Corpus-Based Methods in Chinese Morphology and Phonology

中文構詞法和音韻學語料庫方法

Richard Sproat

AT&T Labs — Research

rws@research.att.com

www.research.att.com/∼rws

Chilin Shih

Bell Labs

cls@research.bell-labs.com

www.bell-labs.com/project/tts/cls.html

Lecture 2: Statistical Measures

2001 LSA Institute – University of California, Santa Barbara

Subinstitute on Chinese Corpus Linguistics

Overview of Lecture 2

- Properties of Word Frequency Distributions
- Measures of Morphological Productivity and a Case Study
  - Mandarin Root Compounds: A Case Study in Productivity
- Measures of Association
  - Probability Estimates
  - (Specific) Mutual Information
  - Frequency-Weighted Mutual Information
  - Pearson's χ-Square
  - Likelihood ratios
  - Extracting Non-Binary Collocations

Zipf's Law

(1) $f(w) \propto \frac{1}{r}$

1995 AP Newswire

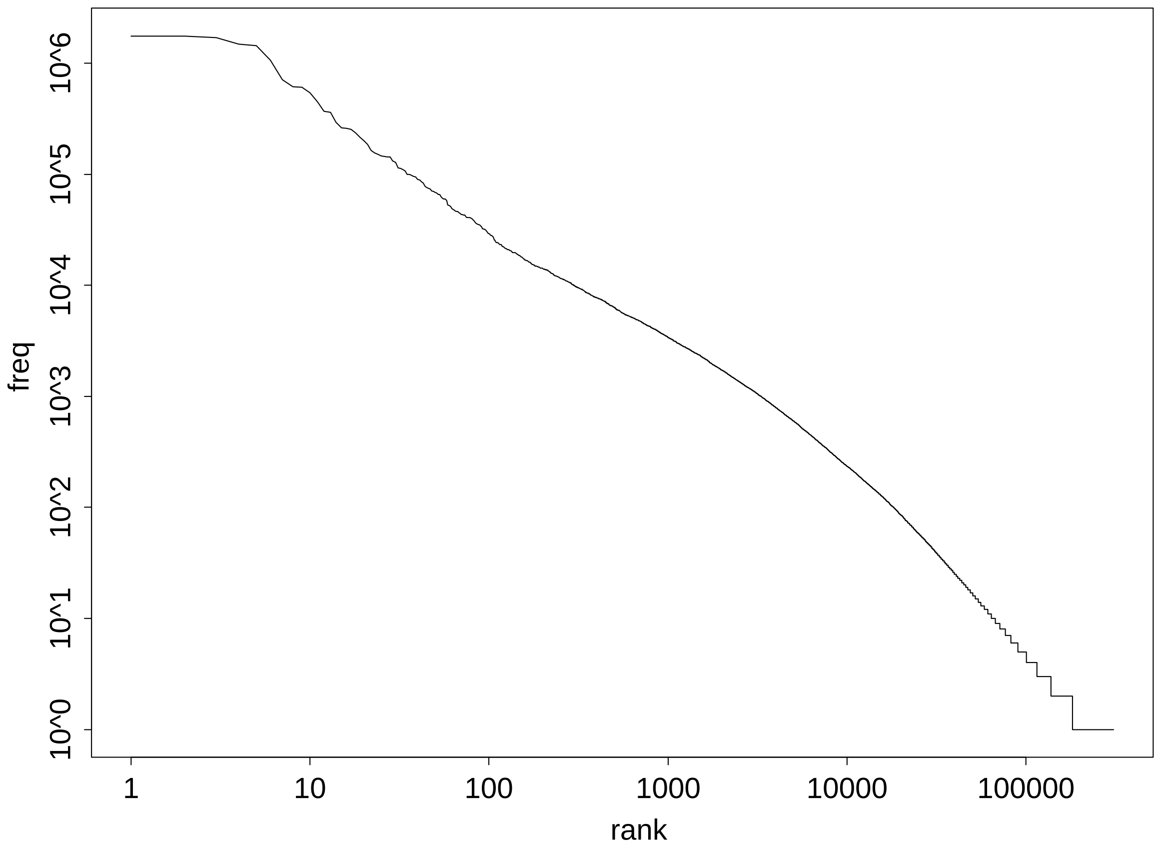

Figure 1: Rank (x) plotted against frequency (y) on a log-log scale for the 1995 Associated Press.

ROCLING Corpus (Characters)

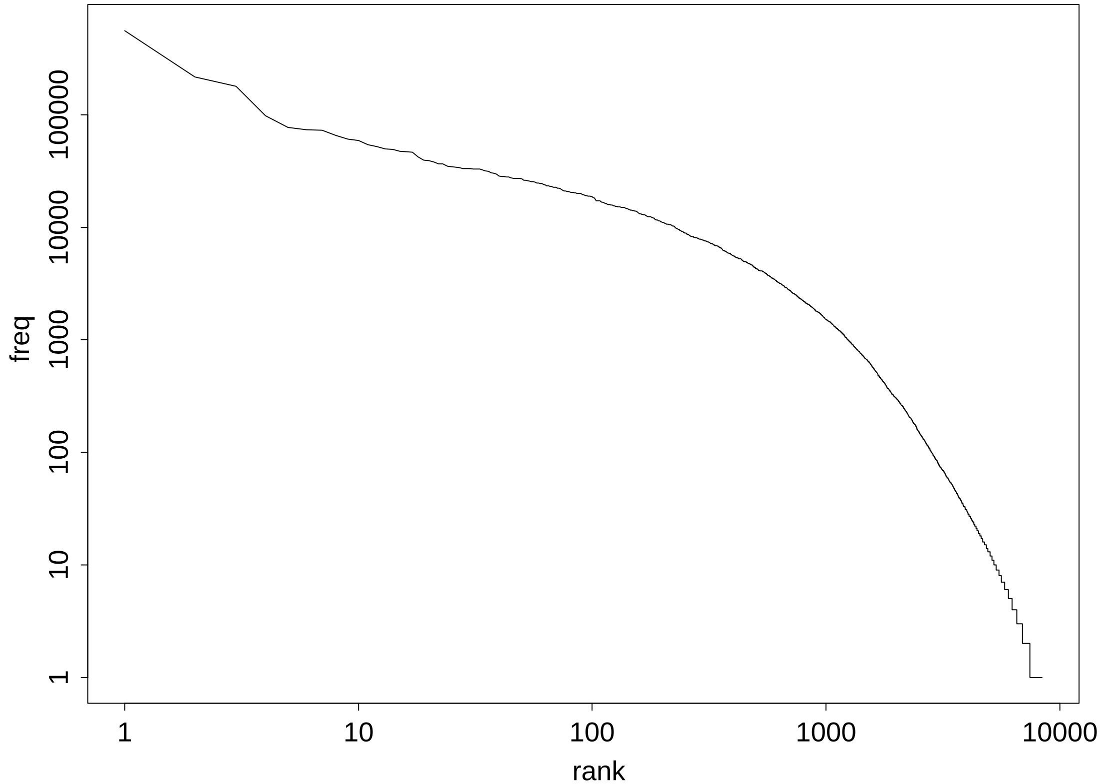

Figure 2: Rank (x) plotted against frequency (y) on a log-log scale over *characters* for the 10 Million character ROCLING corpus.

Important Properties of Word-Frequency Distributions

- Large Number of Rare Events:

  For the 1995 Associated Press corpus 40% of the word types occur just once. (In contrast among the 10 Million character ROCLING corpus, only 11% of the *characters* occur once.) For a smaller corpus, such as the Brown corpus (1 Million words), the amount will be closer to 50%.

  Corollary: words are not normally distributed. Statistical measures that depend on normality (e.g. specific mutual information) are suspect.

- Statistical parameters change with sample size:

  For increasing corpus N the size of the vocabulary $V(N)$ increases. So does the mean frequency $\frac{N}{V(N)}$.

Change of Parameters of Vocabulary with Corpus Size

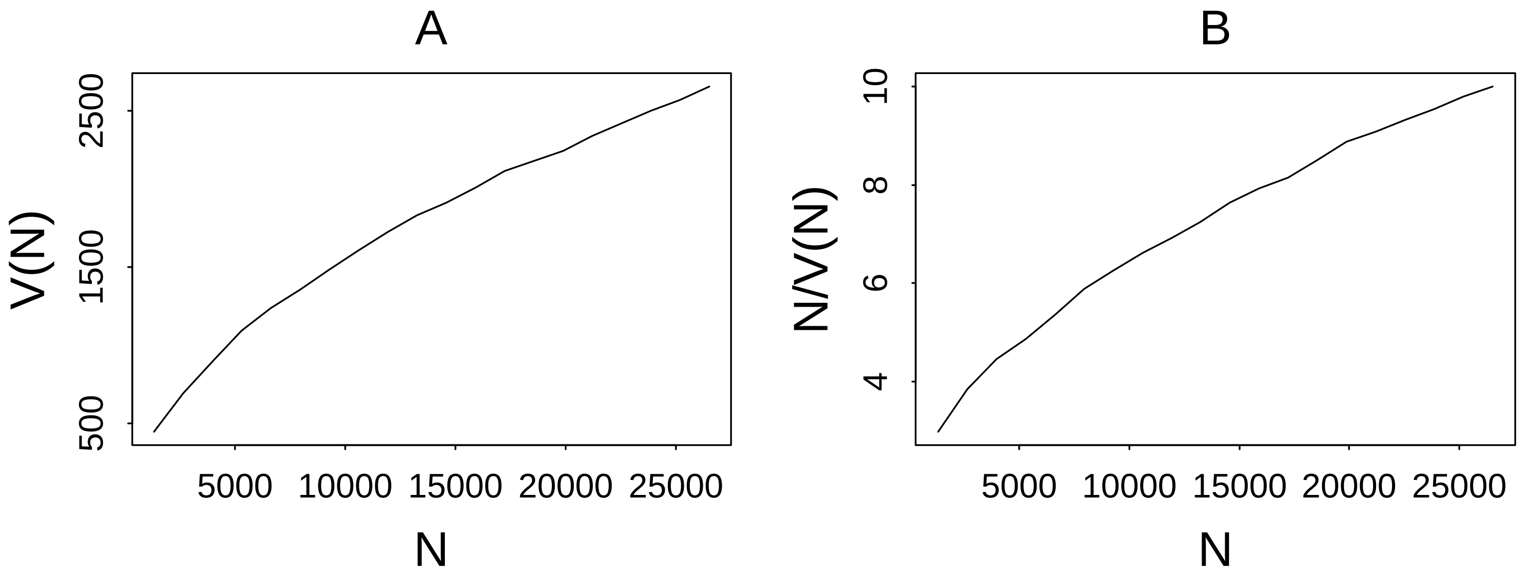

Figure 3: Vocabulary size $V(N)$ (Panel A) and mean word frequency $\frac{N}{V(N)}$ (Panel B) as a function of sample size N in *Alice in Wonderland*, measured at 20 equally spaced intervals. From (Baayen 2001), figure and caption kindly provided by Harald Baayen.

Measures of Morphological Productivity

(2) Aronoff's (1976) proposal: $I = \frac{V}{S}$

(3) Baayen's (1989) proposal: $\mathcal{P} = \dfrac{n_1}{N}$

Intuition behind Good-Turing Measure

Figure 4: A collection of balls where all ball types have been seen.

Intuition behind Good-Turing Measure

Figure 5: A collection of balls where many ball types have not been seen.

Some sample $\mathcal{P}$ scores from Dutch and English

(4)

Morpheme	Gloss	$\mathcal{P}$
-ing	'-ion, -ing'	0.038
-heid	'-ity'	0.114
noun compounds	—	0.225

(5)

Morpheme	$\mathcal{P}$
-ness	0.0044
-ish	0.0034
-ation	0.0006
-ity	0.0007
simplex nouns	0.0001

Some Measures from Mandarin

(6)

Morpheme	Gloss	$\mathcal{P}$
們 *-men*	noun plural	$\frac{247}{5948} = 0.04$
過 *-guò*	experiential	$\frac{330}{1666} = 0.20$

See the first lab exercise for this segment.

Mandarin Root Compounds (Sproat and Shih, 1996)

Some morphological theories, e.g. (Anderson, 1992; Dai, 1992), have a restrictive notion of what can be a compound: only those words formed from two or more other *words*.

(7) 哈密瓜 *hāmìguā* (Hami melon) ‘cantaloupe’

(8) 香腸 *xiāngcháng* (fragrant intestine) ‘sausage’

Mandarin Root Compounds

But bound roots seem to be active in Mandarin morphology:

(9) 螞蟻 *mǎyǐ* ‘ant’
工蟻 *gōngyǐ* ‘worker ant’

(10) 腦子 *nǎozi* ‘brain’
腦水腫 *nǎoshuǐzhǒng* (brain water swelling) ‘hydrocephaly’

(11) 蒼蠅 *cāngyíng* ‘fly’
地中海蠅 *dìzhōnghǎiyíng* ‘Mediterranean fly’

(12) 蘑菇 *mógū* ‘mushroom’
菇傘 *gūsǎn* (mushroom umbrella) ‘pileus’

Mandarin Root Compounds

These formations cannot be an instance of affixation

(13)

蟻 *yǐ*	蟻王	工蟻
	yǐwáng ‘queen ant’	*gōngyǐ* ‘worker ant’
腦 *nǎo*	腦水腫	後腦
	nǎoshuǐzhǒng ‘hydrocephaly’	*hòunǎo* ‘hindbrain’
蠅 *yíng*	蠅屍	地中海蠅
	yíngshī ‘fly corpse’	*dìzhōnghǎiyíng* ‘Mediterranean fly’
菇 *gū*	菇傘	金菇
	gūsǎn ‘pileus’	*jīngū* ‘golden mushroom’

Absent another category, these presumably must be compounds. (NB: Packard (2000) calls them *bound root words*)

Mandarin Root Compounds

224	香菇	104	蘑菇	102	洋菇
23	菇類	16	菇農	9	磨菇
9	草菇	8	菇寮	8	冬菇
8	塊菇	7	金針菇	6	出菇
5	菇價	4	菇傘	4	菇柄
4	鮑魚菇	4	慈菇	3	夏塊菇
3	裸蓋菇	3	茨菇	3	冬季菇
2	菇舍	2	菇木	2	鮮菇
2	發菇	2	採菇	2	夏季菇
2	食用菇	2	食菇	2	春季菇
2	出菇期	2	三菇菜膽	1	菇體
1	菇醇	1	菇腳	1	菇菌
1	菇湯	1	菇場	1	鮑菇
1	褐菇	1	種菇	1	黑菇
1	菌菇	1	產菇量	1	乾菇
1	金菇	1	早發菇	1	生菇
1	可食菇	1	包菇	1	山菇

Table 1: Distribution for the Mandarin nominal root *gū* ‘mushroom’ collected from a 40 million character corpus.

Mandarin Root Compounds

Root	Whole Word	Meaning	n_1	N	V	P_{max}	$\mathcal{P}$
石 *shí*	石頭 *shítou*	'rock'	75	583	147	57	0.129
盒 *hé*	盒子 *hézi*	'box'	82	1205	167	257	0.068
蟻 *yǐ*	螞蟻 *mǎyǐ*	'ant'	21	322	35	173	0.065
蛙 *wā*	青蛙 *qīngwā*	'frog'	22	407	49	199	0.054
龜 *guī*	烏龜 *wūguī*	'turtle'	21	414	46	137	0.051
餃 *jiǎo*	餃子 *jiǎozi*	'dumpling'	8	167	19	66	0.048
蠅 *yíng*	蒼蠅 *cāngyíng*	'fly'	14	325	35	142	0.043
棉 *mián*	棉花 *miánhuā*	'cotton'	52	1283	138	147	0.041
菇 *gū*	蘑菇 *mógū*	'mushroom'	19	598	49	224	0.032
木 *mù*	木頭 *mùtou*	'wood'	265	8904	617	701	0.030
腦 *nǎo*	腦子 *nǎozi*	'brain'	34	1077	75	791	0.032
駝 *tuó*	駱駝 *luòtuó*	'camel'	6	210	16	104	0.029
腸 *cháng*	腸子 *chángzi*	'intestine'	62	2268	148	373	0.027
蜂 *fēng*	蜜蜂 *mìfēng*	'bee'	23	858	63	104	0.027
肚 *dù*	肚子 *dùzi*	'belly'	36	1434	83	734	0.025

Table 2: Productivity measures of some Mandarin nominal roots, measured over a 40 million character corpus.

Measures of Association

- Probability estimates
- Specific Mutual Information
- Frequency-Weighted Mutual Information
- Pearson's χ^2
- Dunning's likelihood ratios
- Non-binary collocations

Probability Estimates

Basic measure of probability is the *maximum likelihood estimate*: $f(t)/N$. This is quite reasonable for frequent words.

	AP92		AP93		AP94	
	f	p	f	p	f	p
the	1,659,949	0.039	1,451,984	0.039	1,311,237	0.039
United	30,883	0.00072	28,336	0.00076	25,456	0.00075
country	21,225	0.00050	18,304	0.00050	15,919	0.00047
night	12,422	0.00029	10,328	0.00028	10,566	0.00031
dog	1,048	2.44e-05	1,045	2.80e-05	992	2.91e-05

Table 3: Maximum likelihood estimates for five common words from three years of the Associated Press (1992–1994). N is 43,012,596, 37,386,960, and 34,041,151, respectively, for these three years.

Probability Estimates

Maximum likelihood estimate becomes less reliable for infrequent words, so typically some *smooth* of the frequency distribution is necessary.

(14) Good-Turing estimate: $r^* = (r+1)\frac{E(n_{r+1})}{E(n_r)}$

The estimates $E(n_r)$ and $E(n_{r+1})$ can be made in various ways, including using the empirical values:

(15) $r^* = (r+1)\frac{n_{r+1}}{n_r}$

For frequencies other than zero this will result in a reestimation downwards. The probabilities will then be reestimated as $\frac{r^*}{N}$, with the probability mass $\frac{n_1}{N}$ reserved for unseen items.

(Specific) Mutual Information

Mutual Information was originally proposed as an information-theoretic measure of channel capacity (Fano 1961).

(16) $I(x;y) = \log_2 \frac{p(x,y)}{p(x)p(y)}$

1995 AP Newswire Collocations

M.I.	f(1,2)	f(1)	f(2)	pair
18.3908	101	104	106	Picket Fences
18.3280	101	101	114	Highly Effective
18.2061	119	121	122	Ku Klux
18.1722	124	124	127	alma mater
18.1574	115	124	119	SPACE CENTER
18.1480	118	127	120	JUDGE LANCE
18.1275	127	131	127	Phnom Penh
18.1266	124	126	129	Velika Kladusa
18.0901	95	103	124	Ginny Terzano
18.0627	134	136	135	Notorious B.I.G
18.0580	116	119	134	Spiritual Laws
18.0486	123	134	127	Deepak Chopra
18.0271	80	105	107	Myriam Sochacki
17.9417	147	149	147	TEL AVIV
17.8421	96	117	131	Reba McEntire
17.7610	108	152	120	Dollar Spin
17.7551	117	124	160	SALT LAKE

Table 4: Sample of highly associated adjacent word pairs from the 37 million words of the 1995 Associated Press. Shown are, from left to right: the mutual information (M.I.); the frequency of the pair f(1,2); the frequency of the first word f(1) and the frequency of the second word f(2). Note that f(1) and f(2) are both greater than 100 for this sample.

1995 AP Newswire Non-Collocations

M.I.	f(1,2)	f(1)	f(2)	pair
-0.000104371	2	44293	1694	But 32
-0.000104371	12	44293	10164	But public
-0.000108717	1	5938	6318	big reports
-0.000122066	7	540363	486	in heroin
-0.000122066	7	486	540363	determination in
-0.000123791	1	20716	1811	says Hall
-0.000135193	2	567	132335	Building from
-0.000135827	1	6639	5651	water given
-0.000135827	1	5651	6639	given water
-0.000137212	1	8365	4485	programs enough
-0.000137212	1	2275	16491	village children
-0.000144883	2	5283	14203	study most
-0.000151325	1	14452	2596	her includes
-0.000163745	1	645	58167	Delaware this

Table 5: Sample of poorly associated adjacent word pairs from the 1995 Associated Press.

Problems with Mutual Information

- It is unreliable for small counts. (But this is really a problem with the MLE)

- The second, and more serious problem is that mutual information relates to estimated probability in a counterintuitive way:

$$I(w_1;w_2) = \log_2\left(\frac{\frac{100}{10{,}000{,}000}}{\frac{100}{10{,}000{,}000}\frac{100}{10{,}000{,}000}}\right) = \log_2(100,000) = 16.6$$

$$I(w_1;w_2) = \log_2\left(\frac{\frac{1{,}000}{10{,}000{,}000}}{\frac{1{,}000}{10{,}000{,}000}\frac{1{,}000}{10{,}000{,}000}}\right) = \log_2(10,000) = 13.3$$

Frequency-Weighted Mutual Information

$$(17)\quad I_{fw}(x;y) = f(x,y)\log_2 \frac{P(x,y)}{P(x)P(y)}$$

1995 AP Newswire Collocations

F.W.M.I.	f(1,2)	f(1)	f(2)	pair
469631	174742	708948	1435262	of the
341288	129132	540363	1435262	in the
184196	16797	16816	18733	Associated Press
182881	17287	24135	17564	United States
181021	66059	258389	1435262	to the
144801	31575	447529	110206	to be
140778	51336	200524	1435262	on the
136748	12956	24177	13364	New York
135747	19968	112171	60003	have been
135610	12452	17858	13778	All Rights
134177	28748	144059	294607	he said
133324	11684	13778	11684	Rights Reserved
120476	17592	95448	60003	has been
118810	50201	254405	1435262	for the
114737	17820	69930	110206	will be
109856	15576	261695	16816	The Associated
107843	36922	127433	1435262	at the
97455	9561	10928	28038	White House
96982	15799	76338	110206	would be

Table 6: Sample of highly associated adjacent word pairs from the 1995 Associated Press. Shown are, from left to right: the frequency weighted mutual information (F.W.M.I.); the frequency of the pair f(1,2); the frequency of the first word f(1) and the frequency of the second word f(2). Again, f(1) and f(2) are both greater than 100 for this sample.

Problems with Frequency-Weighted Mutual Information

Main problem is that it tends to over-reward frequency.

Pearson's χ-Square

- χ-square provides a confidence measure for rejecting an assumption of independence between events.
- χ-square applies to tables:

	$w_1 = \textit{new}$	$w_1 \neq \textit{new}$
$w_1 = \textit{companies}$	8 (*new companies*)	4,667 (e.g. *old companies*)
$w_1 \neq \textit{companies}$	15,820 (*new machines*)	14,287,181 (e.g. *old machines*)

Table 7: A 2×2 table showing the distribution of bigrams in a corpus (from Manning and Schütze, 1999, Table 5.8, page 169). There were 8 instances of *new companies*, 4,667 instances of *X companies*, where *X* is different from *new*, 15,820 instances of *new Y*, where *Y* is different from *companies*, and 14,287,181 instances of *XY*, where *X* and *Y* are different, respectively, from *new* and *companies*.

Pearson's χ-Square

General statement:

(18) $$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Special case for 2×2 table:

(19) $$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

1995 AP Newswire Collocations

χ^2	f(1,2)	f(1)	f(2)	pair
999319	600	6262	2156	attorney general
996401	47	280	297	Connie Mack
993054	192	552	2522	20th century
991404	36	299	164	OR MUSIC
991243	306	818	4330	atomic bomb
990132	145	1709	466	loan guarantees
986765	18	113	109	Geographic Explorer
985647	375	2599	2058	Catholic Church
985038	67	1540	111	racial slur
984217	84	1379	195	highly publicized
983361	147	284	2902	plead guilty
982709	23	159	127	Justices Antonin
975478	289	7415	433	Sen Arlen
972147	22	116	161	Celtic Journey
970830	322	2808	1426	illegal immigrants
968693	126	2984	206	flight attendant
968378	261	1571	1679	pleaded innocent
968056	20	124	125	Joey Buttafuoco
967695	37	147	361	silicone implants

Table 8: Sample of highly associated adjacent word pairs from the 1995 Associated Press, using chi-square. Shown are, from left to right: the chi square value; the frequency of the pair f(1,2); the frequency of the first word f(1) and the frequency of the second word f(2). Again, f(1) and f(2) are both greater than 100 for this sample.

Problems with χ-Square

- χ-square still assumes normality, which can be violated for small counts.
- χ-square is a symmetric measure: it does not distinguish between events that are much more likely than chance from events that are much *less* likely than chance.

  tape-recorded conversations — a plausible collocation — you find that *tape-recorded* occurs 100 times, *conversations* 639 times, and the collocation 7 times, with a high χ-square value of 28,752.8.

  Similarly for *sickening reminder* the breakdown is, 17 for *sickening*, 307 for *reminder* and 2 for the pair, with a χ-square value of 28,747.7.

  However a very similar χ-square value is obtained for *the of*, where *the* occurs 1,435,262 times, *of* 708,948 times, the pair exactly once (due to a typo in the data), yielding a χ-square value of 28,744.5.

Dunning's (1993) Likelihood ratios

- Hypothesis 1: $p(w_2|w_1) = p = p(w_2|\neg w_1)$
- Hypothesis 2: $p(w_2|w_1) = p_1 \neq p_2 = p(w_2|\neg w_1)$
- $p = \frac{c_2}{N}, p_1 = \frac{c_{12}}{c_1}, p_2 = \frac{c_2 - c_{12}}{N - c_1}$
- Assuming a binomial distribution: $b(k; n, p_x) = \binom{n}{k} p_x^k (1 - p_x)^{(n-k)}$

$L(H_1) = b(c_{12}; c_1, p) b(c_2 - c_{12}; N - c_1, p)$

$L(H_2) = b(c_{12}; c_1, p_1) b(c_2 - c_{12}; N - c_1, p_2)$

$$\begin{aligned}
\log \lambda &= \log \frac{L(H_1)}{L(H_2)} \\
&= \log L(c_{12}, c_1, p) + \log L(c_2 - c_{12}, N - c_1, p) \\
&\quad - \log L(c_{12}, c_1, p_1) - \log L(c_2 - c_{12}, N - c_1, p_2)
\end{aligned}$$

(where $L(k, n, x) = x^k (1 - x)^{n-k}$)

1995 AP Newswire Collocations

$-2 \log \lambda$	f(1,2)	f(1)	f(2)	pair
587215.02	174742	708948	1435262	of the
417624.29	129132	540363	1435262	in the
403795.56	16797	16816	18733	Associated Press
387423.99	17287	24135	17564	United States
281913.17	12956	24177	13364	New York
279548.21	12452	17858	13778	All Rights
230483.10	19968	112171	60003	have been
223206.08	66059	258389	1435262	to the
218798.60	31575	447529	110206	to be
211761.70	15576	261695	16816	The Associated
203728.79	17592	95448	60003	has been
200819.84	28748	144059	294607	he said
192063.84	9561	10928	28038	White House
190007.67	17820	69930	110206	will be
173072.45	51336	200524	1435262	on the
161027.74	194	1435262	1435262	the the
157326.20	15799	76338	110206	would be
143998.60	9530	24496	40710	more than
143592.45	10387	19586	89982	did not
142904.19	18933	1435262	24135	the United

Table 9: Sample of highly associated adjacent word pairs from the 1995 Associated Press, using log likelihood ratios. Shown are, from left to right: the value of $-2 \log \lambda$ (which is asymptotically χ-square distributed; the frequency of the pair f(1,2); the frequency of the first word f(1) and the frequency of the second word f(2). Again, f(1) and f(2) are both greater than 100 for this sample.

Problems with Likelihood Ratios

Shares with χ-square the problem that it is symmetric

$-2\log\lambda$	f(1,2)	f(1)	f(2)	pair
4988.08	340	14203	2247	most powerful
1959.01	420	607952	2247	a powerful
1336.07	131	24496	2247	more powerful
1093.76	89	7967	2247	most powerful
532.11	57	14183	2247	very powerful
527.49	36	2247	1410	powerful earthquake
438.62	285	1435262	2247	the powerful
363.11	31	2247	3345	powerful lower
309.31	22	1053	2247	politically powerful
293.39	20	2247	776	powerful storms
249.88	30	10575	2247	so powerful
237.09	1	2247	1435262	powerful the
225.00	31	15989	2247	as powerful

Table 10: Possible collocations of *powerful*, along with their log likelihood ratios, from the 1995 Associated Press.

Extracting Non-Binary Collocations

- Can extend approaches for the binary case. For example, mutual information can be defined for triples (Chang and Su, 1997):

  $$I(x, y, z) = \log \frac{P(x, y, z)}{P_I(x, y, z)}$$

  where

  $$P_I(x, y, z) = P(x)P(y)P(z) + P(x)P(y, z) + P(x, y)P(z)$$

- Or, more commonly, different approaches have been used.

Smadja's 1993 *Xtract*

- Compute a concordance of all instances of a given word w in a corpus.

- For each w_i found in the context of w, a profile is computed of how often each w_i occurs in each position in a window ranging from five to the left to five to the right of w.

Smadja's 1993 *Xtract*

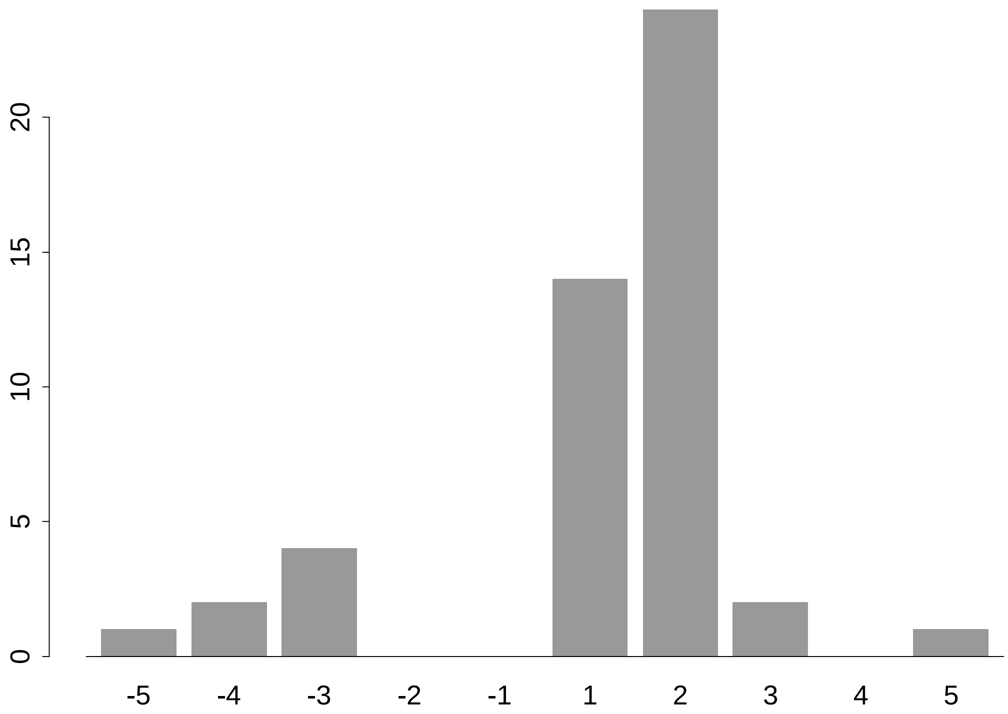

Figure 6: Occurrences of *forecast* in a window around *weather*, computed on the the 1995 Associated Press corpus.

Smadja's 1993 *Xtract*

- Remove "uninteresting" words based on property of profile. e.g.: the profile must show at least one peak. A flat distribution suggests a pair of words that is only semantically associated, such as *doctor* and *nurse*.

- Once interesting two-word collocates are found, compute further concordances for each pair of words found in the first phase, for each relative position. Recurring sequences of words are kept as potential collocates. For example, a collocation *blue. . . stocks* discovered during the first phase would be replaced in the second phase by *blue* **chip** *stocks*, since *chip* would occur frequently in this context.

- Fung and Wu (1994) have applied the Xtract system to Chinese.

Readings for Lecture 2

- Baayen (2001): Chapter 1.
- Sproat and Shih (1996).
- Manning and Schütze (1999): Chapter 5.
- Fung and Wu (1994).

Lab Assignment for Lecture 2

1. Using the measure $\mathcal{P}$ introduced in this lecture, measure the productivity of the following Mandarin affixes, using the Penn Chinese Treebank:
   (a) -們 -*men*
   (b) -子 -*zi*
   (c) -率 -*l`ü*
   (d) -家 -*jiā*
   (e) -了 -*le* (perfective affix)
   (f) -過 -*guò* (experiential affix)
   (g) -見 -*jiàn* (experiential affix)
   (h) -下 -*xià* (resultative affix)

2. Consider the four association measures:

   - mutual information
   - χ-square
   - weighted mutual information
   - likelihood ratios

   For each measure, rate its efficacy at extracting reasonable binary terms from the ROCLING corpus (10-million characters). It is suggested that, for each of the three measures, you select the top ranked N, where N is a reasonable number such as 100. Then simply mark each example as to how much it feels like a word.