Bayesian Models of Lexicon Learning

Timothy J. O’Donnell
McGill Linguistics
Language Learning and the Lexicon

Language learning consists mostly (entirely?) of learning the lexicon (words, morphemes, etc.)?
The Problem of Lexical Uncertainty

How do learners identify the lexical units in their language?
The Problem of Lexical Uncertainty

<table>
<thead>
<tr>
<th>Suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>-ness</td>
</tr>
<tr>
<td>-th</td>
</tr>
<tr>
<td>-ity</td>
</tr>
</tbody>
</table>
## The Problem of Lexical Uncertainty

<table>
<thead>
<tr>
<th>Nouns (N)</th>
<th>Attach to adjectives (Adj)</th>
<th>Mean abstract quality or state</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Goodness, cheapness, forgiveness, circuitousness, grandness, orderliness, business, goodness, …</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Suffix

<table>
<thead>
<tr>
<th>-ness</th>
</tr>
</thead>
<tbody>
<tr>
<td>-th</td>
</tr>
<tr>
<td>-ity</td>
</tr>
</tbody>
</table>
The Problem of Lexical Uncertainty

<table>
<thead>
<tr>
<th>Nouns (N)</th>
<th>Attach to adjectives (Adj)</th>
<th>Suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>goodness, cheapness, forgiveness, circuitousness, grandness, orderliness, business, goodness, …</td>
<td>truth, warmth, width, depth, filth, sloth, strength, death, dearth, wealth, length, youth, …</td>
<td>-ness</td>
</tr>
<tr>
<td>-th</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-ity</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The Problem of Lexical Uncertainty

<table>
<thead>
<tr>
<th>Nouns (N)</th>
<th>Attach to adjectives (Adj)</th>
<th>Suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>goodness, cheapness, forgiveness,</td>
<td></td>
<td>-ness</td>
</tr>
<tr>
<td>circuitousness, grandness, orderliness,</td>
<td></td>
<td>-th</td>
</tr>
<tr>
<td>business, goodness, …</td>
<td></td>
<td>-ity</td>
</tr>
<tr>
<td>truth, warmth, width, depth, filth, sloth,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>strength, death, dearth, wealth, length,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>youth, …</td>
<td></td>
<td></td>
</tr>
<tr>
<td>verticality, tractability, severity, seniority,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>inanity, electricity, parity, scarcity, reality,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>…</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## The Problem of Lexical Uncertainty

<table>
<thead>
<tr>
<th></th>
<th>Suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Productive</td>
<td>-ness</td>
</tr>
<tr>
<td>Unproductive</td>
<td>-th</td>
</tr>
<tr>
<td>Productive in Combination</td>
<td>-ity</td>
</tr>
</tbody>
</table>
The Problem of Lexical Uncertainty

<table>
<thead>
<tr>
<th>Productive</th>
<th>Suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>pine-scented</td>
<td>-ness</td>
</tr>
</tbody>
</table>

pine-scentedness
### The Problem of Lexical Uncertainty

<table>
<thead>
<tr>
<th>Productive</th>
<th>Suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>-ness</td>
<td></td>
</tr>
</tbody>
</table>

- goodness, cheapness, forgiveness, circuitousness, grandness, orderliness, pretentiousness, business, goodness, greenness, …
The Problem of Lexical Uncertainty

<table>
<thead>
<tr>
<th></th>
<th>Suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Productive</td>
<td>-ness</td>
</tr>
</tbody>
</table>

- *goodness, cheapness, forgiveness, circuitousness, grandness, orderliness, pretentiousness, business, greenness, …*

Phonological\ Morpho-Syntactic\ Semantic
Regularities\ Regularities\ Regularities
The Problem of Lexical Uncertainty

<table>
<thead>
<tr>
<th>Suffix</th>
<th>Productive</th>
</tr>
</thead>
<tbody>
<tr>
<td>-ness</td>
<td></td>
</tr>
</tbody>
</table>

N

Adj

-ness
## The Problem of Lexical Uncertainty

<table>
<thead>
<tr>
<th></th>
<th>Suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Productive</strong></td>
<td>-ness</td>
</tr>
<tr>
<td><strong>Unproductive</strong></td>
<td>-th</td>
</tr>
</tbody>
</table>

- *coolth*
The Problem of Lexical Uncertainty

<table>
<thead>
<tr>
<th>Suffix</th>
<th>Productive</th>
<th>Unproductive</th>
</tr>
</thead>
<tbody>
<tr>
<td>-ness</td>
<td>truth, warmth, width, depth, filth, sloth, strength, death, dearth, wealth, length, youth, …</td>
<td>-th</td>
</tr>
</tbody>
</table>
The Problem of Lexical Uncertainty

<table>
<thead>
<tr>
<th></th>
<th>Suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Productive</td>
<td>-ness</td>
</tr>
<tr>
<td>Unproductive</td>
<td>-th</td>
</tr>
</tbody>
</table>

*truth, warmth, width, depth, filth, sloth, strength, death, dearth, wealth, length, youth, …*
The Problem of Lexical Uncertainty

<table>
<thead>
<tr>
<th>Productive in Combination</th>
<th>Suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Productive</td>
<td>-ness</td>
</tr>
<tr>
<td>Unproductive</td>
<td>-th</td>
</tr>
</tbody>
</table>

- pine-scented
- cool

- *pine-scentedity
- *coolity
## The Problem of Lexical Uncertainty

<table>
<thead>
<tr>
<th></th>
<th>Suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Productive</td>
<td>-ness</td>
</tr>
<tr>
<td>Unproductive</td>
<td>-th</td>
</tr>
<tr>
<td>Productive in Combination</td>
<td>-ity</td>
</tr>
</tbody>
</table>
# The Problem of Lexical Uncertainty

<table>
<thead>
<tr>
<th>Productive</th>
<th>Suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Productive in Combination</td>
<td>-ity</td>
</tr>
<tr>
<td>Unproductive</td>
<td>-th</td>
</tr>
</tbody>
</table>

- -ile, -al, -able, -ic, -(i)an
The Problem of Lexical Uncertainty

<table>
<thead>
<tr>
<th></th>
<th>Suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Productive</td>
<td>-ness</td>
</tr>
<tr>
<td>Unproductive</td>
<td>-th</td>
</tr>
<tr>
<td>Productive in Combination</td>
<td>-ity</td>
</tr>
</tbody>
</table>

sequentiable sequentiability
The Problem of Lexical Uncertainty

<table>
<thead>
<tr>
<th></th>
<th>Suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Productive</td>
<td>-ness</td>
</tr>
<tr>
<td>Unproductive</td>
<td>-th</td>
</tr>
<tr>
<td>Productive in Combination</td>
<td>-ity</td>
</tr>
</tbody>
</table>

-Adj

\[\text{N} \rightarrow \text{Adj} \rightarrow \text{ity}\]

\[\text{N} \rightarrow \text{V} \rightarrow \text{-ability}\]
The Problem of Lexical Uncertainty

<table>
<thead>
<tr>
<th></th>
<th>Suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Productive</td>
<td>-ness</td>
</tr>
<tr>
<td>Unproductive</td>
<td>-th</td>
</tr>
<tr>
<td>Productive in Combination</td>
<td>-ity</td>
</tr>
</tbody>
</table>
The Problem of Lexical Uncertainty

<table>
<thead>
<tr>
<th>Productive</th>
<th>-ness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unproductive</td>
<td>-th</td>
</tr>
<tr>
<td>Productive in Combination</td>
<td>-ity</td>
</tr>
</tbody>
</table>
Overview

Part I: An approach to lexical uncertainty.

Part II: Detailed case study in morphology.

Part III: Unsupervised models of language learning.
Overview

Part I: An approach to lexical uncertainty.

Part II: Detailed case study in morphology.

Part III: Unsupervised models of language learning.
Bayesian Approach

• Use probabilistic models to define prior distribution over possible lexicons.

• Find posterior distribution over lexicons given actual data using probabilistic conditioning.

\[ P(L \mid D) \propto P(D \mid L) \, P(L) \]
The Framework

1. An underlying computational system that defines space of possible structures.

2. Manage uncertainty over the ways in which forms can be analyzed with a probabilistic model implementing a tradeoff between storage and computation (prior and likelihood).

3. Use probabilistic inference to derive language-specific predictions.
The Framework

1. An underlying computational system that defines space of possible structures.

2. Manage uncertainty over the ways in which forms can be analyzed with a probabilistic model implementing a tradeoff between storage and computation (prior and likelihood).

3. Use probabilistic inference to derive languagespecific predictions.
Underlying Computational System

1. Inventory of structured units (lexicon).

2. Structure-building operations.
Lexical Items

N

Adj -ity

Stems, suffixes, prefixes with combinatorial restrictions.
Structure Building

N

Adj -ity

Unit-driven selection.
Structure Building

N                    Unit-driven selection.
    /\                  /   \
   Adj    -ity
            /\     \   /
           V     -able
Structure Building

Unit-driven selection.

Diagram:

```
N
  /\  \
Adj -ity
  /\  \
V -able
    /\  \
agree
```
The Framework

1. An underlying computational system that defines space of possible structures.

2. Manage uncertainty over the ways in which forms can be analyzed with a probabilistic model implementing a tradeoff between storage and computation (prior and likelihood).

3. Use probabilistic inference to derive language-specific predictions.
Uncertainty

Phonological, semantic, and morphosyntactic processes generate candidate analyses.

```
adj
  /\  -able
 V   -ity
    /\  \
   agree
```
Uncertainty

\[
\text{N} \rightarrow \text{Adj} \rightarrow \text{V} \\
\text{N} \rightarrow \text{Adj} \rightarrow \text{V} \\
\text{V} \rightarrow \text{able} \\
\text{Adj} \rightarrow \text{able} \\
\text{N} \rightarrow \text{ity} \\
\text{V} \rightarrow \text{agree} \\
\text{Adj} \rightarrow \text{ity} \\
\text{V} \rightarrow \text{agree}
\]
Uncertainty

N

Adj

-able

-it

N

V

agree

N

V

-ability

agree
Probabilistic Model

• **Goal:** Define a probability distribution that quantifies uncertainty about likely and unlikely lexical units.

• Prior and likelihood lead to a **tradeoff** between storage and computation.

1. Prior over lexicon: $P(L)$.
2. Likelihood over derivations of individual forms: $P(D \mid L)$. 
Prior on Lexicon

• Prefer a small number of highly reusable (i.e., generalizable) units.

• Dirichlet Process (Ferguson, 1973; Aldous, 1983; Sethuraman, 1994; Pitman 1995).
  • Allows unbounded number of stored units (non-parametric).
  • Prefers fewer units.
  • Prefers more reusable units.

\[ P(\text{unit}) \propto \text{frequency of use} \]
Likelihood of Derived Forms

• Simple observed forms with few parts.

• Probability of derivation: Product of probability of individual units.
  
  • Probabilities between 0 and 1.
    
    • Geometric decrease in probability.

• Prefers fewer lexical items per derivation.

• Prefers to store more complex units.
The Framework

1. An underlying computational system that defines space of possible structures.

2. Manage uncertainty over the ways in which forms can be analyzed with a probabilistic implementing a tradeoff between storage and computation.

3. Use probabilistic inference to derive language-specific predictions.
The Mathematical Model: *Fragment Grammars*

- Generalization of *Adaptor Grammars* (Johnson et al., 2007).
- Allows storing of partial trees.
- Can be generalized with stochastic variants of memoization and order of evaluation.
Inference Problem

Search for sets of stored units and derivations of individual forms which best explain the input taking into account preferences for a small inventory of highly reusable units and simple derivations of individual forms.
Example Input

```
Example Input

N
  Adj
  -ity
  V
  agree

N
  Adj
  -ness
  V
  agree

N
  Adj
  -ity
  V
  count

N
  Adj
  -ity
  V
  agree
```
Maximal Decomposition
Derivation Complexity

```
   N
  / \  /  \
 Adj -ity
  /  /  /
 V -able
   |   |
 agree
```
Derivation Complexity

P(Adj -ity)

N

Adj

-ity

V

-able

agree
Derivation Complexity

\[ P(N) \times P(V) \]

\[ \text{agree} \]
Derivation Complexity

\[
P(\text{N}) \times P(\text{Adj} - \text{ity}) \times P(\text{V} - \text{able}) 
\]

Geometric Decrease in Probability
Sharing Across Expressions

N  Adj  -ity
V  able  agree

N  Adj  -ness
V  able  agree

N  Adj  -ity
V  able  count

N  Adj  -ity
V  able  agree
Sharing Across Expressions

\[
P(\text{Adj} \rightarrow \text{N}) \propto 3
\]
No Decomposition
Derivation Complexity

\[
\begin{array}{c}
N \\
\text{Adj} & -ity \\
V & -able \\
\mid & \\
agree
\end{array}
\]
Derivation Complexity

P( agree )
Sharing Across Expressions

```
N
  Adj -ity
    V -able
      agree

N
  Adj -ness
    V -able
      agree

N
  Adj -ity
    V -able
      count

N
  Adj -ity
    V -able
      agree
```
Sharing Across Expressions

\[ P(\text{agree}) \propto 2 \]
Intermediate Decomposition

- Adj - able
- N - ity
- V - agree

- Adj - able
- N - ness
- V - agree

- Adj - able
- N - ity
- V - count

- Adj - able
- N - ity
- V - agree
Derivation Complexity

N

Adj

-ity

V

-able

agree
Derivation Complexity

P( agree )
Derivation Complexity

\[ P(\text{\textit{agree}}) \times P(\text{\textit{able}}) \times P(\text{\textit{ity}}) \]

Diagram: Tree structure with nodes labeled as follows:
- **N** (Noun)
- **Adj** (Adjective)
- **V** (Verb)
- **-able**
- **-ity**
- **agree**
Sharing Across Expressions

```
N
  Adj
  -ity
V
  able
  agree

N
  Adj
  -ness
V
  able
  agree

N
  Adj
  -ity
V
  able
  count

N
  Adj
  -ity
V
  able
  agree
```
Sharing Across Expressions
Computation/Storage Tradeoff
Computation/Storage Tradeoff

Builds on both of classic and recent work in non-parametric Bayesian and minimum description length framework.s

Has connections to theory of programming languages, Kolmogorov complexity, and other areas, via probabilistic programming and program induction frameworks.
Overview

Part I: An approach to lexical uncertainty.

Part II: Detailed case study in morphology.

Part III: Unsupervised models of language learning.
## English Derivational Morphology

<table>
<thead>
<tr>
<th></th>
<th>Suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Productive</td>
<td>-ness</td>
</tr>
<tr>
<td>Unproductive</td>
<td>-th</td>
</tr>
<tr>
<td>Productive in Combination</td>
<td>-ity</td>
</tr>
</tbody>
</table>
Morphological Productivity

1. Background
2. Productivity and frequency
3. Affix ordering
Morphological Productivity

1. Background
2. Productivity and frequency
3. Affix ordering
Idealized Learner Study

- Not intended to correspond to a particular stage of development or learning procedure used by children.

- **Question:** Does the computation/storage tradeoff provide an effective way of distinguishing between productive units like -ness and false generalizations like -th?
Input Representations

Phonologically, semantically, and morphosyntactically plausible candidate analyses, erring on the side of shareable structure (decomposition).
Models

1. Differ in storage strategy, otherwise as closely matched as possible.

2. Same inputs and representation space.

3. Represent state-of-the-art implementations from natural language processing.

---

**Full-Parsing**

( DMPCFG )

**Full-Storage**

(MAG)

**Inference-Based**

(FG)
Full-Parsing

(MAP Dirichlet-Multinomial Context-Free Grammars: DMPCFG)

Johnson, Griffiths, & Goldwater (2007a)

- **Storage Strategy**: Maximally Decompose.

- Implements radical version of decompositional theories from psychology (e.g., Taft, 1988) and linguistics (e.g., Halle & Marantz, 1983).
Full-Parsing

(_MAP Dirichlet-Multinomial Context-Free Grammars: DMPCFG)

Johnson, Griffiths, & Goldwater (2007a)

- Storage Strategy: Maximally Decompose.

- Best thought of as a baseline system for morphology.
Full-Parsing

(MAP Dirichlet-Multinomial Context-Free Grammars: DMPCFG)

Johnson, Griffiths, & Goldwater (2007a)

- **Storage Strategy:** Maximally Decompose.

Productivity

\[ P(\text{Adj} \rightarrow \text{N} - \text{ness}) \propto \text{Token Frequency} \]
Models

Full-Parsing (DMPCFG)

Full-Storage (MAG)

Inference-Based (FG)
Full-Storage

(MAP All-Adapted Adaptor Grammars: MAG)

Johnson, Griffiths, & Goldwater (2007)

- **Storage Strategy:** Store structures in their entirety after first time generated.
Full-Storage
(MAP All-Adapted Adaptor Grammars: \texttt{MAG})

Johnson, Griffiths, & Goldwater (2007)

- **Storage Strategy:** Store structures in their entirety after first time generated.

All full sub-trees are stored recursively.
**Full-Storage**

(MAP All-Adapted Adaptor Grammars: MAG)

Johnson, Griffiths, & Goldwater (2007)

- **Storage Strategy:** Store structures in their entirety after first time generated.

All full sub-trees are stored recursively.
Full-Storage
(MAP All-Adapted Adaptor Grammars: MAG)
Johnson, Griffiths, & Goldwater (2007)

- **Storage Strategy**: Store structures in their entirety after first time generated.

All full sub-trees are stored recursively.
Full-Storage

(MAP All-Adapted Adaptor Grammars: MAG)

Johnson, Griffiths, & Goldwater (2007)

- **Storage Strategy:** Store structures in their entirety after first time generated.

All full sub-trees are stored recursively.
Full-Storage
(MAP All-Adapted Adaptor Grammars: MAG)
Johnson, Griffiths, & Goldwater (2007)

- **Storage Strategy:** Store structures in their entirety after first time generated.

- Modern probabilistic implementation of classical lexical redundancy rules (e.g., Jackendoff, 1975; Aronoff, 1976).
**Full-Storage**

(MAP All-Adapted Adaptor Grammars: MAG)

Johnson, Griffiths, & Goldwater (2007)

- **Storage Strategy:** Store structures in their entirety after first time generated.

\[
P(A) \propto \text{Type frequency}
\]
Inference-Based

(Fragment Grammars: FG)


• **Storage Strategy:** Store set of units that best explains data (only inference-based storage proposal).
Inference-Based
(Fragment Grammars: FG)

• **Storage Strategy**: Store set of units that best explains data (only inference-based storage proposal).

• Generalization of **Full-Storage** (AG) model (Johnson et al., 2007).
Models

Full-Parsing (DMPCFG)

Full-Storage (MAG)

Inference-Based (FG)
Morphological Productivity

1. Background
2. Productivity and frequency
3. Affix ordering
Productivity and the Distribution of Forms

What aspects of the distribution of words and part-words signal that some structure is an independent, productive lexical unit?
## Five Most Productive Suffixes (Predicted)

### Full-Parsing (MDPCFG)

<table>
<thead>
<tr>
<th>Suffix</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>ion:V&gt;N</td>
<td>regression</td>
</tr>
<tr>
<td>ly:Adj&gt;Adv</td>
<td>quickly</td>
</tr>
<tr>
<td>ate:BND&gt;V</td>
<td>segregate</td>
</tr>
<tr>
<td>ment:V&gt;N</td>
<td>development</td>
</tr>
<tr>
<td>er:V&gt;N</td>
<td>talker</td>
</tr>
</tbody>
</table>

### Inference-Based (FG)

<table>
<thead>
<tr>
<th>Suffix</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>ly:Adj&gt;Adv</td>
<td>quickly</td>
</tr>
<tr>
<td>er:V&gt;N</td>
<td>talker</td>
</tr>
<tr>
<td>ncss:Adj&gt;N</td>
<td>tallness</td>
</tr>
<tr>
<td>y:N&gt;Adj</td>
<td>mousey</td>
</tr>
<tr>
<td>er:N&gt;N</td>
<td>prisoner</td>
</tr>
</tbody>
</table>

### Full-Storage (MAG)

<table>
<thead>
<tr>
<th>Suffix</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>ly:Adj&gt;Adv</td>
<td>quickly</td>
</tr>
<tr>
<td>ion:V&gt;N</td>
<td>regression</td>
</tr>
<tr>
<td>er:V&gt;N</td>
<td>talker</td>
</tr>
<tr>
<td>ly:V&gt;Adv</td>
<td>bitingly</td>
</tr>
<tr>
<td>y:N&gt;Adj</td>
<td>mousey</td>
</tr>
</tbody>
</table>
Five Most Productive Suffixes (Predicted)

-ness

<table>
<thead>
<tr>
<th>Full-Parsing (MDPCFG)</th>
<th>Inference-Based (FG)</th>
<th>Full-Storage (MAG)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Suffix</strong></td>
<td><strong>Example</strong></td>
<td><strong>Suffix</strong></td>
</tr>
<tr>
<td>ion:V&gt;N</td>
<td>regression</td>
<td>ly:Adj&gt;Adv</td>
</tr>
<tr>
<td>ly:Adj&gt;Adv</td>
<td>quickly</td>
<td>er:V&gt;N</td>
</tr>
<tr>
<td>ate:BND&gt;V</td>
<td>segregate</td>
<td>ness:Adj&gt;N</td>
</tr>
<tr>
<td>ment:V&gt;N</td>
<td>development</td>
<td>y:N&gt;Adj</td>
</tr>
<tr>
<td>er:V&gt;N</td>
<td>talker</td>
<td>er:N&gt;N</td>
</tr>
</tbody>
</table>
Five Most Productive Suffixes (Predicted)

**-ion**

<table>
<thead>
<tr>
<th><strong>Full-Parsing</strong> (MDPCFG)</th>
<th><strong>Inference-Based</strong> (FG)</th>
<th><strong>Full-Storage</strong> (MAG)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Suffix</strong></td>
<td><strong>Example</strong></td>
<td><strong>Suffix</strong></td>
</tr>
<tr>
<td>ion:V&gt;N</td>
<td>regression</td>
<td>ly:Adj&gt;Adv</td>
</tr>
<tr>
<td>ly:Adj&gt;Adv</td>
<td>quickly</td>
<td>er:V&gt;N</td>
</tr>
<tr>
<td>ate:BND&gt;V</td>
<td>segregate</td>
<td>ncss:Adj&gt;N</td>
</tr>
<tr>
<td>ment:V&gt;N</td>
<td>development</td>
<td>y:N&gt;Adj</td>
</tr>
<tr>
<td>er:V&gt;N</td>
<td>talker</td>
<td>er:N&gt;N</td>
</tr>
</tbody>
</table>

**eat : *eation**
Productivity and Frequency Distributions

- Productive suffixes give rise to new forms.
- **Distributional consequence**: Large proportion of low-frequency forms.
- Large number of rare event distributions (LNRE) investigated mathematically (Khmaladze, 1987; Baayen, 2001).

- Used to develop useful statistical tools to quantify and study productivity in corpus samples (e.g., Baayen, 1992).
  - More below.
Productivity and The Proportion of Low-Frequency Forms

High Proportion of Low-Frequency Forms

Word Frequency (Log Scale)
Productivity and The Proportion of Low-Frequency Forms

- **ness**
  - 34% Freq-One Forms

- **-ion**
  - .05% Freq-One Forms
Type and Token Frequencies

-ness

-ion

15,568 Tokens

1024 Types

162,573 Tokens

1117 Types

Word Frequency (Log Scale)
## Five Most Productive Suffixes (Predicted)

### Full-Parsing (MDPCFG)

<table>
<thead>
<tr>
<th>Suffix</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>ion:V&gt;N</td>
<td>regression</td>
</tr>
<tr>
<td>ly:Adj&gt;Adv</td>
<td>quickly</td>
</tr>
<tr>
<td>ate:BND&gt;V</td>
<td>segregate</td>
</tr>
<tr>
<td>ment:V&gt;N</td>
<td>development</td>
</tr>
<tr>
<td>er:V&gt;N</td>
<td>talker</td>
</tr>
</tbody>
</table>

### Inference-Based (FG)

<table>
<thead>
<tr>
<th>Suffix</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>ly:Adj&gt;Adv</td>
<td>quickly</td>
</tr>
<tr>
<td>er:V&gt;N</td>
<td>talker</td>
</tr>
</tbody>
</table>

### High Proportion of Low-Frequency Forms

- **ness:Adj>N**
  - tallness
- **y:N>Adj**
  - mousey
- **er:N>N**
  - prisoner
Five Most Productive Suffixes (Predicted)

- **ion**

<table>
<thead>
<tr>
<th>Suffix</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>ion:V&gt;N</td>
<td>regression</td>
</tr>
<tr>
<td>ly:Adj&gt;Adv</td>
<td>quickly</td>
</tr>
<tr>
<td>ate:BND&gt;V</td>
<td>segregate</td>
</tr>
<tr>
<td>ment:V&gt;N</td>
<td>development</td>
</tr>
<tr>
<td>er:V&gt;N</td>
<td>talker</td>
</tr>
</tbody>
</table>

**High Token Frequency**

**Inference-Based (FG)**

<table>
<thead>
<tr>
<th>Suffix</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>ly:Adj&gt;Adv</td>
<td>quickly</td>
</tr>
<tr>
<td>er:V&gt;N</td>
<td>talker</td>
</tr>
<tr>
<td>ncss:Adj&gt;N</td>
<td>tallness</td>
</tr>
<tr>
<td>y:N&gt;Adj</td>
<td>mousey</td>
</tr>
<tr>
<td>er:N&gt;N</td>
<td>prisoner</td>
</tr>
</tbody>
</table>

**High Type Frequency**

<table>
<thead>
<tr>
<th>Suffix</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>ion:V&gt;N</td>
<td>regression</td>
</tr>
<tr>
<td>er:V&gt;N</td>
<td>talker</td>
</tr>
<tr>
<td>ly:V&gt;Adv</td>
<td>bitingly</td>
</tr>
<tr>
<td>y:N&gt;Adj</td>
<td>mousey</td>
</tr>
</tbody>
</table>
Productivity and The Proportion of Low-Frequency Forms

• The *Inference-Based* (FG) makes correct use of distributional facts about productivity.

• Why is the model sensitive to the proportion of low-frequency forms?
Ideal Input for Productivity

- busy-ness
- cheap-ness
- good-ness
- grand-ness
Ideal Input for Productivity
Ideal Input for Storage

```
N             N             N             N
  Adj -th     Adj -th     Adj -th     Adj -th
    |             |             |             |
   true       true       true       true
```
Ideal Input for Storage
Productivity and the Proportion of Low-Frequency Forms
Productivity and the Proportion of Low-Frequency Forms

- busy-ness
- busy-ness
- good-ness
- grand-ness
Productivity and the Proportion of Low-Frequency Forms
Productivity and the Proportion of Low-Frequency Forms

N
-ness

Adj
busy

N
-ness

Adj
busy

N
-ness

Adj
busy

N
-ness

Adj
grand
Productivity and the Proportion of Low-Frequency Forms
Baayen’s Hapax-Based Estimators

- Baayen’s $\frac{P}{P^*}$ (e.g., Baayen, 1992)
- Most well-studied statistical corpus-oriented estimators of productivity.
- Estimators of *single-affix* productivity based on the proportion hapax.
- Various mathematical derivations.
  - Rate of vocabulary change.
  - Good-Turing.
- Non-psychological.
The *Inference-Based* Model and Baayen’s Measures

- For individual suffixes predictions strongly correlate.

<table>
<thead>
<tr>
<th>Measure</th>
<th>FG</th>
<th>MDPCFG</th>
<th>MAG</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{P}$</td>
<td>0.907</td>
<td>-0.003</td>
<td>0.692</td>
</tr>
<tr>
<td>$\mathcal{P}^*$</td>
<td>0.662</td>
<td>0.480</td>
<td>0.568</td>
</tr>
</tbody>
</table>

- *Inference-Based (FG)* model makes no explicit assumption about low frequency forms.

- *Inference-Based (FG)* model derives relationship between productivity inferences and distributional facts from storage-computation tradeoff applied to the problem of lexicon learning.
Morphological Productivity

1. Background
2. Productivity and frequency
3. Affix ordering
The Affix-Ordering Problem

- Tiny fraction of logically possible affix combinations occur in practice.

- Fabb (1988)
  - 43 suffixes
  - 663 predicted possible.
  - ~50 attested.
The Affix-Ordering Problem

• Numerous accounts (see, O’Donnell, 2015 for review).
  • None entirely successful.

• Main empirical generalization: Ordering correlates with a number of other properties.
Correlated Properties

Earlier Suffixes
- ion, -ity, -y, -al, -ic,
- ate, -ous, -ive, ...

Later Suffixes
- ness, -less, -hood, -ful,
- ly, -y, -like, -ist, ...

1. Origin (Latin/Germanic)
2. Phonological regularity
3. Transparency of meaning
4. Productivity
Productivity and Ordering
Generalization

On average more productive affixes appear outside of less productive affixes
(Hay, 2002; Hay & Plag, 2004; Plag et al., 2009).
Productivity and Ordering Generalization

• Follows as a consequence of *Inference-Based* (FG) model’s pattern of storage and computation.

• **Crucial point:** By definition, productive affixes are just independent units which can combine freely; unproductive affixes are subparts of other stored forms.
Productivity and Ordering

Simple words

Words with unproductive suffixes
Productivity and Ordering
Correlated Properties

• Latinate v. Germanic reflects diachronic fact that Latinate affixes were borrowed as parts of loan words.

• Regularity and Transparency differences are accounted for if stored items can preserve or accrue idiosyncrasies.
Paradoxical Suffix Combinations

• Exceptions to the productivity and ordering generalization and all current theoretical accounts (Hay, 2003).

  • -ability, -ation, -istic

• Idea: -ability, -ation, and -istic are single, stored lexical units.
## Predicted Generalizable Units

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>-ate -ion</td>
<td>N</td>
</tr>
<tr>
<td>-ic -al</td>
<td>A</td>
</tr>
<tr>
<td>-ate -ive</td>
<td>A</td>
</tr>
<tr>
<td>-al -ity</td>
<td>N</td>
</tr>
<tr>
<td>-al -ize</td>
<td>V</td>
</tr>
<tr>
<td>-ology -ist</td>
<td>N</td>
</tr>
<tr>
<td>-ment -al</td>
<td>A</td>
</tr>
<tr>
<td>-able -ity</td>
<td>N</td>
</tr>
<tr>
<td>-ist -ic</td>
<td>A</td>
</tr>
<tr>
<td>-ous -ity</td>
<td>N</td>
</tr>
</tbody>
</table>
Experimental Evidence

(Aronoff & Schvaneveldt, 1978; Anshen and Aronoff, 1981)

Suffix 1

novel stem + -ive v. -(a)ble

Suffix 2

+ -ity v. -ness

depulsive

(Aronoff & Schvaneveldt, 1978)

depulsivity depulsiveness

remortible

(Anshen & Aronoff, 1981)

remortibility remortibleness
-ivity v. -bility

Predicted

-ive
-ble

-ness

-ity
-ivity v. -bility

Preference for -ness

-ive

-ble
-ivity v. -bility

Preference for -ity
-ivity v. -bility

-ive
-ble

-ness

-ity

Predicted

Preceding suffix -ive
-ivity v. -bility

{ -ive
  -ble
}

Preceding suffix -ble
Full-Parsing (DMPCFG)

-ive
-ble

-ness

-ity

Predicted

Full-Parsing (DMPCFG)

ble
ive

ble
ive
Inference-Based (FG)

Predicted

Full-Parsing (DMPCFG)

Full-Storage (MAG)

Inference (FG)

-ive

-ble

-ness

-ity
Overview

Part I: An approach to lexical uncertainty.

Part II: Detailed case study in morphology.

Part III: Unsupervised models of language learning.
Morphology Studies

Phonological, morphosyntactic, and semantic processes generate candidate analyses.

N

Adj -able

V -ity

agree
Morphology Studies

Phonological, morphosyntactic, and semantic processes generate candidate analyses.

\[
\begin{align*}
\text{N} & \quad \text{-ity} \\
\text{Adj} & \quad \text{-able} \\
\text{V} & \quad \text{agree}
\end{align*}
\]
Unsupervised Learning

Phonological, morphosyntactic, and semantic processes generate candidate analyses.

Unsupervised Learning

N

Adj

V

agree

-able

-ity
Unsupervised Learning

Phonological, morphosyntactic, and semantic processes generate candidate analyses.

No Productivity Inference

agree -able -ity
Unsupervised Lexicon Discovery from Speech
(Lee, O’Donnell, & Glass, 2015)

• First step in integrating models of phonetic, phonological, and morphological structure learning.

• Completely unsupervised learning of words, and morphemes from speech (acoustic input).

• Uses similar storage-computation tradeoffs in each component.
Unsupervised Lexicon Discovery from Speech
(Lee, O’Donnell, & Glass, 2015)

1. Model of unsupervised learning of phonological units from acoustic data.

2. Model of mapping between underlying (lexical) to surface phonological units.

3. Model of morphological structure.

Unsupervised Lexicon Discovery from Speech
(Lee, O’Donnell, & Glass, 2015)
Unsupervised Lexicon Discovery from Speech
(Lee, O’Donnell, & Glass, 2015)
1. Learning Phonemes (Lee & Glass, 2012)

- Phones are hidden Markov models generating acoustics.
- Unbounded number of phones.
- Dirichlet process prior over phone inventory.
- Fewer more reusable phones.
Unsupervised Lexicon Discovery from Speech
(Lee, O’Donnell, & Glass, 2015)

I. Learning Phonemes (Lee & Glass, 2012)

and MIT’s open university
Unsupervised Lexicon Discovery from Speech
(Lee, O’Donnell, & Glass, 2015)

1. Learning Phonemes (Lee & Glass, 2012)

banana

/b/ /æ/ /n/ /æ/ /n/ /ə/

and MIT’s open university
Unsupervised Lexicon Discovery from Speech
(Lee, O’Donnell, & Glass, 2015)
Unsupervised Lexicon Discovery from Speech
(Lee, O’Donnell, & Glass, 2015)

2. Underlying and Surface Phones

- Highly simplified phonetics/phonology.
- Allophony, phonetic variation, coarticulation, noise, etc.
Unsupervised Lexicon Discovery from Speech
(Lee, O’Donnell, & Glass, 2015)

2. Underlying and Surface Phones

• Highly simplified phonetics/phonology.
• Substitutions.
• Splits.
• Deletions.
Unsupervised Lexicon Discovery from Speech
(Lee, O’Donnell, & Glass, 2015)
Unsupervised Lexicon Discovery from Speech
(Lee, O’Donnell, & Glass, 2015)

3. Model of Morphological Structure

• Utterances consist of words.
• Words consist of sub-words.
• Sub-words consist of phonemes.
Unsupervised Lexicon Discovery from Speech
(Lee, O’Donnell, & Glass, 2015)

3. Model of Morphological Structure

\[
\begin{align*}
U & \rightarrow W^* \\
W & \rightarrow Sw^* \\
Sw & \rightarrow P^* \\
P & \rightarrow l \\
\end{align*}
\]
Unsupervised Lexicon Discovery from Speech
(Lee, O’Donnell, & Glass, 2015)

and MIT's open university
Unsupervised Lexicon Discovery from Speech
(Lee, O’Donnell, & Glass, 2015)
Unsupervised Lexicon Discovery from Speech

(Lee, O’Donnell, & Glass, 2015)
Unsupervised Lexicon Discovery from Speech
(Lee, O’Donnell, & Glass, 2015)
Unsupervised Lexicon Discovery from Speech
(Lee, O’Donnell, & Glass, 2015)

4. Model of Lexical Storage

• Full-Storage Model (Adaptor Grammars; Johnson, et al., 2007)
  • First pass don’t infer productivity.
  • Store all sub-word and word units.
Unsupervised Lexicon Discovery from Speech
(Lee, O’Donnell, & Glass, 2015)
Unsupervised Lexicon Discovery from Speech

(Lee, O’Donnell, & Glass, 2015)
Unsupervised Lexicon Discovery from Speech

(Lee, O’Donnell, & Glass, 2015)
Unsupervised Lexicon Discovery from Speech
(Lee, O’Donnell, & Glass, 2015)
Key Scientific Results

- First unsupervised model of this scope.


- Phone segmentation ~75%, word segmentation less great <20%.

- Lexical units.

- Synergistic Interactions.
Key Scientific Results

- Lexical units.
- Synergistic Interactions.
Lexical Unit Study

<1.5 Hours of Speech

U → W*
W → Sw*
Sw → P*
P → l
## Lexical Units

<table>
<thead>
<tr>
<th>Transcription</th>
<th>Discovered lexical units</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>/iy l iy/ (really, willy, billion)</td>
<td>[35] [31 4]</td>
<td>68</td>
</tr>
<tr>
<td>/ey sh ax n/ (innovation, imagination)</td>
<td>[67 30] [49]</td>
<td>43</td>
</tr>
<tr>
<td>/ax bcl ax l/ (able, cable, incredible)</td>
<td>[34 18] [38 91]</td>
<td>18</td>
</tr>
<tr>
<td>discovered</td>
<td>[26] [70 110 3] [9 99] [31]</td>
<td>9</td>
</tr>
<tr>
<td>individual</td>
<td>[49 146] [34 99] [154] [54 7] [35 48]</td>
<td>7</td>
</tr>
<tr>
<td>powerful</td>
<td>[50 57 145] [145] [81 39 38]</td>
<td>5</td>
</tr>
<tr>
<td>open university</td>
<td>[48 91] [4 67] [25 8 99 29] [44 22] [103 4]</td>
<td>4</td>
</tr>
<tr>
<td>the arab muslim world</td>
<td>[28 32] [41] [67] [25 35] [1 27] [13 173] [8 139] [38 91]</td>
<td>2</td>
</tr>
</tbody>
</table>
## Lexical Units

<table>
<thead>
<tr>
<th>Transcription</th>
<th>Discovered lexical units</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ɪ-liː/ (really, willy, billion)</td>
<td>[35] [31 4]</td>
<td>68</td>
</tr>
<tr>
<td>/eɪˈʃeɪ.zən/ (innovation, imagination)</td>
<td>[6 7 30] [49]</td>
<td>43</td>
</tr>
<tr>
<td>/ˈæbl əbl/ (able, cable, incredible)</td>
<td>[34 18] [38 91]</td>
<td>18</td>
</tr>
<tr>
<td>discovered</td>
<td>[26] [70 110 3] [9 99] [31]</td>
<td>9</td>
</tr>
<tr>
<td>individual</td>
<td>[49 146] [34 99] [154] [54 7] [35 48]</td>
<td>7</td>
</tr>
<tr>
<td>powerful</td>
<td>[50 57 145] [145] [81 39 38]</td>
<td>5</td>
</tr>
<tr>
<td>open university</td>
<td>[48 91] [4 67] [25 8 99 29] [44 22] [103 4]</td>
<td>4</td>
</tr>
<tr>
<td>the arab muslim world</td>
<td>[28 32] [41] [67] [25 35] [1 27] [13 173] [8 139] [38 91]</td>
<td>2</td>
</tr>
</tbody>
</table>
## Predicted Generalizable Combinations

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>-ate -ion</td>
<td>N</td>
</tr>
<tr>
<td>-ic -al</td>
<td>A</td>
</tr>
<tr>
<td>-ate -ive</td>
<td>A</td>
</tr>
<tr>
<td>-al -ity</td>
<td>N</td>
</tr>
<tr>
<td>-al -ize</td>
<td>V</td>
</tr>
<tr>
<td>-ology -ist</td>
<td>N</td>
</tr>
<tr>
<td>-ment -al</td>
<td>A</td>
</tr>
<tr>
<td>-able -ity</td>
<td>N</td>
</tr>
<tr>
<td>-ist -ic</td>
<td>A</td>
</tr>
<tr>
<td>-ous -ity</td>
<td>N</td>
</tr>
</tbody>
</table>
Lexical Units

<table>
<thead>
<tr>
<th>Transcription</th>
<th>Discovered lexical units</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>/iy 1 iy/ (really, willy, billion)</td>
<td>[35] [31 4]</td>
<td>68</td>
</tr>
<tr>
<td>/ey sh ax n/ (innovation, imagination)</td>
<td>[6 7 30] [49]</td>
<td>43</td>
</tr>
<tr>
<td>/ax bcl ax 1/ (able, cable, incredible)</td>
<td>[34 18] [38 91]</td>
<td>18</td>
</tr>
<tr>
<td>discovered</td>
<td>[26] [70 110 3] [9 99] [31]</td>
<td>9</td>
</tr>
<tr>
<td>individual</td>
<td>[49 146] [34 99] [154] [54 7] [35 48]</td>
<td>7</td>
</tr>
<tr>
<td>powerful</td>
<td>[50 57 145] [145] [81 39 38]</td>
<td>5</td>
</tr>
<tr>
<td>open university</td>
<td>[48 91] [4 67] [25 8 99 29] [44 22] [103 4]</td>
<td>4</td>
</tr>
<tr>
<td>the arab muslim world</td>
<td>[28 32] [41] [67] [25 35] [1 27] [13 173] [8 139] [38 91]</td>
<td>2</td>
</tr>
</tbody>
</table>
Lexical Units

<table>
<thead>
<tr>
<th>Transcription</th>
<th>Discovered lexical units</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>/iy l iy/ (really, willy, billion)</td>
<td>[35] [31 4]</td>
<td>68</td>
</tr>
<tr>
<td>/ey sh ax n/ (innovation, imagination)</td>
<td>[6 7 30] [49]</td>
<td>43</td>
</tr>
<tr>
<td>/ax bcl ax l/ (able, cable, incredible)</td>
<td>[34 18] [38 91]</td>
<td>18</td>
</tr>
<tr>
<td>discovered</td>
<td>[26] [70 110 3] [9 99] [31]</td>
<td>9</td>
</tr>
<tr>
<td>individual</td>
<td>[49 146] [34 99] [154] [54 7] [35 48]</td>
<td>7</td>
</tr>
<tr>
<td>powerful</td>
<td>[50 57 145] [115] [81 33 28]</td>
<td>5</td>
</tr>
<tr>
<td>open university</td>
<td>[48 91] [48 91]</td>
<td>4</td>
</tr>
<tr>
<td>the arab muslim world</td>
<td>[28 32] [41] [67] [25 35] [127] [13 173] [8 139] [38 91]</td>
<td>2</td>
</tr>
</tbody>
</table>

*The arab muslim world.*
Key Scientific Results

- Lexical units.
- Synergistic Interactions.
Synergistic Interactions

Integrating multiple components often makes learning better. Synergistic interactions of system mutually constrain one another in ways that improve linguistic structures they learn. These structures are more accurate than if they were learned independently.

- **Lexical units and Phones:** Top-down influence of lexical unit learning improves phone unit learning.

- **Underlying-to-Surface Mapping and Words:** Modeling underlying-to-surface mapping improves ability to find correct words and sub-words.
Lesion Study

- **Lesioned** model components to study importance of difference components.

- Examined top 20 most important words in each lecture (term frequency inverse-document frequency)
Baseline term detection system.

Park & Glass, 2008
Full Model

- # discovered keywords
- Lecture topics: Economics, Signal processing, Clustering, Speaker adaptation, Physics, Linear algebra
- Park & Glass, 2008 vs. Full model
Fixed Phones
Joint Model

- Park & Glass, 2008
- Full model

Fixed phones

Top-down influence on phone inferences.

Lecture topics

- Economics
- Signal processing
- Clustering
- Speaker adaptation
- Physics
- Linear algebra

# discovered keywords
No Noisy Channel
No Noisy-Channel

- Park & Glass, 2008
- Full model
- Fixed Phones
- No noisy-channel

Regularizing phone variations is critical

Lecture topics

# discovered keywords

- Economics
- Signal processing
- Clustering
- Speaker adaptation
- Physics
- Linear algebra
Overall Conclusion

- Can use tradeoff-based approach to learn lexical units.
- Makes fine-grained predictions for linguistics and psychology.
- Can scale up to more unsupervised settings.
Thanks