Lexical learning and lexical diffusion: studies on dispersion, social factors, and cultural consumption

Paula Chesley
ches0045@umn.edu

Institute of Linguistics
University of Minnesota
USA

Quantitative Measures in Morphology and Morphological Development
15 January 2011
University of California at San Diego
Q1: How are new words learned?

- Mostly examined in child, pre-adolescent, L2, and/or cognitively impaired populations
- Usually studied from a cognitive perspective, not from a social network perspective
- This talk: examine social factors involved in learning new words
  - Ingroup/outgroup and free recall, dispersion/contextual diversity, and cultural consumption
Lexical learning and lexical diffusion

Q2: What is the relationship between lexical learning and lexical diffusion?

- Experience-based model of mental lexicon
  - Lexical access is a function of input (comprehension)
  - Output (production) is a function of lexical access
  - Output is determined by input

- Lexical input is always changing
  - \[ \vdash \text{mental lexicon, lexical access, and output are also changing} \]

- This talk: provide a link between an individual’s lexicon and the lexicon of her speech community
  - Examining language as a complex adaptive system
    - Essential component: feedback loop between individual speakers and their speech communities
Lexical processes and quantitative techniques

<table>
<thead>
<tr>
<th>Study</th>
<th>Word type</th>
<th>Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dispersion</td>
<td>Borrowings</td>
<td>Linear regression, random forest</td>
</tr>
<tr>
<td>Ingroup/outgroup</td>
<td>Acronyms, blends, derived words</td>
<td>Principal components analysis, mixed-effects models</td>
</tr>
<tr>
<td>Cultural consumption</td>
<td>Lots: clippings, blends...</td>
<td>Sliced inverse regression, mixed-effects models, random forest</td>
</tr>
</tbody>
</table>

**Table:** Classification of studies.
Act 1: Social factors in the recall of new words

Act 2: Dispersion in lexical learning and lexical diffusion

Act 3: Lexical learning through cultural consumption

Technical intermezzo: focus on quantitative methods

Act 4: General discussion

Appendices: Data and models
How does input translate to output?

Input/output: speakers won’t adopt (output) a new word unless they can freely recall it from input

Q: What makes speakers recall previously unseen words?

- 2-part study on recalling new words (N = 45)
  - Learning phase: exposure and questions about a word's socio-contextual profile
  - Test phase: 3 minutes to freely recall all words from learning phase
  - Words: acronyms, blends, and derived words

- Chesley and Baayen (under revision)
Questions asked in learning phase

1. Have you seen this word before? (subjective frequency)
2. How emotional do you think this example text is?
3. How funny do you think this example text is?
4. How educated does the writer of this example text sound?
5. Would you use this word at a party with your friends?
6. Would you use this word at school or at work?
Example word-question pairing

**Word:** workmare

**Definition:** A nightmare that is derived directly from your place of employment, including your job, co-workers, duties and/or responsibilities while on the job.

**Example:** I must be working too much lately, I'm having some wicked workmares at night. The other night I had one where I botched my presentation and got fired - kept me up half the night.

**How funny do you think this example text is?**

1  2  3  4  5  6  7  very funny

not funny at all
Predictor variables

- Principal Components Analysis (PCA) done for responses to learning-phase questions

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PC5</th>
<th>PC6</th>
</tr>
</thead>
<tbody>
<tr>
<td>familiarity</td>
<td>0.17</td>
<td>0.88</td>
<td>0.31</td>
<td>0.20</td>
<td>0.25</td>
<td>-0.06</td>
</tr>
<tr>
<td>emotional</td>
<td>0.34</td>
<td>-0.28</td>
<td>0.46</td>
<td>0.68</td>
<td>-0.36</td>
<td>0.04</td>
</tr>
<tr>
<td>funny</td>
<td>0.47</td>
<td>-0.25</td>
<td>-0.29</td>
<td>0.23</td>
<td>0.66</td>
<td>-0.39</td>
</tr>
<tr>
<td>educated</td>
<td>0.31</td>
<td>-0.22</td>
<td>0.67</td>
<td>-0.58</td>
<td>0.24</td>
<td>0.09</td>
</tr>
<tr>
<td>party</td>
<td>0.54</td>
<td>0.09</td>
<td>-0.35</td>
<td>-0.06</td>
<td>-0.04</td>
<td>0.76</td>
</tr>
<tr>
<td>work</td>
<td>0.50</td>
<td>0.17</td>
<td>-0.19</td>
<td>-0.33</td>
<td>-0.56</td>
<td>-0.52</td>
</tr>
</tbody>
</table>

Table: Loadings of the six familiarization phase questions on the principal components of the elicited ratings.
Act 1: Social factors in the recall of new words

Results: ingroup/outgroup usage

- Mixed-effects model with free recall as response variable
- Ingroup usage, in informal contexts, aids recall
- Outgroup usage, in formal contexts, is detrimental to recall
Act 1 discussion: Free recall and diffusion patterns

- Results highlight the importance of social factors (e.g., ingroup/outgroup distinctions) in free recall
- For existing words: the more you think a word describes a close social tie, the more likely you are to remember it (Bower and Gilligan, 1979)
  - Similar to the Self-Reference Effect
- For new words, ingroup/outgroup associations during lexical learning could be a factor in lexical diffusion and lexical change
  - One way in which an individual’s lexicon can impact the lexicon of a speech community
- Better recall for ingroup words could be a key in understanding faster turnover rates for ingroup lexical items (slang)
Outline

Act 1: Social factors in the recall of new words

Act 2: Dispersion in lexical learning and lexical diffusion

Act 3: Lexical learning through cultural consumption

Technical intermezzo: focus on quantitative methods

Act 4: General discussion

Appendices: Data and models
Act 1: Dispersion in lexical diffusion and lexical memory

Goal: Examine input speakers are actually getting. What is the means by which input affects output?

Claim: Input affects output through dispersion

▶ Dispersion: measure of how evenly spread out a word is in a speaker’s linguistic input

Dispersion studies

▶ Psycholinguistic study: Adelman et al. (2006)
▶ Longitudinal corpus studies: Chesley and Baayen (2010), Altmann et al. (submitted)
Psycholinguistic study on dispersion and lexical access

Adelman et al. (2006): dispersion named as *contextual diversity*: number of documents in which a word occurs in a corpus

- For existing words, dispersion a better predictor of word naming and lexical decision latencies than frequency
- Possible explanation: working memory collapses across all occurrences in the same context (cf. Murphy 2003)
- The word frequency effect on lexical access is perhaps due to dispersion!
- Implication: Dispersion is an essential component to lexical learning
Longitudinal corpus study on dispersion and lexical diffusion


- Used frequency as a measure of entrenchment at T2
- Major finding: dispersion at T1 a better predictor of T2 frequency than T1 frequency
- What drives lexical entrenchment? The number of contexts a word occurs in is more important than sheer frequency
Dispersion seems to be driving both lexical access and entrenchment in the lexicon.

This argues for a dispersion feedback loop between individual speakers and the speech community:

- Speakers are sensitive to dispersion input, which affects lexical access, which affects probability of output, and output changes input for other speakers.
- Lexical learning is dependent on lexical diffusion processes (and vice versa!)
- This is one explanation of the power-law Zipfian distribution.
Outline

Act 1: Social factors in the recall of new words

Act 2: Dispersion in lexical learning and lexical diffusion

Act 3: Lexical learning through cultural consumption

Technical intermezzo: focus on quantitative methods

Act 4: General discussion

Appendices: Data and models
Examples of lexical processes in AAE music

- Metaphoric/metonymic extension of meaning: *shawty*
  Trey Songz, “I need a girl” (2009)
  Shawty where you at?

- Morphophonological reduction: *Imma*, as in “Imma let you finish...” (2009); http://www.youtube.com/watch?v=1z8gCZ7zpsQ, 0:43 - 0:59

- *izzle-, eezy-speak*: *foshizzle my nizzle, fo’ sheezy my neezy*

- Derivational morphology: *hater*
  H to the izz-O, V to the izz-A
  Fo’ sheezy my neezy keep my arms so freezy
  Can’t leave rap alone the game needs me
  Haters want me clapped and chromed it ain’t easy
Lexical learning through cultural consumption

- Chesley and Abdurrahman (in prep)
- Why do my younger siblings know features of AAE?
  - Do they have African-American friends?
  - Is it their knowledge of pop culture?
  - Is it their knowledge of musical genres typically associated with African-Americans?

Figure: My brother dresses as a hipster for Halloween.
Survey on knowledge of African-American lexical items

- Online survey asking for free-response definitions of 64 vocabulary items (N = 168)
- Follow-up questions: age, sex, hometown, social network, musical preferences, pop-culture knowledge
- Social network questions elicited both strong and weak ties
- Musical preferences: number of artists listed for each genre (9 genres total)
- Definitions transformed to a Likert scale (1-5) by two raters
Outline

Act 1: Social factors in the recall of new words

Act 2: Dispersion in lexical learning and lexical diffusion

Act 3: Lexical learning through cultural consumption

Technical intermezzo: focus on quantitative methods

Act 4: General discussion

Appendices: Data and models
Predicting AAE lexical knowledge

Basic idea: use demographic, musical preferences, and pop-culture questions to predict participant vocabulary scores with regression

Problem: lots of predictor variables (dimensionality reduction)

- Use principal components analysis (PCA) to minimize number of predictors? But we have categorical predictors too
- Use PCA on a subset of predictor variables? Problems with PCA:
  - Response variable not taken into account when choosing components
  - How to interpret loadings of predictor variables on components?
One solution: Sliced inverse regression (SIR)

- Tries to find the linear combination of predictors necessary to explain the response variable
- Uses response variable to determine whether each predictor is necessary once all other variables are taken into account
- Unlike PCA, uses best linear combination of predictors with respect to the response variable

SIR: dimensionality reduction for predictor variables
Predictor selection with SIR

- SIR can also determine whether higher-order predictor terms like interactions and quadratic terms are needed
- Stepwise backward selection with SIR similar to `stepAIC` with linear models (function `dr.step()` in `dr` package in R)
- Stopping test used: dimensionality reduction coordinate test
  - Tests for conditional independence of all predictors
- Next, create a model with predictors given in final iteration
  - For the present study, a linear model was appropriate
Results from linear model with predictors from SIR

Added-Variable Plots

- hiphop | others
- country | others
- blackWeekly | others
- barkley | others
- boondocks1 | others
Convergent evidence for effects of hiphop, weak ties

- Results with non-parametric model: random forest

**Figure**: AAE lexical knowledge: partial dependence on (a) hiphop music, (b) weak ties to African-Americans.
Act 3 discussion: learning AAE lexical items

- Robust associations between musical preferences (and pop-culture knowledge) and AAE vocabulary scores
- Increased weak social ties to African-Americans also associated with higher AAE vocabulary scores
  - Finding consistent with the Strength of Weak Ties theory (Granovetter 1973)
- Significant predictors are those that a speaker has some control over (agency)
- Evidence that broadcast nodes have more connections than previously thought (for previous claims, see e.g. Labov 2001: 356-357)
Outline

Act 1: Social factors in the recall of new words

Act 2: Dispersion in lexical learning and lexical diffusion

Act 3: Lexical learning through cultural consumption

Technical intermezzo: focus on quantitative methods

Act 4: General discussion

Appendices: Data and models
How to relate these findings?

- Importance of social context in lexical learning
  - Lexical access dependent upon dispersion/number of speakers using the term
    - For new words, we can infer similar processes
  - Free recall improved for hypothetical ingroup usage

- Importance of lexical learning for diffusion of new words
  - A new word needs to be recalled before it can be adopted
  - Possible relation between memory and age of lexical innovators
    - Younger people have better memory; they are also more likely to be linguistic innovators
Language as a complex adaptive system

Figure: Language as a complex adaptive system.

- Feedback between individual and speech community
  - Input to individual affects lexical access, which affects individual output
    - Model needed that gives linguistic output as a function of linguistic input
Language as a complex adaptive system II

- Broadcast nodes in social networks
  - Have more connections than previously assumed
  - Asymmetric links

Figure: A broadcast node in a social network.
Methodological notes

- Use of Sliced Inverse Regression as a dimensionality reduction tool
- Use of multiple methods to establish convergent evidence for results
- ... suggestions?
Conclusion and future directions

- Lexical learning is a highly social process
- Lexical access is based on dispersion input
- Lexical output based on lexical access
- This way, we can account for the interaction between an individual’s lexicon and the lexicon of her speech community
- Future lines of inquiry
  - More longitudinal experimental studies
  - Better modeling capabilities of input/output processes and language as a complex adaptive system
References I


E.G. Altmann, J.B. Pierrehumbert, and A.E. Motter. Niche as a determinant of word fate in online groups. submitted.


P. Chesley and R. H. Baayen. Structural and social determinants of memory for new words. under revision.
M.S. Granovetter.
The Strength of Weak Ties.

W. Labov.
*Principles of Linguistic Change: Social Factors.*

G.L. Murphy.
The Big book of concepts.
Outline

Act 1: Social factors in the recall of new words

Act 2: Dispersion in lexical learning and lexical diffusion

Act 3: Lexical learning through cultural consumption

Technical intermezzo: focus on quantitative methods

Act 4: General discussion

Appendices: Data and models
## Appendix I: Dispersion study model

|               | $\beta$ | S.E. | t value | Pr($>|t|)$ |
|---------------|---------|------|---------|-----------|
| (Intercept)   | 1.496   | 0.734| 2.039   | 0.042     |
| Dispersion    | 2.322   | 0.123| 18.934  | $<0.001$  |
| Frequency     | −0.801  | 1.012| −0.791  | 0.429     |
| Length        | −0.599  | 0.355| −1.688  | 0.093     |
| Context (restricted) | 0.564  | 0.833| 0.677   | 0.499     |
| Sense (poly)  | 2.230   | 0.513| 4.347   | $<0.001$  |
| Language (eng)| −0.755  | 0.530| −1.425  | 0.155     |
| Frequency*Dispersion | −3.324 | 0.692| −4.806  | $<0.001$  |
| Frequency*Context (restricted) | 2.531  | 0.865| 2.927   | 0.004     |
| Length*Context (restricted)   | −1.721  | 0.468| −3.676  | $<0.001$  |
| Sense (poly)*Context (restricted) | −2.016 | 0.744| −2.710  | 0.007     |
| Language (eng)*Context (restricted) | 1.837  | 0.586| 3.137   | 0.002     |

**Table:** A multiple regression model for predicting entrenchment of lexical borrowings into French.
Appendix I: Dispersion study data

<table>
<thead>
<tr>
<th>Borrowing</th>
<th>T1 Freq</th>
<th>T2 Freq</th>
<th>Borrowing</th>
<th>T1 Freq</th>
<th>T2 Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>popiwek</td>
<td>1</td>
<td>0</td>
<td>lobbying</td>
<td>3</td>
<td>865</td>
</tr>
<tr>
<td>taref</td>
<td>1</td>
<td>0</td>
<td>come-back</td>
<td>1</td>
<td>333</td>
</tr>
<tr>
<td>the</td>
<td>1</td>
<td>0</td>
<td>hedge funds</td>
<td>3</td>
<td>368</td>
</tr>
<tr>
<td>huasipongo</td>
<td>2</td>
<td>0</td>
<td>perestroïka</td>
<td>8</td>
<td>210</td>
</tr>
<tr>
<td>ejido</td>
<td>9</td>
<td>0</td>
<td>success story</td>
<td>2</td>
<td>382</td>
</tr>
<tr>
<td>classless society</td>
<td>1</td>
<td>0</td>
<td>running</td>
<td>1</td>
<td>53</td>
</tr>
</tbody>
</table>

Table: Dispersion study: examples of nonce and productive borrowings and their respective frequencies in the T1 and T2 corpora.
Appendix II: Social factors and recall model

Mixed-effects logistic regression model (0 = not recalled, 1 = recalled)

|                          | Estimate | Std. Error | z value | Pr(>|z|) |
|--------------------------|----------|------------|---------|----------|
| Intercept                | -2.9740  | 0.3194     | -9.3123 | 0.0000   |
| Trial Index              | 0.1459   | 0.0528     | 2.7620  | 0.0057   |
| Response = no            | -1.2914  | 0.2293     | -5.6309 | 0.0000   |
| RT (neg inv transformation) | -0.6924 | 0.1950     | -3.5506 | 0.0004   |
| PC3                      | -0.2015  | 0.0685     | -2.9423 | 0.0033   |
| PC5                      | 0.2100   | 0.0813     | 2.5822  | 0.0098   |
| PC6                      | 0.2810   | 0.1064     | 2.6418  | 0.0082   |
| Sleep Condition = sleep  | -0.4717  | 0.2914     | -1.6187 | 0.1055   |
| Type = blend             | 0.4782   | 0.3090     | 1.5476  | 0.1217   |
| Type = derivation        | 0.4249   | 0.3104     | 1.3688  | 0.1711   |
| Sleep Condition = sleep : Type = blend | 0.5662 | 0.3047 | 1.8580  | 0.0632   |
| Sleep Condition = sleep : Type = derivation | 1.0160 | 0.3019 | 3.3651  | 0.0008   |

**Table:** Model coefficients — Recall. Reference level = not recalled.
Appendix II: Social factors and recall stimuli – acronyms

AATP  ADF  BGP  BITGOD  CLM
CVOC  DBI  DLS  DTR  FD
FSBO  FSP  FUSSDIRAG  ICTYIAS  LIMH
MIRF  NCMO  NGL  OMS  PINO
SMV  UDI  URST  VOCD  WAM
## Appendix II: Social factors and recall stimuli – blends

<table>
<thead>
<tr>
<th>anablog</th>
<th>chairdrobe</th>
<th>chickchismo</th>
</tr>
</thead>
<tbody>
<tr>
<td>compunicate</td>
<td>dejabrew</td>
<td>destinesia</td>
</tr>
<tr>
<td>econnoisseur</td>
<td>elecoustic</td>
<td>enterdrainment</td>
</tr>
<tr>
<td>gaybie</td>
<td>grade digger</td>
<td>hater tots</td>
</tr>
<tr>
<td>man stand</td>
<td>manther</td>
<td>multi-slacking</td>
</tr>
<tr>
<td>mysterectomy</td>
<td>pornocchio</td>
<td>premature evacuation</td>
</tr>
<tr>
<td>showmance</td>
<td>spim</td>
<td>stoptional</td>
</tr>
<tr>
<td>testosterphone</td>
<td>tomorrow</td>
<td>trumor</td>
</tr>
<tr>
<td>workmare</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix II: Social factors and recall stimuli – derived words

<table>
<thead>
<tr>
<th>Halophile</th>
<th>McDad</th>
<th>anti-anti-Semite</th>
<th>bio-accessory</th>
</tr>
</thead>
<tbody>
<tr>
<td>cannabista</td>
<td>de-eat</td>
<td>e-haircut</td>
<td>e-tact</td>
</tr>
<tr>
<td>e-void</td>
<td>errorist</td>
<td>fiscalize</td>
<td>hypertasking</td>
</tr>
<tr>
<td>innuendo-ish</td>
<td>obliviation</td>
<td>oldiephile</td>
<td>playlistism</td>
</tr>
<tr>
<td>pre-eat</td>
<td>pre-walk</td>
<td>re-cop</td>
<td>resolutionary</td>
</tr>
<tr>
<td>sapiosexual</td>
<td>un-trade-upable</td>
<td>underdrunk</td>
<td>unfull</td>
</tr>
<tr>
<td>wordanista</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Appendix III: AAE lexical learning model

**Table**: Linear regression model for predicting AAE lexical knowledge.

|            | Estimate | Std. Error | t value | Pr(>|t|) |
|------------|----------|------------|---------|----------|
| (Intercept)| 67.9463  | 6.1654     | 11.02   | 0.0000   |
| hiphop     | 3.2459   | 0.9265     | 3.50    | 0.0006   |
| country    | -3.5464  | 1.4673     | -2.42   | 0.0168   |
| barkley    | 6.5717   | 1.6998     | 3.87    | 0.0002   |
| boondocks  | 5.7915   | 1.4354     | 4.03    | 0.0001   |
| monique    | 5.6488   | 1.8417     | 3.07    | 0.0025   |
| jayz       | 2.7452   | 1.3185     | 2.08    | 0.0389   |
### Appendix III: AAE lexical learning items

<table>
<thead>
<tr>
<th>ballin’</th>
<th>boughie</th>
<th>chitlins</th>
<th>cheese</th>
</tr>
</thead>
<tbody>
<tr>
<td>crump</td>
<td>dead presidents</td>
<td>dollar cab</td>
<td>facheezie</td>
</tr>
<tr>
<td>finna</td>
<td>good hair</td>
<td>grip</td>
<td>heezy</td>
</tr>
<tr>
<td>player hater</td>
<td>road dog</td>
<td>roll deep</td>
<td>saditty</td>
</tr>
<tr>
<td>straight cash</td>
<td>toe up</td>
<td>trill</td>
<td>wile out</td>
</tr>
</tbody>
</table>

**Table:** Examples of AAE lexical items.