Word distance distribution in literary texts

Gemma Boleda\(^1\)  Alvaro Corral\(^2\)
Ramon Ferrer i Cancho\(^1\)  Albert Díaz-Guilera\(^3\)

Universitat Politècnica de Catalunya\(^1\)
Centre de Recerca Matemàtica\(^2\)
Universitat de Barcelona\(^3\)

Corpus Linguistics Conference, Liverpool.
20-23rd July 2009
Goal and motivation

- we are interested in modeling dynamic properties of language
  - as opposed to, e.g., static properties captured by Zipf’s law
- the distribution of **distances between word repetitions**
  - distance = number of words + 1

Example

I have a rabbit. My rabbit eats lettuce. Actually, rabbits usually like lettuce. Have you ever encountered a rabbit who doesn’t like lettuce? My rabbit does.

- distances(rabbit) = 2, 4, 8, 6
Goal and motivation

- we are interested in modeling dynamic properties of language
  - as opposed to, e.g., static properties captured by Zipf’s law
- the distribution of distances between word repetitions
  - distance = number of words + 1

Example

I have a rabbit. My rabbit eats lettuce. Actually, rabbits usually like lettuce. Have you ever encountered a rabbit who doesn’t like lettuce? My rabbit does.

- distances(rabbit) = 2, 4, 8, 6
Goal and motivation

- identify patterns in distance distributions
- research questions
  - can we observe universal (language- and frequency-independent) patterns?
  - if so, are they to be traced in . . .
    - rhetorics?
    - communication?
    - brain functioning?
    - language?
    - more general physical phenomena?
  - why? how?
Related work

- initial work
  - Zipf 1945: number of pages between repetitions of low-frequency words in English
  - Herdan 1956: number of words between repetitions of high frequency words in Russian
- Church and Gale 1995
  - number of documents a word appears in
  - if a word is unfrequent its occurrences are concentrated in a few documents (burstiness)
- Lambiotte, Ausloos, and Thelwall (2007)
  - time units between word repetitions in blogs and RSS feeds
- our approach
  - all frequency scales
  - distances in number of words within a single document
    - need large documents: novels
Measuring word distances

- distances $\ell$ for *rabbit*: $\ell = 2, 4, 8, 6$
- huge variation in frequency $\rightarrow$ rescale $\ell$ by the mean:
  - for each word $w$, define $\theta = \ell/\bar{\ell}_w$
  - example: $\bar{\ell}_{rabbit} = 5$
  - $\theta = 0.4, 0.8, 1.6, 1.2$
  - $\rightarrow$ comparison across frequencies

- compute probability density for sets of words $s$ with close frequencies
Data

- 9 novels in 4 languages
  - English, French, Spanish, Finnish
- varying length
- from 80,000 to 2 million words
- processing
  - automatic tokenization, lemmatization, part-of-speech tagging
  - FreeLing, TreeTagger, Conexor’s tools
What we are studying

Texts as sequencies of ranks: *Moby Dick* (about 200,000 words)

*Moby* and *Ishmael* are specially shown
Verbs in *Clarissa*, by S. Richardson (year 1748, 1 million words)

→ Burstiness (attraction) and scaling!
Statistics for word repetitions

Adjectives in *Clarissa*

![Graph showing distribution of adjectives in Clarissa]
Statistics for word repetitions

Adjectives in *Don Quijote*, by M. Cervantes (Spanish, 1605)
Statistics for word repetitions

Verbs in *Kevä ja Takatalvi*, by J. Aho (Finnish, 1906)
A universal exception?

Most nouns and pronouns in *Clarissa*
A universal exception?

Excluded nouns and pronouns: 12 out of 315 ($\theta \gg 1$)

Also proper nouns (not shown)
Pronouns

- automatic grouping of words according to distribution
  - community detection algorithm, graph representation

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>Pronouns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td><em>Moby Dick</em></td>
<td>all, itself, many, none, that, them, themselves, this, those, who, whom</td>
</tr>
<tr>
<td></td>
<td><em>Clarissa</em></td>
<td>it, all, each, itself, many, none, that, themselves, this, those, what, who, whom</td>
</tr>
<tr>
<td>Special</td>
<td><em>Moby Dick</em></td>
<td>he, her, him, himself, I, it, me, myself, she, they, us, we, what, you</td>
</tr>
<tr>
<td></td>
<td><em>Clarissa</em></td>
<td>he, her, herself, him, himself, I, me, myself, ourselves, she, them, they, us, we, you, yourself</td>
</tr>
</tbody>
</table>
Pronouns

- automatic grouping of words according to distribution
- community detection algorithm, graph representation

<table>
<thead>
<tr>
<th></th>
<th>Moby Dick</th>
<th>Clarissa</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Basic</strong></td>
<td>all, itself, many, none, that, them, themselves, this, those, who, whom</td>
<td>it, all, each, itself, many, none, that, themselves, this, those, what, who, whom</td>
</tr>
<tr>
<td><strong>Special</strong></td>
<td>he, her, him, himself, I, it, me, myself, she, they, us, we, what, you</td>
<td>he, her, herself, him, himself, I, me, myself, ourselves, she, them, they, us, we, you, yourself</td>
</tr>
</tbody>
</table>
Burstiness

- we observe burstiness or attraction
- is this trivial?
  - we talk about something until we change topic
  - BUT: also function words!
  - AND: two regimes!
Scaling

- pattern that is independent of word frequency
  - in general, same process for common words and rare words
  - common words are a model for rarer words (useful for language; data sparseness)
  - BUT different process for rhetorically relevant words

- is the function related to neurolinguistic mechanisms?
  - priming, constraints on discourse coherence?
Scaling

- pattern that is independent of word frequency
  - in general, same process for common words and rare words
    → common words are a model for rarer words (useful for language; data sparseness)
  - BUT different process for rhetorically relevant words
- is the function related to neurolinguistic mechanisms?
  - priming, constraints on discourse coherence?
General discussion

- we observe universal patterns in word distance distributions
  - language- and frequency-independent patterns
  - burstiness, repulsion, and scaling
  - burstiness also affects function words!
- hypothesis: they are related to
  - rhetorics: patterns in word repetitions
  - communication: focus on topic causes burstiness
  - brain functioning: activation of a word after utterance (priming)
    → not language-specific?
  - more general physical phenomena? (earthquakes!)
we observe universal patterns in word distance distributions
- language- and frequency-independent patterns
- burstiness, repulsion, and scaling
- burstiness also affects function words!

hypothesis: they are related to
- rhetorics: patterns in word repetitions
- communication: focus on topic causes burstiness
- brain functioning: activation of a word after utterance (priming)
  → not language-specific?
  - more general physical phenomena? (earthquakes!)
we observe universal patterns in word distance distributions
- language- and frequency-independent patterns
- burstiness, repulsion, and scaling
- burstiness also affects function words!

hypothesis: they are related to
- rhetorics: patterns in word repetitions
- communication: focus on topic causes burstiness
- brain functioning: activation of a word after utterance (priming)
  → not language-specific?
- more general physical phenomena? (earthquakes!)
Word distance distribution in literary texts

Gemma Boleda\textsuperscript{1}   Alvaro Corral\textsuperscript{2}
Ramon Ferrer i Cancho\textsuperscript{1}   Albert Díaz-Guilera\textsuperscript{3}

Universitat Politècnica de Catalunya\textsuperscript{1}
Centre de Recerca Matemàtica\textsuperscript{2}
Universitat de Barcelona\textsuperscript{3}

Corpus Linguistics Conference, Liverpool.
20-23rd July 2009
Verbs and earthquakes

Verbs in *Clarissa*
Comparing verbs in *Clarissa* with earthquakes in S. California, 1995-1998

\[ \theta = \frac{\text{time}}{\text{time}} \text{ for earthquakes} \]