Collocations

- The definition of Collocations
- Frequency
- Mean and Variance

Collocations

“Collocations of a given word are statements of the habitual or customary places of that word”

J.R. Firth

Examples of collocations

	+	-
Adverb + Adjective	*completely satisfied*	*downright satisfied*
Adjective + Noun	*excruciating pain*	*excruciating joy*
Noun + Noun	*a surge of anger*	*a rush of anger*
Noun + Verb	*lions roar*	*lions shout*
Verb + Noun	*commit suicide*	*undertake suicide*
Verb + expression with Preposition	*burst into tears*	*blow up in tears*
Verb + Adverb	*wave frenetically*	*wave feverishly*

Compositionality

Collocations are characterized by limited compositionality.

*Natural language expression is called compositional if the meaning of the expression can be predicted from the meaning of the parts.

Compositionality

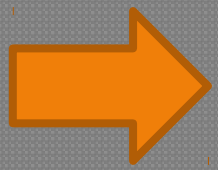
The most extreme examples of non-compositionality are idioms.

- *To hear it through the grapevine*
- *To kick the bucket*
- *To rain cats and dogs*

Most collocations exhibit, however, milder forms of non-compositionality.

Theoretical Approaches

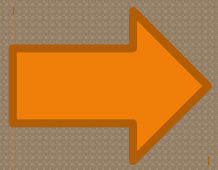
Structural linguistic tradition
(Saussure and Chomsky)



No attention to collocations

General abstraction about phrases and sentences

British linguistics
(Firth, Halliday, Sinclair)



Close attention to collocations

Emphasis on the importance of the context

Contextual Theory of Meaning

Firth’s Contextual Theory of Meaning:

- social setting

(as opposed to the idealized speaker)

- spoken and textual discourse

(as opposed to the isolated sentence)

- surrounding words

(“You shall know a word by the company it keeps" Firth, 1957)

Applications

Collocations are important for a number of applications:

- Natural language generation

(to make sure that the output sounds natural)

- Computational lexicography

(to automatically identify the important collocations to be listed in a dictionary entry)

- Parsing

(to give preference to parses with natural collocations)

- Corpus linguistic research

(to study the social phenomena)

Finding collocations

- Frequency
- Mean and variance
- Hypothesis testing
- Mutual information

Frequency

Main idea

If the two words occur together a lot, then that is evidence that they have special function which results from their combination.

Main problem

Sequences of two adjacent words:

Of the, in the, he said, has been etc. – NOT collocations

Solution

Justeson and Katz' part of speech filter

(to identify likely collocations among frequently occurring word sequences)

Mean and Variance

Main idea

If the pattern of varying distances between two words is relatively predictable, then we have an evidence for a collocation (not necessary a fixed phrase)

Mean and Variance

Fixed phrases ≠ Frequency-based method

Flexible phrases (e.g. *knock*/*hit*/*beat*/*rap* + *door*) ≠ ?

a) *She **knocked** on his **door***
b) *They **knocked** at the **door***
c) *100 woman **knocked** on Donaldson's **door***
d) *A man **knocked** on the metal front **door***

Mean

Compute the mean and variance of the offsets (signed distances)

The mean is the average offset:

a) *She knocked_on_his_door (3)*
b) *They knocked_at_the_door (3)*
c) *100 woman knocked on Donaldson ' s_door (5)*
d) *A man kn or (5)*

¼(3+3+5+5)=4.0

Variance

The variance measures how much the individual offsets deviate from the mean.

n - number of times

the two words co-occur

di - the offset

of the co-occurrence *I*

d - the sample mean

of the offsets

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

Sample deviation

$$s = \sqrt{s^2}$$

The mean and the deviation characterize the distribution of distances between two words in a corpus.

$$s = \sqrt{\frac{1^2}{3}\left((3-4.0)^2 + (3-4.0)^2 + (5-4.0)^2 + (5-4.0)^2\right)} \approx 1.15$$

Mean and Variance

- M ≈1.0, s is low ⇒ fixed phrase

(≈ frequency-based approach)

New York, next year, vice president

- M >1.0, s is low ⇒ interesting phrase

(? collocation ?)

Previous/games, hundreds/dollars

- s is high ⇒ not a collocation

Ring New, editorial Atlanta

Frequency vs. Mean and Variance

Fixed phrases ⇒ Frequency-based approach

Flexible phrases ⇒

Variance-based collocation discovery