# CS 565: Intelligent Systems and Interfaces

Lecture: Words - Collocations

$20^{th}$ – $21^{st}$ Jan, 2016

Semester: Jan - May 2016

Ashish Anand

IIT Guwahati

# Continuing with words ….

- Start up, Stand Up India, New Delhi, Take a break, Strong tea, Stiff breeze

# Understanding Collocation

# Collocations: Definition

- Expressing ourselves with the help of two or more words corresponding to conventional way of saying things or the combination of words which usually go together.

- Examples
  - Strong Tea, Stiff breeze, Take a risk, Start up, New Delhi

# Collocations: Definition

- [Choueka, 1988]: "*A sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact, unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components*“

- Limitation
  - We may do away with the requirement of words being consecutive.

- Example
  - *Knocked* on the *door*
  - *Knocked* on my *door*
  - *Knocked* at the class-room *door*

# Characteristics

- Limited compositionality
  - Example: Strong Tea

An expression is ***compositional*** if its meaning can be predicted from the meaning of the parts.

- Non substitutability
  - Example: *yellow* cannot replace *white* in "*white wine*".
- Non-modifiability: can't be modified using additional lexical materials or through grammatical transformations.
  - Example: *people as poor as church mice; to get an ugly frog in one's throat.*

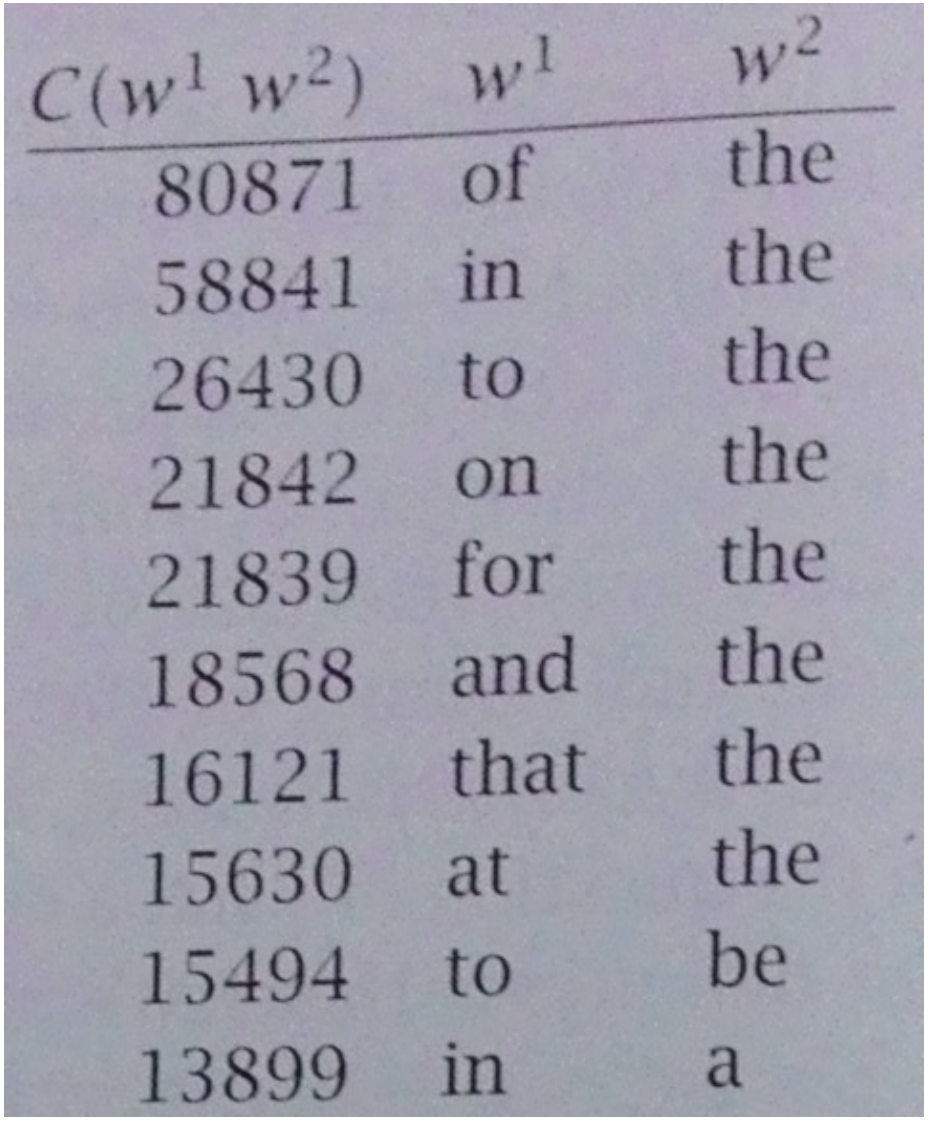
# Why it is important?

- Computational lexicography
- Parsing
- Machine Translation

# Finding Collocations

# Frequency

- Assumption: More frequent occurrence of two words together may imply special function or property which can’t be simply explained



| $C(w^1\ w^2)$ | $w^1$ | $w^2$ |
|---|---|---|
| 80871 | of | the |
| 58841 | in | the |
| 26430 | to | the |
| 21842 | on | the |
| 21839 | for | the |
| 18568 | and | the |
| 16121 | that | the |
| 15630 | at | the |
| 15494 | to | be |
| 13899 | in | a |

Frequency based methods for finding collocations

Source: Table 5.1[FSNLP: Page 154]

Corpus: New York Times newswire-Aug to Nov 1990.
Statistics: 115 MB text with roughly 14 million words

# Adding linguistic knowledge to Frequency

| **Tag Pattern** |
| --- |
| A N |
| N N |
| A A N |
| A N N |
| N A N |
| N N N |
| N P N |

1. Part of Speech (PoS) tag patterns for collocation filtering.
2. Patterns were proposed by *Justeson and Katz* (1995).
3. [*A*]djective; [*N*]oun; [*P*]reposition

Source: Table 5.2 [FSNLP: 154]

| $C(w^1\ w^2)$ | $w^1$ | $w^2$ | Tag Pattern |
|---|---|---|---|
| 11487 | New | York | A N |
| 7261 | United | States | A N |
| 5412 | Los | Angeles | N N |
| 3301 | last | year | A N |
| 3191 | Saudi | Arabia | N N |
| 2699 | last | week | A N |
| 2514 | vice | president | A N |
| 2378 | Persian | Gulf | A N |
| 2161 | San | Francisco | N N |
| 2106 | President | Bush | N N |
| 2001 | Middle | East | A N |
| 1942 | Saddam | Hussein | N N |
| 1867 | Soviet | Union | A N |
| 1850 | White | House | A N |
| 1633 | United | Nations | A N |
| 1337 | York | City | N N |
| 1328 | oil | prices | N N |
| 1210 | next | year | A N |
| 1074 | chief | executive | A N |
| 1073 | real | estate | A N |

**Table 5.3** Finding Collocations: Justeson and Katz' part-of-speech filter.

Source: Table 5.3 [FSNLP: Page 155]

# Pros and Cons of Frequency+PoS Filter

- Advantages
  - Simple method

- Disadvantages
  - Too much dependency on hand-designed filter
  - High frequency can be random without any specific meaning
  - Works well for fixed phrases but will not work for cases where variable number of words may exist between two words
  - Example
    - She *knocked* on his *door*
    - They *knocked* at the *door*
    - 100 women *knocked* on Donaldson's *door*
    - a man *knocked* on the metal front *door*

# Sliding window could be savior

*Sentence:*

*man knocked on the front door*

*Bigrams:*

| | | | | |
|---|---|---|---|---|
| *man knocked* | *man on* | *man the* | *man front* | |
| | *knocked on* | *knocked the* | *knocked front* | *knocked door* |
| | | *on the* | *on front* | *on door* |
| | | | *the front* | *the door* |
| | | | | *front door* |

Four word collocational window to capture bigrams at a distance

# Mean and Variance

- Can implicitly take care of varying distance issue
- Method
  - Calculate mean of *offsets* (signed distance) between the two words.

She *knocked* on his *door*
They *knocked* at the *door*
100 women *knocked* on Donaldson's *door*
a man *knocked* on the metal front *door*

  - Mean, $\bar{d}$ = ¼(3 + 3 + 5 + 5)

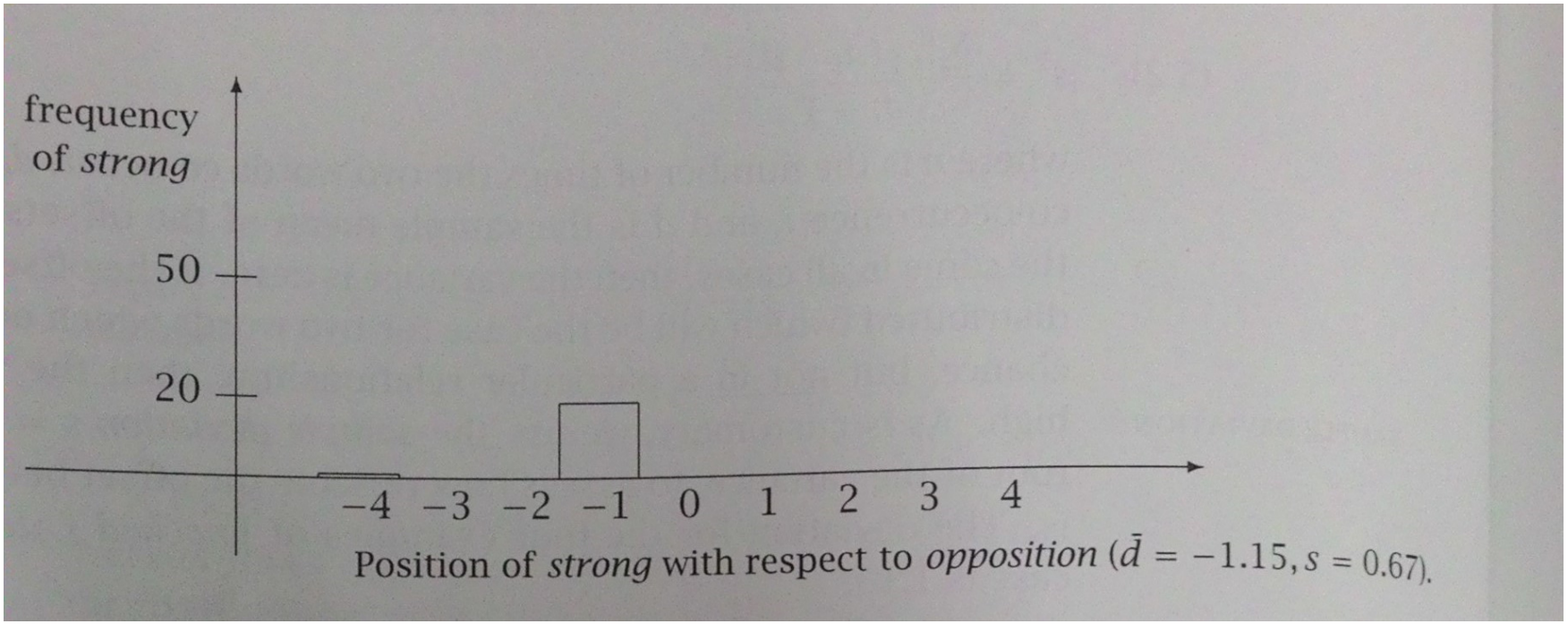
    [Donaldson's tokenized as : Donaldson, apostrophe, s]

  - Variance, $s^2 = \frac{\sum_{i=1}^{n}(d_i - \overline{d})^2}{n-1}$

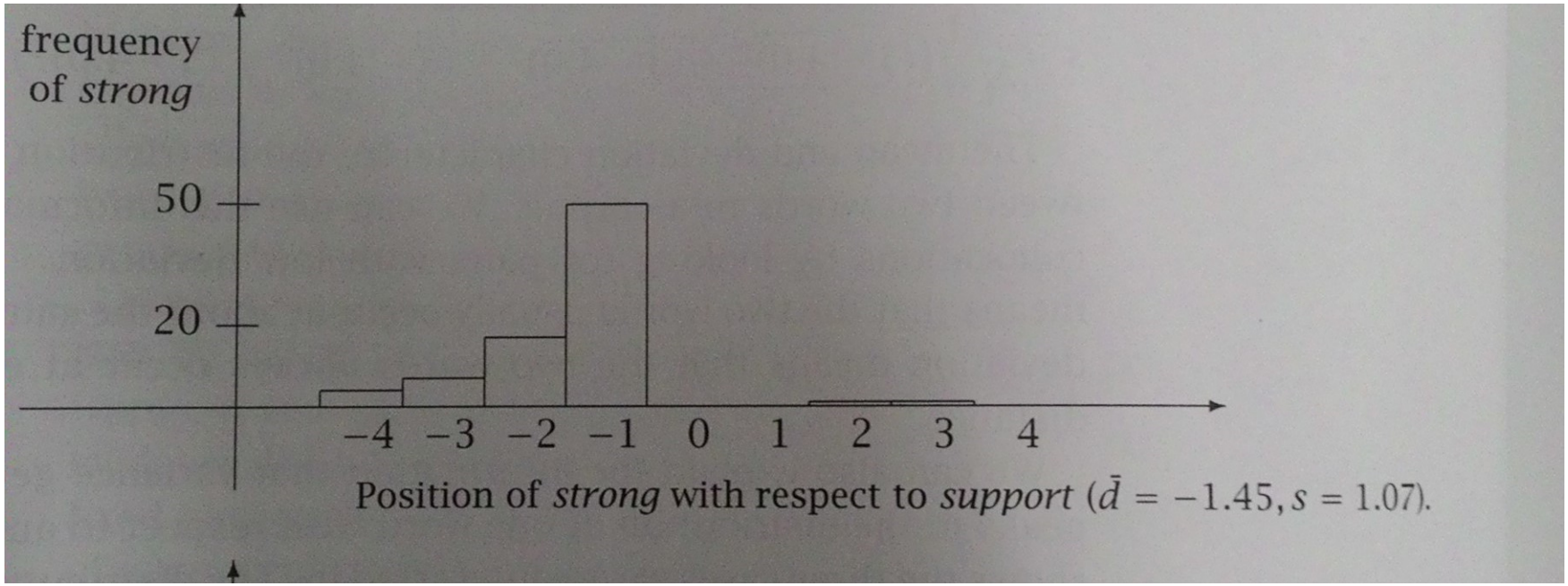
| $s$ | $\bar{d}$ | Count | Word 1 | Word 2 |
|---|---|---|---|---|
| 0.43 | 0.97 | 11657 | New | York |
| 0.48 | 1.83 | 24 | previous | games |
| 0.15 | 2.98 | 46 | minus | points |
| 0.49 | 3.87 | 131 | hundreds | dollars |
| 4.03 | 0.44 | 36 | editorial | Atlanta |
| 4.03 | 0.00 | 78 | ring | New |
| 3.96 | 0.19 | 119 | point | hundredth |
| 3.96 | 0.29 | 106 | subscribers | by |
| 1.07 | 1.45 | 80 | strong | support |
| 1.13 | 2.57 | 7 | powerful | organizations |
| 1.01 | 2.00 | 112 | Richard | Nixon |
| 1.05 | 0.00 | 10 | Garrison | said |

**Table 5.5** Finding collocations based on mean and variance. Sample deviation $s$ and sample mean $\bar{d}$ of the distances between 12 word pairs.

Source: Table 5.5 [FSNLP: page 161]




Source: Figure 5.2 [FSNLP: page 160]




Source: Figure 5.2 [FSNLP: page 160]

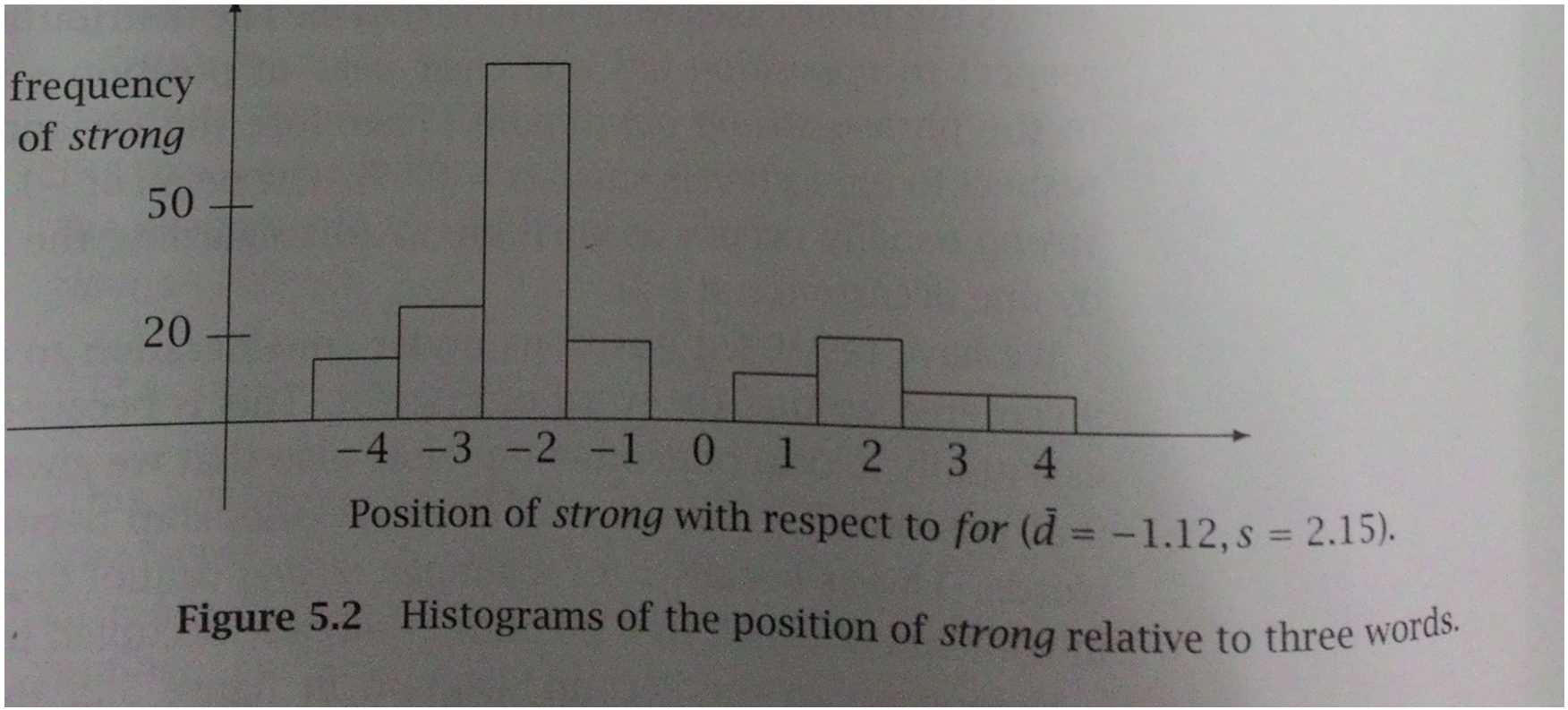



**Figure 5.2** Histograms of the position of *strong* relative to three words.

Source: Figure 5.2 [FSNLP: page 160]

# Reference

- Chapter 5: Until 5.2 FSNLP
- FSNLP: Foundations of Statistical Natural Language Processing, Manning & Schütze