A complex-systems view on language (text analysis)

Eduardo G. Altmann

School of Mathematics and Statistics
The University of Sydney
Australia
The faculty of Language: What Is It, Who Has It, and How Did It Evolve
Hauser, Chomsky, Fitch (Science 2002)

The Evolution of Universal Grammar
Nowak, Komarova, Niyogi (Science 2001)

\[ \dot{x}_i = \sum_{j=1}^{n} x_j f_j Q_{ji} - \phi x_i \quad i = 1, \ldots, n \]

The Mystery of Language Evolution, Hauser et al. (Frontiers in Psychology 2014)

“We argue instead that the richness of ideas is accompanied by a poverty of evidence...”
Human language as a culturally transmitted replicator
Pagel (Nature Rev. Genetics 2009)

Modelling the dynamics of language death

\[
\frac{dx}{dt} = yP_{yx}(x,s) - xP_{xy}(x,s)
\]

\[
P_{yx}(x,s) = cx^{a}s \quad \text{and} \quad P_{xy}(x,s) = c(1-x)^{a}(1-s)
\]

The origin and evolution of word order
Gell-Mann and Ruhlen (PNAS 2011)
In the evening the yawl return'd from fishing having caught two Sting rays weighing near 600 pounds. The great quantity of New Plants & Ca Mr Banks & Dr Solander collected in this place occasioned my giving it the name of Botany Bay. It is situated in the Latitude of 34°.0' S° Longitude 20°.8°.37' W° it is Capacious safe and commodious - it may be known by the land on the Sea-coast which is of a pretty even and moderate height and rather higher than it is farther inland with steep rocky cliffs next the Sea and looks like a long Island lying close under the Shore: the entrance of the harbour lies about the Middle of this land - in coming from the Southward it is discover'd before you are abreast of it which you cannot do in coming from the northward...


Cook's Diary
Sunday 6th May 1770

Utterance selection model of language change
Baxter, Blythe, Croft, McKane (Phys Rev E 2006)

Quantifying the evolutionary dynamics of language
Lieberman, Michel, Jackson, Tang, Nowak (Nature 2007)
Universal statistical laws?

War and Peace, by Leo Tolstoy

Well, Prince, so Genoa and Lucca are now just family estates of the Buonapartes. But I warn you, if you don't tell me that this means war, if you still try to defend the infamies and horrors perpetrated by that Antichrist—I really believe he is Antichrist--

...
Language Dynamics

Humans 100,000 years
Languages 10,000 years
Change 1,000 years

100 years 10 years

One book Twitt

Google n-gram corpus: 500y, >1M Books
English Wikipedia: 10y, 4M Articles
Scientific Papers in WoS: 30y, 2M Articles

Usenet Discussion group: ~30y, 5M posts/group
Twitter: ~10y, 750M tweets/day

Data is available!

Complex Systems

Applications (e.g., data mining)

Models / Methods
Counting Words

Quantifying Language Change

Mining Texts With Networks

**Problem:** dependence of vocabulary on database size?

**Report on the state of the German language** (March 2013)
German Academy for Language and Literature
Union of German Academies of Sciences and Humanities

<table>
<thead>
<tr>
<th>Year</th>
<th>1905-1914</th>
<th>1948-1957</th>
<th>1995-2004</th>
</tr>
</thead>
<tbody>
<tr>
<td># distinct words</td>
<td>3,715,000</td>
<td>5,045,000</td>
<td>5,238,000</td>
</tr>
</tbody>
</table>

**Quantitative Analysis of Culture Using Millions of Digitized Books**
Michel et. al., Science (2011) *English*

<table>
<thead>
<tr>
<th>Year</th>
<th>1900</th>
<th>1950</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td># distinct words</td>
<td>544,000</td>
<td>597,000</td>
<td>1,022,000</td>
</tr>
</tbody>
</table>
Example of applications:
- invert indexing (document classification, text mining, etc.)
- vocabulary richness of texts / authors (different document lengths)
Vocabulary growth with database size

Limit vocabulary?

- 5 M books from \[1520, 2000\]
- 100 B words

Vocabulary size \(N\) vs. database size \(M\)

- English
- German
- French
- Spanish
- Russian

\[N = M\] Heaps law

\[\Delta\] Wikipedia
\[\bullet\] Cumulative
\[\times\times\times\] Yearly data
Vocabulary growth with database size

**Simple model**: usage of each word follows a Poisson process with fixed frequency

\[
\langle N(M) \rangle = \sum_{r} 1 - e^{-F(r)M}
\]

where \(F(r)\) is the frequency of the \(r\)-th most frequent word (\(r = \text{rank}\)).

Zipf’s law?

**Rank-frequency distribution**

*rank (r-th most frequent word)*
Zipf’s law?

<table>
<thead>
<tr>
<th>language</th>
<th>$b^*$</th>
<th>$\gamma^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>7,873</td>
<td>1.77</td>
</tr>
<tr>
<td>French</td>
<td>8,208</td>
<td>1.78</td>
</tr>
<tr>
<td>Spanish</td>
<td>8,757</td>
<td>1.78</td>
</tr>
<tr>
<td>German</td>
<td>19,863</td>
<td>1.62</td>
</tr>
<tr>
<td>Russian</td>
<td>62,238</td>
<td>1.94</td>
</tr>
</tbody>
</table>

$F_{dp}(r; \gamma, b) = \begin{cases} 
  r^{-1}, & r \leq b, \\
  r^{-\gamma}, & r > b 
\end{cases}$
**Simple mode**: usage of each word follows a Poisson process with fixed frequency

$$\langle N(M) \rangle = \sum_r 1 - e^{-F(r)} M$$

where $F(r)$ is the frequency of the $r$-th most frequent word ($r = \text{rank}$).

$$F_{dp}(r; \gamma, b) = \begin{cases} r^{-1}, & r \leq b, \\ r^{-\gamma}, & r > b \end{cases}$$

$$N_{dp}(N_c) = \begin{cases} M, & M \ll M_b, \\ M^{1/\gamma}, & M \gg M_b \end{cases}$$

Extension of the Zipf-Heaps connection [<Mandelbrot 1950’s]!
Vocabulary growth with database size

\[ N(M_b) = 7873 \]

\[ N(M) \sim \begin{cases} M & M \ll M_b, \\ M^{1/(\alpha+1)} & M \gg M_b \end{cases} \]
What is changing?
$f(t, \Delta t)$: fraction of core words at time $t$ which remain core at time $t + \Delta t$
Change in the core vocabulary

Replacement in the core vocabulary:
$N_c/\kappa \approx 30$ words/year

Accelerating in time!

1900
majesty, doubtless, furnished, monsieur, Napoleon, hitherto

Most frequent replaced words

2000
cultural, context, technology, programs, environmental, computer

1900

2000
Jensen-Shanon Divergence, $D_{\alpha}$

$$H_\alpha(H(p)) = -\sum_{i} p_i \log p_i$$

$$H_{\alpha=1}(p) = H(p)$$

Havrda&Chrvát, Kybernetika (1967)

$$D_{\alpha}(p, q) = H\left(\frac{p + q}{2}\right) - \frac{1}{2} H(p) - \frac{1}{2} H(q)$$

$$D_{\alpha=1}(p, q) = D(p, q)$$


$$-\sqrt{D_{\alpha}}$$ is a metric for $\alpha \in [0, 2]$


Slow convergence of statistical estimators due to Zipf’s law: $F_r \sim r^{-\gamma}$

<table>
<thead>
<tr>
<th>$H_\alpha$</th>
<th>$D_{\alpha}$</th>
<th>$\tilde{D}_{\alpha}(p \neq q)$</th>
<th>$\tilde{D}_{\alpha}(p = q)$</th>
<th>$V^{(\alpha)} / N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias:</td>
<td>$V^{(\alpha)} / N$</td>
<td>$V^{(\alpha)} / N$</td>
<td>$V^{(\alpha)} / N$</td>
<td>$V^{(\alpha)} / N$</td>
</tr>
<tr>
<td>Fluctuations:</td>
<td>$V^{(2\alpha)} / N$</td>
<td>$V^{(2\alpha)} / N$</td>
<td>$V^{(2\alpha-1)} / N^2$</td>
<td>$V^{(\alpha)} \propto \left{ \begin{align*} N^{-\alpha+1+1/\gamma} &amp; \quad \alpha &lt; 1 + 1/\gamma \ constant &amp; \quad \alpha &gt; 1 + 1/\gamma \end{align*} \right.$</td>
</tr>
</tbody>
</table>
Change of English
(Google n-gram database 1520-2010)
Change of English
(Google n-gram database 1520-2010)
Similarity of Scientific Disciplines
(title and abstract of all Web of Science papers 1990-2014)

<table>
<thead>
<tr>
<th>DOMAINS</th>
<th>DISCIPLINES</th>
<th>SPECIALTIES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural Sciences</td>
<td>Computer sciences</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Physical sciences</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chemical sciences</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Earth sciences</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Biological sciences</td>
<td></td>
</tr>
<tr>
<td>Engineering</td>
<td>Electrical eng.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Materials eng.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Medical eng.</td>
<td></td>
</tr>
<tr>
<td>Medical Sciences</td>
<td>Basic medicine</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Clinical medicine</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Health sciences</td>
<td></td>
</tr>
<tr>
<td>Agriculture</td>
<td>Veterinary science</td>
<td></td>
</tr>
<tr>
<td>Social Sciences</td>
<td>Psychology</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Economics and business</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sociology</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Soc. and econ. geography</td>
<td></td>
</tr>
<tr>
<td>Humanities</td>
<td>Lang. and literature</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Arts</td>
<td></td>
</tr>
</tbody>
</table>

Kendall Correlation between WoS and text

\[ \alpha_{max} = 1.40 \]
\[ \langle \Delta \tilde{D}_{\alpha}^{(i,j)} = 2 \rangle_{i,j} \approx 0 \] 

Not significantly different from zero  
(T-test, \( p = 0.056 \); Wilcoxon test \( p = 0.17 \))
“The progress of language change through a community follows a lawful course, an S-curve from minority to majority to totality.”


What is the empirical support?

“...up to a dozen points for a single change”


- Are all changes following S-curves? No!
- Are all S-curves the same? No!
- Can we extract from S-curves information about the process of change? Yes!
Adoption of new words

Ortography reform (1996): \( \beta \rightarrow ss \)

2,000 different words (e.g., Kongreß \( \rightarrow \) Kongress)
Adoption of new words

\[ \frac{d\rho(t)}{dt} = (a + b \rho(t)) (1 - \rho(t)) \]

\[ \begin{cases} 
  b = 0 \Rightarrow \rho(t) = \text{exponential} \\
  a = 0 \Rightarrow \rho(t) = \text{symmetric S-curve} 
\end{cases} \]
Text mining

A
UNIVERSALIS
DE JURE
HOMINUM
DECLARATIO

B
Cum dignitatis
infixae
omnibus
humanae

C
Cum dignitatis
infixae
omnibus
humanae

D
familiae
partibus et
eorum jurum
aequalium,

Documents

|   | A  | B  | C  | D  | ...
|---|----|----|----|----|-----
| the of science sport networks physics biology | 156 | 85 | 111 | 35 | 56 |
| | 59 | 65 | 75 | 33 | 40 |
| | ... | ... | ... | ... | ... |
| | 0 | 5 | 2 | 0 | 0 |
| | 4 | 0 | 0 | 0 | 0 |
| | 2 | 0 | 0 | 0 | 0 |
| | 0 | 0 | 1 | 0 | 0 |
| | 0 | 0 | 0 | 5 | 0 |
| | ... | ... | ... | ... | ... |
Latent Dirichlet Allocation (LDA)

Blei, Ng, Jordan (Journal of Machine Learning 2003), >20k citations
Implementation: McCallum's MALLET (http://mallet.cs.umass.edu)

- Fixed number of topics K
- Dirichlet Priors
- Inference problem:

\[
P(Model|Data) = P(Data|Model) \frac{P(Model)}{P(Data)}
\]

\[
\begin{align*}
\text{Data} &= A_{\omega,d} \\
\text{Model} &= \{\varphi_{j,w}, \theta_{d,j}\}
\end{align*}
\]

\[
P(Model) = Prior = \left\{ \begin{array}{c}
\varphi_{j,w} \sim Dir(\beta) \\
\theta_{d,j} \sim Dir(\alpha)
\end{array} \right. \]
Stochastic Block Models (SBM)
Holland, Laskey, Leinhardt (Social Networks 1983)

- Probability of connection between nodes depends on the blocks they belong
- Number of Blocks << Number of nodes (links)

Generative model: non-parametric hierarchical SBM

- number of blocks (topics) not fixed
- prior at one level is set by the upper hierarchy level
- each link (word token in a document) is assigned to a pair of blocks
Which model compacts better the data in terms of coding or description length (DL)?

Grünwald (*The Minimum Description Length Principle*, 2007)

\[
\Sigma = DL(\text{data}|\text{model}) + DL(\text{model})
\]

Minimum description length (MDL) for probabilistic models:

\[
\hat{\Sigma} = - \log P(D|\hat{\theta}) - \log P(\hat{\theta})
\]

- D = data
- \( \theta \) = discrete parameters of the model

\[
\hat{\theta} = \arg \max_{\theta} P(D|\theta) P(\theta)
\]

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Docs.</th>
<th>Words</th>
<th>Word Tokens</th>
<th>( \Sigma_{\text{LDA}} ) (hyperf)</th>
<th>( \Sigma_{\text{hSBM}} )</th>
<th>hSBM groups</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>50</td>
<td>100</td>
<td>500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Twitter</td>
<td>10,000</td>
<td>12,258</td>
<td>196,625</td>
<td>1,140,357 1,110,186 1,091,998 1,056,321</td>
<td>963,260</td>
<td>365 359</td>
</tr>
<tr>
<td>Reuters</td>
<td>1,000</td>
<td>8,692</td>
<td>117,661</td>
<td>879,684 876,656 881,107 879,321</td>
<td>341,199</td>
<td>54 55</td>
</tr>
<tr>
<td>Web of Science</td>
<td>1,000</td>
<td>11,198</td>
<td>126,313</td>
<td>1,035,555 1,057,491 1,065,584 1,075,433</td>
<td>426,529</td>
<td>16 18</td>
</tr>
<tr>
<td>New York Times</td>
<td>1,000</td>
<td>32,415</td>
<td>335,749</td>
<td>2,701,001 2,699,711 2,695,955 2,693,749</td>
<td>1,414,631</td>
<td>124 125</td>
</tr>
<tr>
<td>PlosONE</td>
<td>1,000</td>
<td>68,188</td>
<td>5,172,908</td>
<td>9,782,605 49,497,904 49,326,867 48,741,824</td>
<td>8,475,866</td>
<td>897 972</td>
</tr>
</tbody>
</table>
LDA generated documents:
10 topics, 1M documents, following Heaps’ and Zipf’s laws
Conclusions

Thank you for your attention!


Data is available!

Language as a Complex System

Applications (e.g., data mining)

Models / Methods

Thank you for your attention!
