

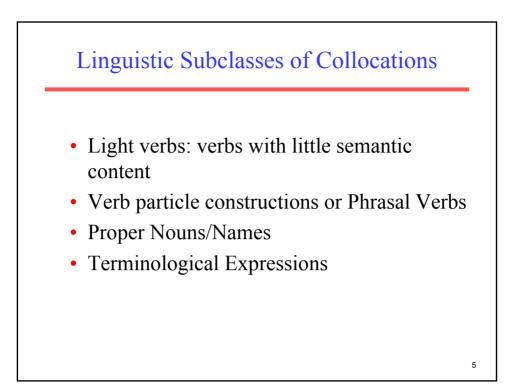
#### Definition (w.r.t Computational and Statistical Literature)

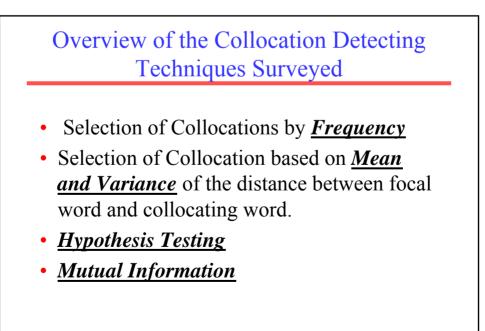
• [A collocation is defined as] a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components. [Chouekra, 1988]

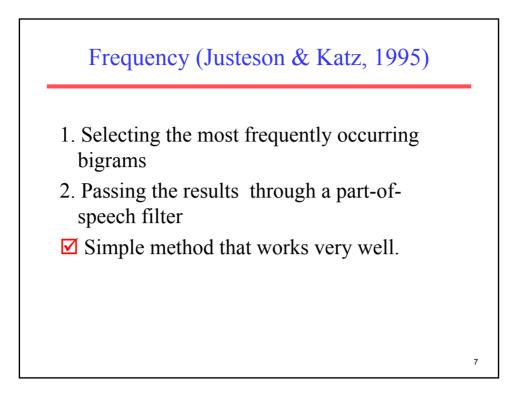
#### Other Definitions/Notions (w.r.t. Linguistic Literature)

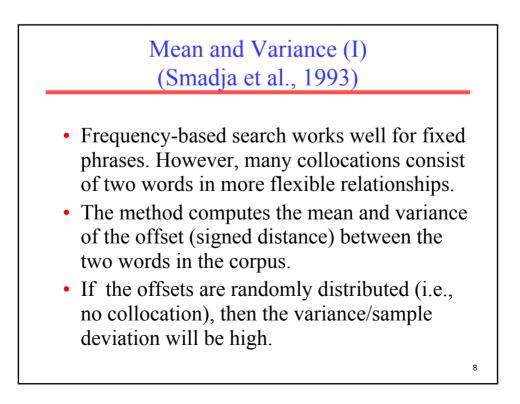
3

- Collocations are not necessarily adjacent
- Typical criteria for collocations: noncompositionality, non-substitutability, nonmodifiability.
- Collocations cannot be translated into other languages.
- Generalization to weaker cases (strong association of words, but not necessarily fixed occurrence.

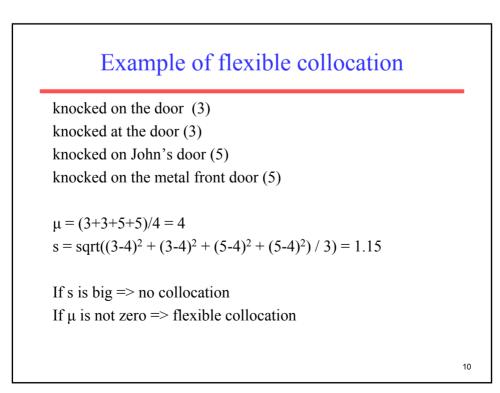


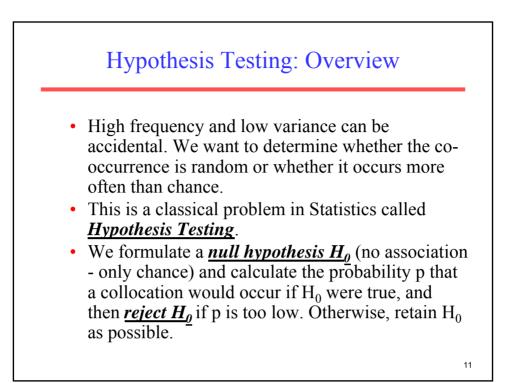


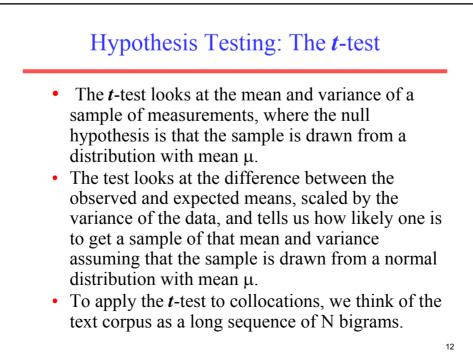




## Mean and Variance (II) • n = number of times two words collocate • $\mu$ = sample mean • $d_i$ = the value of each sample • Sample deviation: $s^2 = \sum_{i=1}^{n} \frac{(d_i - \mu)^2}{n - 1}$









N = number of bigrams  $\mu$  = sample mean for H<sub>0</sub>

 $\overline{\mathbf{X}}$  = observed sample mean



p = probability that the event would occur if H<sub>0</sub> were true
Significance level
p < 0.05 means 95% confidence</li>
p < 0.01 means 99% confidence</li>

new com	Examp panies – col	le llocation or no	t?
	$w_1 = new$	$w_1 \neq new$	
$w_2 = companies$	O <sub>11</sub> = 8	O <sub>12</sub> = 4667	
$w_2 \neq companies$	O <sub>21</sub> = 15820	O <sub>22</sub> = 14287173	
P(new) = $(15820 + 8) /$ P(companies) = $(4667 +$ H <sub>0</sub> : P(new companies) $\overline{x}$ = 8 / 14307668 = 0.0 s <sup>2</sup> = p(1-p) ≈ p	+ 8) / 14307668 = P(new) * P(comp 0000005591		
$t = \frac{0.0000005591 - 0.00000036}{\sqrt{\frac{0.0000005591}{14307668}}}$	15 = 0.999932 < 2.576	=> We cannot reject	null hypothesis

#### Hypothesis testing of differences (Church & Hanks, 1989)

- We may also want to find words whose cooccurrence patterns best distinguish between two words. This application can be useful for lexicography.
- The *t*-test is extended to the comparison of the means of two normal populations.
- Here, the null hypothesis is that the average difference is 0.

15

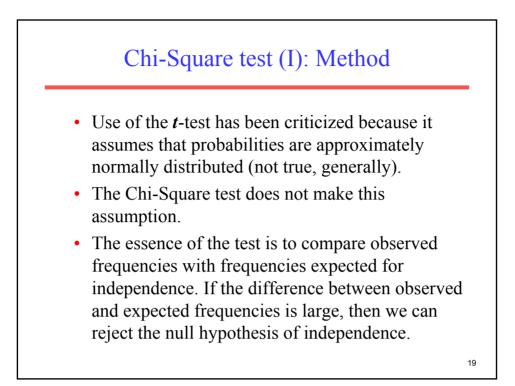
t-test for statistical significance of the difference between two systems				
	System 1	System 2		
scores	71,61,55,60,68,49, 42,72,76,55,64	42,55,75,45,54,51, 55,36,58,55,67		
total	673	593		
n	11	11		
Mean $\overline{\mathbf{X}}_{i}$	61.2	53.9		
$s_i^2 = sum(x_{ij} - \overline{x}_i)^2$	1081.6	1186.9		
df	10	10		

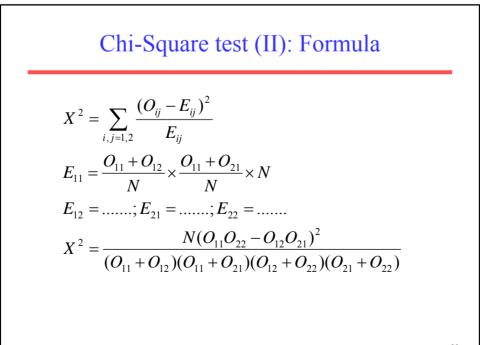
### t-test for differences (continued)

• Pooled  $s^2 = (1081.6 + 1186.9) / (10 + 10) = 113.4$ 

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{2 s^2}{n}}} = \frac{61.2 - 53.9}{\sqrt{\frac{2 113.4}{11}}} = 1.60$$

- For rejecting the hypothesis that System 1 is better then System 2 with a probability level of  $\alpha = 0.05$ , the critical value is t=1.725 (from statistics table)
- We cannot conclude the superiority of System 1 because of the large variance in scores





#### Chi-Square test (III): Applications

- One of the early uses of the Chi square test in Statistical NLP was the identification of translation pairs in aligned corpora (Church & Gale, 1991).
- A more recent application is to use Chi square as a metric for corpus similarity (Kilgariff and Rose, 1998)
- Nevertheless, the Chi-Square test should not be used in small corpora.

21

Example <i>new companies</i> – collocation or not?				
	$w_1 = new$	$w_1 \neq new$		
$w_2 = companies$	O <sub>11</sub> = 8	O <sub>12</sub> = 4667		
$w_2 = companies$	O <sub>21</sub> = 15820	O <sub>22</sub> = 14287173		

 $E_{ij}$  = marginal probabilities = totals of row i and column j converted into proportions = expected values for independence

 $X^2$  = 1.55 < 3.841 needed for p < 0.05, one degree of freedom for 2x2 table

#### Likelihood Ratios I: Within a single corpus (Dunning, 1993)

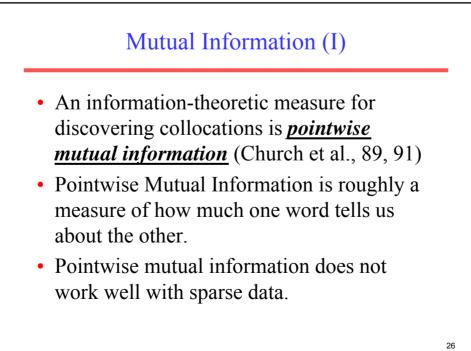
- Likelihood ratios are more appropriate for sparse data than the Chi-Square test. In addition, they are easier to interpret than the Chi-Square statistic.
- In applying the likelihood ratio test to collocation discovery, we examine the following two alternative explanations for the occurrence frequency of a bigram w1 w2:
  - The occurrence of w2 is independent of the previous occurrence of w1
  - The occurrence of w2 is dependent of the previous occurrence of w1

23

# Log likelihood $H_{1}: P(w_{2} | w_{1}) = P(w_{2} | \neg w_{1}) = p$ $H_{2}: P(w_{2} | w_{1}) = p_{1} \neq p_{2} = P(w_{2} | \neg w_{1})$ $p = \frac{c_{2}}{N}; p_{1} = \frac{c_{12}}{c_{1}}; p_{2} = \frac{c_{2} - c_{12}}{N - c_{1}}; c_{1} = C(w_{1}); c_{2} = C(w_{2}); c_{12} = C(w_{1}w_{2});$ $\log \lambda = \log \frac{L(H_{1})}{L(H_{2})} = \frac{b(c_{12}; c_{1}, p)b(c_{2} - c_{12}; N - c_{1}, p)}{b(c_{12}; c_{1}, p_{1})b(c_{2} - c_{12}; N - c_{1}, p_{2})}$ $= \log L(c_{12}, c_{1}, p) + \log L(c_{2} - c_{12}, N - c_{1}, p_{2})$ where $L(k, n, x) = x^{k}(1 - x)^{n-k}$ and b - binomial distrib.

#### Likelihood Ratios II: Between two or more corpora (Damerau, 1993)

- Ratios of *relative frequencies* between two or more different corpora can be used to discover collocations that are characteristic of a corpus when compared to other corpora.
- This approach is most useful for the discovery of subject-specific collocations.



#### Mutual Information (II)

$$MI(x, y) = P(X, Y) \log \frac{P(x, y)}{P(x)P(y)}$$
$$PMI(x, y) = \log \frac{P(x, y)}{P(x)P(y)}$$
$$PMI(x, y) = \log \frac{C(x, y)N}{C(x)C(y)}$$

PMI = E(MI)

#### Example

PMI(new, companies) = = log ((8 \* 14307668) / (4675 \* 15828)) = 1.546

PMI(house, commons) = 4.2 PMI(videocasette, recorder) = 15.94