reduction

- What is the relation between shortening [e.g. Aylett et al., 2004, 2006; Bell et al., 2003, 2009], reduced articulation [e.g. Bybee and Scheibman, 1999; Bell et al., 2003; van Son and Pols, 2003; van Son and van Santen, 2005], phoneme omission [e.g. Bybee and Scheibman, 1999; Bell et al., 2003; Cohen Priva, 2008; Gahl and Garnsey, 2004], massive phonological reduction [e.g. Johnson, 2004], and morphological contraction [e.g. Frank & Jaeger, 2008]?

<table>
<thead>
<tr>
<th>Preceding Type</th>
<th>stop + o Group 1</th>
<th>flap + o Group 2</th>
<th>flap + θ Group 3</th>
<th>θ Group 4</th>
<th>Total no.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>16</td>
<td>22</td>
<td>38</td>
<td>12</td>
<td>88</td>
<td>63</td>
</tr>
<tr>
<td>you</td>
<td>7</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td>14</td>
</tr>
<tr>
<td>we</td>
<td>2</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>they</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>lexical NP</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>pause</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>adverb</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>neg. imper.</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>interrogative</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td></td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>43</td>
<td>44</td>
<td>39</td>
<td>12</td>
<td>138</td>
<td>100</td>
</tr>
</tbody>
</table>

[Table 1 from Bybee and Scheibman, 1999]