Computational Approaches to Collocations

Vienna, July 2002

STS: Mathematical Properties of AMs

by Stefan Evert

Collocation Extraction Procedure

- **source text**, e.g. *Frankfurter Rundschau* corpus ($\approx$ 40 million words)
- **pre-processing**: reformatting/conversion, tokenisation, spelling corrections (?)
- **linguistic annotations**:
  part-of-speech, lemma (citation forms), morphosyntactic features,
  chunk parsing ($\rightarrow$ YAC), full parsing with complex grammar
- **collocation candidates**:
  syntactic patterns based on part-of-speech and chunk annotations,
  or direct extraction from syntax trees
- **large number of candidates**: e.g. Adj+Noun pairs from *Frankfurter Rundschau*:
  $N = 1\,505\,192$ **tokens** (instances) and $V = 537\,743$ **types** (different pairs)
  $\rightarrow$ need for **filtering** or **ranking** techniques

What is a collocation?

- "collocation" can be defined in many different ways, depending on the application
- Manning and Schütze (1999) identify three major criteria used in NLP applications: **non-compositionality**, **non-substitutability**, and **non-modifiability**
- statistical approaches are based on J. R. Firth's notion of collocations:

You shall know a word by the company it keeps!

> Collocations of a given word are statements of the habitual or customary places of that word . . . The collocation of a word or a 'piece' is not to be regarded as mere juxtaposition, it is an order of *mutual expectancy*.
>
> Firth (1957), A synopsis of linguistic theory 1930–55

- in this STS (and related work) we make further restrictions on the candidate data:

collocation candidates are **lexical arguments of binary syntactic relations**

Co-occurrence Frequencies

- the citation from Firth (1957) suggests that collocations are characterised by **high co-occurrence frequency**

$\rightarrow$ **rank** candidates by frequency or apply frequency **threshold**

- initial results are fairly good, but **Zipf's Law** leads to low recall:

	$f=1$	$f=2$	$f=3$	$f=4$	$f=5$	$f=6$	$f=7$	$f=8$
# types	377 881	77 413	25 487	14 243	8 193	5 945	4 090	3 315

- the $3\,315$ candidates with $f=8$ include *beifälliges Nicken* (approving nod) and *vegetatives Nervensystem* (vegetative nervous system), but also obviously random combinations such as *erste Partei* and *schöner Teil*
- $f(\text{beifällig}) = 16$ and $f(\text{Nicken}) = 11$, but $f(\text{schön}) = 3\,594$ and $f(\text{Teil}) = 4\,536$
  $\rightarrow$ frequency of *beifälliges Nicken* is **higher than expected**

Contingency Table (observed frequencies)

	$w_2 =$Nicken	$w_2 \neq$Nicken	
$w_1 =$beifällig	O_{11}	O_{12}	
$w_1 \neq$beifällig	O_{21}	O_{22}	

$$O_{11} + O_{12} + O_{21} + O_{22} = N$$

Contingency Table (observed frequencies)

	$w_2 =$Nicken		$w_2 \neq$Nicken	
$w_1 =$beifällig	O_{11}	$+$	O_{12}	$= R_1$
	$+$		$+$	
$w_1 \neq$beifällig	O_{21}	$+$	O_{22}	$= R_2$
	$= C_1$		$= C_2$	

$$O_{11} + O_{12} + O_{21} + O_{22} = N$$

Contingency Table (observed frequencies)

	$w_2 =$Nicken		$w_2 \neq$Nicken	
$w_1 =$beifällig	8	$+$	8	$= 16$
	$+$		$+$	
$w_1 \neq$beifällig	3	$+$	$1\,505\,173$	$= 1\,505\,176$
	$= 11$		$= 1\,505\,181$	

$N = 1\,505\,192$ Adj+N pairs (instances) extracted from YAC-parsed *Frankfurter Rundschau* corpus ($\approx$ 40 million tokens)

Expected vs. Observed Frequencies

	$w_2 = B$	$w_2 \neq B$
$w_1 = A$	$E_{11} = \frac{R_1 C_1}{N}$	$E_{12} = \frac{R_1 C_2}{N}$
$w_1 \neq A$	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$

expected frequencies

	$w_2 = B$	$w_2 \neq B$
$w_1 = A$	O_{11}	O_{12}
$w_1 \neq A$	O_{21}	O_{22}

observed frequencies

Mutual Information

- assuming random combinations, the expected co-occurrence frequency is $E_{11} = \frac{R_1 C_1}{N}$
- use observed-to-expected ratio as **measure of association** between lexemes

$$\textbf{MI} = \log \frac{O_{11}}{E_{11}}$$

this measure has become known as **mutual information** (from information theory)

- however, in applications MI has been shown
to overestimate association between low-frequency pairs dramatically

→ measures derived from statistical hypothesis tests correct for "small sample size"

- **definition**: an association measure (**AM**) is a formula
which computes an association score for a candidate pair from its contingency table

Corpus as a Random Sample

Population	$w_2 = B$	$w_2 \neq B$
$w_1 = A$	T_{11}	T_{12}
$w_1 \neq A$	T_{21}	T_{22}

$$N_0 = T_{11} + T_{12} + T_{21} + T_{22}$$

Sample	$w_2 = B$	$w_2 \neq B$
$w_1 = A$	O_{11}	O_{12}
$w_1 \neq A$	O_{21}	O_{22}

$$N = O_{11} + O_{12} + O_{21} + O_{22}$$

→ random variables $(\boldsymbol{X_{11}, X_{12}, X_{21}, X_{22}})$ are **multinomially distributed**
with sample size N and probability parameters $\frac{T_{11}}{N_0}, \frac{T_{12}}{N_0}, \frac{T_{21}}{N_0}, \frac{T_{22}}{N_0}$

Multinomial Sampling Distribution

- for a random sample of size N from the population,
the random variables $(X_{11}, X_{12}, X_{21}, X_{22})$ are **multinomially distributed**:

$$P\big(X_{11} = k_{11} \wedge X_{12} = k_{12} \wedge X_{21} = k_{21} \wedge X_{22} = k_{22}\big) = \frac{N!}{k_{11}!\,k_{12}!\,k_{21}!\,k_{22}!} \cdot \left(\frac{T_{11}}{N_0}\right)^{k_{11}} \cdot \left(\frac{T_{12}}{N_0}\right)^{k_{12}} \cdot \left(\frac{T_{21}}{N_0}\right)^{k_{21}} \cdot \left(\frac{T_{22}}{N_0}\right)^{k_{22}}$$

- each X_{ij} is **binomially distributed**:

$$P(X_{ij} = k) = \binom{N}{k} \cdot \left(\frac{T_{ij}}{N_0}\right)^{k} \cdot \left(1 - \frac{T_{ij}}{N_0}\right)^{N-k}$$

but the X_{ij} are **not independent** of each other

Relative Frequencies

$$\pi = \frac{T_{11}}{N_0} \qquad \pi_1 = \frac{T_{11} + T_{12}}{N_0} \qquad \pi_2 = \frac{T_{11} + T_{21}}{N_0}$$

true relative frequencies
(population)

$$p = \frac{O_{11}}{N} \qquad p_1 = \frac{R_1}{N} = \frac{O_{11} + O_{12}}{N} \qquad p_2 = \frac{C_1}{N} = \frac{O_{11} + O_{21}}{N}$$

observed relative frequencies
(sample)

Statistical Hypothesis Tests

- **null hypothesis** H_0 and **alternative hypothesis** H_1 are statements about relative frequencies (= probabilities) in the population

- test for **independence**:
  H_0 stipulates that a given candidate pair is a random combination of two lexemes

$$H_0: \ \pi = \pi_1 \cdot \pi_2$$

- unknown parameters are estimated from sample: $\pi_1 \approx p_1$ and $\pi_2 \approx p_2$

$$\boxed{H_0: \ \pi = \pi_0 \ := \pi_1 \cdot \pi_2 \approx p_1 \cdot p_2}$$

- test decides whether sample provides sufficient evidence to reject null hypothesis, by comparison with sampling distribution under H_0 (written as $P_0(\ldots)$ and $E_0[\ldots]$)

One-Sided and Two-Sided Tests

- **two-sided test** rejects H_0 if true value of π is different from π_0

$$H_0^{\text{(two-sided)}}: \ \pi \neq \pi_0$$

- **one-sided test** rejects H_0 *only* if frequency is higher than expected

$$H_0^{\text{(one-sided)}}: \ \pi > \pi_0$$

- in our situation, one-sided test is appropriate

- some tests are inherently two-sided → candidates with $p < \pi_0$ must be excluded

- one-sided tests are slightly less conservative than two-sided tests
  → best solution is to use two-sided test and discard candidates with $p < \pi_0$

Exact Hypothesis Tests

- hypothesis test is based on **sampling distribution** of X_{ij} with expected frequencies

$$E_0[X_{ij}] = E_{ij} = \frac{R_i C_j}{N}$$

- **significance** (or p-**value**) of a given sample is the probability of observing a deviation from the expected frequencies that is at least as great as in the sample

- H_0 is rejected if p-value is smaller than a pre-defined **significance level** α:

$$P_0(X_{11} \geq O_{11}) < \alpha$$

  (this test only compares O_{11} to E_{11} → most immediate evidence against H_0)

- low **significance level** = high degree of certainty = conservative test
  (typical values are $\alpha = 0.05$ (95%), $\alpha = 0.01$ (99%), or $\alpha = 0.001$ (99.9%))

Binomial Test

- correct binomial distribution for X_{11} leads to **binomial test**

$$\textbf{binomial} = \sum_{k=O_{11}}^{N} \binom{N}{k} \pi_0^k (1-\pi_0)^{N-k}$$
$$= \sum_{k=O_{11}}^{N} \binom{N}{k} \left(\frac{E_{11}}{N}\right)^k \left(1 - \frac{E_{11}}{N}\right)^{N-k}$$
$$= 1 - \sum_{k=0}^{O_{11}-1} \binom{N}{k} \left(\frac{E_{11}}{N}\right)^k \left(1 - \frac{E_{11}}{N}\right)^{N-k}$$

  where $P_0(X_{11} \geq O_{11}) = \sum_{k=O_{11}}^{N} P_0(X_{11} = k)$ is expanded

- computation of exact probabilities for large samples may lead to numerical difficulties

Poisson Test

- for large sample size N and comparatively small E_{11}, the binomial distribution can be approximated with the numerically easier Poisson distribution → **Poisson test**

$$\textbf{Poisson} = \sum_{k=O_{11}}^{\infty} e^{-E_{11}} \frac{(E_{11})^k}{k!} = 1 - \sum_{k=0}^{O_{11}-1} e^{-E_{11}} \frac{(E_{11})^k}{k!}$$

- no upper limit for X_{11}, but probabilities are vanishingly small when $X_{11} > N$
- small p-values indicate strong rejection of H_0
→ it is convenient to show the negative decadic logarithm: $-\log_{10}(p\text{-value})$
- **convention**: higher AM scores indicate stronger association
- exact p-values for binomial test and Poisson test are still difficult to compute for high-frequency candidates ($O_{11} > 100$, perhaps even lower)

Exact and Asymptotic Tests

- if N and E_{11} are sufficiently large, the binomial (or Poisson) distribution of X_{11} is approximately **normal**, with parameters $\mu = E_{11}$ and $\sigma^2 \approx E_{11}$
- the standardised **z-score** of X_{11} approximates a standard normal distribution:

$$\textbf{z-score} = \frac{O_{11} - E_{11}}{\sqrt{E_{11}}}$$

- unlike the p-value obtained from an **exact test**, an **asymptotic test** computes a **test statistic**, which approximates a known distribution for $N \to \infty$
- the z-score statistic can be converted into a p-value using tables (traditionally) or software (sensibly) for the limiting normal distribution
- for a one-sided asymptotic test like z-score,
multiply p-values by 2 to obtain the more conservative behaviour of a two-sided test

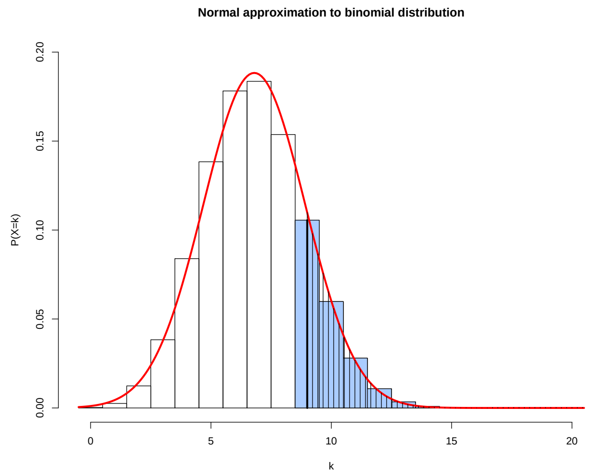
Yates' Continuity Correction

- the z-score measure uses a continuous distribution (normal distribution) to approximate discrete distributions (binomial or Poisson)
- Yates' continuity correction reduces $|O_{ij} - E_{ij}|$ by 0.5 in order to correct for quantisation error when computing p-values from the continuous approximation:

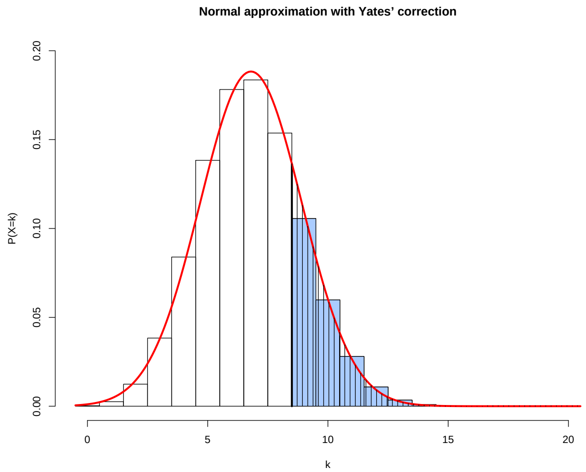
$$O_{ij} := O_{ij} - 0.5 \quad \text{if } O_{ij} > E_{ij}$$
$$O_{ij} := O_{ij} + 0.5 \quad \text{if } O_{ij} < E_{ij}$$

- Yates' correction greatly improves the normal approximation of z-score, but its applicability in other situations is less obvious, and statisticians disagree whether it should be used at all (Motulsky, 1995, Chapter 37)
- in many situations, Yates' does not lead to a better approximation of the limiting distribution, but it makes the test more conservative (Agresti, 1990, p. 68)

Yates' Continuity Correction



Yates' Continuity Correction



More Asymptotic Tests

- Pearson's **chi-squared** test X^2 (for independence of rows and columns) approximates χ^2 distribution with df=1 (degrees of freedom): 4 squares – 1 constraint – 2 estimates

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

- when Yates's continuity correction is applied, the chi-squared formula becomes

$$\textbf{chi-squared}_i = \sum_{i,j} \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}}$$

- the **t-score** obtained from a t-test approximates Student's t distribution with df=N ($\approx$ df=∞); assumes normal distributions for binary indicator variables $\rightarrow$ questionable

$$\textbf{t-score} = \frac{O_{11} - E_{11}}{\sqrt{O_{11}}}$$

Indicator Variables

$$I_{11}^{(m)} = \begin{cases} 1 & \text{if } w_1 = A \wedge w_2 = B \text{ for the } m\text{-th pair in the sample} \\ 0 & \text{otherwise} \end{cases}$$

$$I_{12}^{(m)} = \begin{cases} 1 & \text{if } w_1 = A \wedge w_2 \neq B \text{ for the } m\text{-th pair in the sample} \\ 0 & \text{otherwise} \end{cases}$$

$$I_{21}^{(m)} = \begin{cases} 1 & \text{if } w_1 \neq A \wedge w_2 = B \text{ for the } m\text{-th pair in the sample} \\ 0 & \text{otherwise} \end{cases}$$

$$I_{22}^{(m)} = \begin{cases} 1 & \text{if } w_1 \neq A \wedge w_2 \neq B \text{ for the } m\text{-th pair in the sample} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{ij} = \sum_{m=1}^{N} I_{ij}^{(m)}$$

Homogeneity Tests

$$\boxed{\begin{aligned} \rho &= \frac{T_{11} + T_{12}}{N_0} \\ \rho_1 &= \frac{T_{11}}{T_{11} + T_{21}} \\ \rho_2 &= \frac{T_{12}}{T_{12} + T_{22}} \end{aligned}} \qquad \boxed{\begin{aligned} r &= \frac{R_1}{N} = \frac{O_{11} + O_{12}}{C_1 + C_2} \\ r_1 &= \frac{O_{11}}{C_1} = \frac{O_{11}}{O_{11} + O_{21}} \\ r_2 &= \frac{O_{12}}{C_2} = \frac{O_{12}}{O_{12} + O_{22}} \end{aligned}}$$

population ratios   observed ratios

$$\boxed{H_0 : \ \rho_1 = \rho_2 = \rho \ \approx r}$$

- Pearson's chi-squared test for homogeneity is equivalent to the test for independence

$$\textbf{chi-squared}_h = \frac{N\left(O_{11}O_{22} - O_{12}O_{21}\right)^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})}$$

- **log-likelihood** $G^2 = -2\log\lambda$ (likelihood-ratio test, χ^2 distribution with df=1)

$$\lambda = \frac{L(O_{11},C_1,r)\cdot L(O_{12},C_2,r)}{L(O_{11},C_1,r_1)\cdot L(O_{12},C_2,r_2)} \quad \text{where } L(k,n,r) = r^k(1-r)^{n-k}$$

- G^2 has a much simpler equivalent form (known as the entropy version):

$$\textbf{log-likelihood} = 2\sum_{ij} O_{ij}\log\frac{O_{ij}}{E_{ij}}$$

(note that log-likelihood is a two-sided test!)

Assessing the Quality of a Test

- most important mathematical criterion for asymptotic tests:
  How well does the test statistic approximate its limiting distribution?
- Dunning (1993) shows that chi-squared statistic X^2 gives poor approximation of the χ^2 distribution for low-frequency candidates (any $E_{ij} < 5$) and suggests to use G^2
- according to textbooks, Pearson's X^2 converges more quickly to a χ^2 distribution than the G^2 statistic obtained from a likelihood-ratio test (Agresti, 1990, p. 49)
  $\rightarrow$ for small sample sizes, G^2 gives a poor approximation
- but we have a *large* sample (size $= N$) with a highly skewed distribution
- Pedersen (1996) recommends **Fisher's exact test** for very low frequency pairs (this does *not* necessarily imply a poor approximation of the χ^2 distribution by G^2, since Fisher's test is based on a different null hypothesis)

Fisher's Exact Test

- the null hypothesis of Fisher's exact test is *not* a statement about the full population
  $\rightarrow$ only the observed sample is considered
- assumes fixed row and column totals (= marginal frequencies)
- under H_0 the fixed numbers of lexemes are randomly combined into pairs, leading to a **hypergeometric distribution** for X_{11}

$$\textbf{Fisher} = \sum_{k=O_{11}}^{\min\{R_1,C_1\}} \frac{\binom{C_1}{k}\cdot\binom{C_2}{R_1-k}}{\binom{N}{R_1}}$$

- the row and column totals in the formula above can be exchanged
- Fisher's test is one-sided;
  as an exact test it yields p-values and suffers from numerical complexity

Directions for Future Research

- ongoing research for my PhD project (and joint work with Brigitte Krenn)
- empirical investigations into the mathematical properties of AMs
- know your numbers: the question of numerical accuracy
- do we need yet another association measure?
- statistics (association measures) for fractional counts
- beyond bigrams: n-gram statistcs and the influence of categorical variables

The PhD Thingy

- my project: **Understanding Collocation Statistics** (working title)
- current goals
  - restriction to *lexical arguments of binary syntactic relations*
  - a reference including all widely-used AMs, with explanation of their background, connections between AMs, and analysis of their mathematical properties
  - implementation guidelines and details, ensuring numerical accuracy
  - methods and tools for the empirical evaluation of AMs, based on manual annotation (includes techniques for evaluation of random samples to reduce workload)
  - significance tests for (empirical) differences between AMs
  - what factors influence the performance of an AM? (e.g. corpus size, pre-processing, extraction, filtering, type of collocation)
  - examples: Adj+N and PP+V pairs extracted from *Frankfurter Rundschau*
- the companion website (work in progress): http://www.collocations.de/

Short-Term Goals

- put a short HTML version of this presentation on the website at

  `http://www.collocations.de/AM/`

  which supersedes *On lexial association measures* written in June 2001 (available from `http://www.collocations.de/EK/`)
- start a central repository of association measures, including short explanation, references, formula in terms of O_{ij} and E_{ij}, and connections to other AMs
  → send input to `evert@ims.uni-stuttgart.de`
- software for comparative empirical evaluation:
  a collection of Perl scripts and R code called the `UCS` system
  - no support for bigram extraction → complement to Pedersen's `BSP/NSP`
  - early release version will hopefully be available soon (Unix only)

Empirical Investigations

- precise mathematical analysis of the properties of AMs is tedious
  → obtain empirical results (cf. Monte Carlo and randomisation methods)
- method: compute AM scores for a large number of random contingency tables, then compare results for different AMs, formulae, frequency layers etc.
- lazy man's approach: construct mock data set where the O_{ij} vary systematically, then use UCS tools to annotate data set with AM scores and compare results
- data set should cover wide frequency ranges, with higher density for small frequencies
- need to choose fixed sample size to avoid having too many candidates
  suggested representative sample size is $N = 1\,000\,000$
- note that many AMs (practically all asymptotic tests) are size-invariant

Know Your Numbers

- we usually take a cavalier approach towards numerical accuracy — *at least I do* (i.e., we ignore the issue completely and use standard floating-point arithmetic)
- another example: the `cephes` library of special mathematical functions
  → Perl version includes regression tests, which fail miserably on Solaris 2.8
- theory: Fisher's exact test or binomial test should give most accurate results
  evaluation: performance of Fisher AM breaks down for highest ranks
  (a closer look reveals *negative probabilites* for some candidates!)
- *What Every Computer Scientist Should Know About Floating Point Arithmetic* (Goldberg, 1991)
- easy: **high-precision arithmetic** (e.g. GMP library, `http://www.swox.com/gmp/`)
- more professional: **interval arithmetic** (Kearfott, 1996) → MuPAD 2.5

Yet Another Association Measure

- Aren't there enough yet? Isn't Fisher's exact test the best solution, if we can get the numerics right? Is there room for substantial improvement, or are we just twiddling?
- all statistics-based AMs attempt to measure the same quantity:
  the significance of evidence obtained from the sample against the null hypothesis of independence (random combination of lexemes into pairs)
- is this really the correct translation of Firth's definition into mathematical terms?
- H_0 is rejected for at least half of the candidates, even at $\alpha = 0.001$
- the difference between high-ranking and low-ranking candidates is just that between a very low probability under H_0 and an *incredibly* low one
- suggestion: try different alternative $H_1 : \ \pi \gg \pi_0$ (against $H_0 : \ \pi \leq K \cdot \pi_0$)

Statistics for Fractional Counts

- we are now beginning to obtain **fractional** co-occurrence counts from stochastic grammars (cf. the presentation by Zinsmeister and Heid)
- we can simply insert the fractional counts O_{ij} into AM equations
  (for all AMs based on asymptotic tests)
- however, there is no a-priori **theoretical** justification for this approach,
  which amounts to interpolation between the grid of integer frequencies
  (unproblematic when $O_{ij} \geq 5$, but interpolation for $O_{11} \ll 1$ is just a wild guess)
- the actual **data** from the parser are instances of lexeme pairs,
  each annotated with a **probability** weight = parser's confidence in the analysis
- if these confidence estimates were correct, then among ten instances of a pair (A, B) with weight 0.2 each ($\rightarrow O_{11} = 2.0$) there should *on average* be two correct ones

Statistics for Fractional Counts

- interpret fractional counts as **estimates** for the number of correct instances
  $\rightarrow$ justifies interpolation approach for high-frequency candidates
- a possible interpretation of co-occurrence frequencies $O_{11} < 1$:
  - for a pair (A, B) with $O_{11} = 0.3$, think of an idealised corpus 10 times as large, which contains exactly 10 times as many instances of (A, B) with the same weights
  - in this hypothetical corpus, $O'_{11} = 3.0$, i.e. the parser expects 3 correct instances
  - multiplying the corpus size by 10^k, we can always obtain integer counts
  - relative frequencies p, p_1, p_2 remain *the same* for the hypothetical larger corpus
- an AM g is **size-invariant** iff multiplying all observed frequencies with a constant factor does not change the AM scores (or only by a constant factor):

$$g\big(k \cdot O_{11},\ k \cdot O_{12},\ k \cdot O_{21},\ k \cdot O_{22}\big) = \gamma(k) \cdot g\big(O_{11}, O_{12}, O_{21}, O_{22}\big)$$

- surprisingly, most association measures *are* size-invariant

Agresti, Alan (1990). *Categorical Data Analysis*. John Wiley & Sons, New York.

Dunning, Ted (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.

Firth, J.R. (1957). A synopsis of linguistic theory 1930–55. In *Studies in linguistic analysis*, pages 1–32. Oxford. (special volume of the Philological Society).

Goldberg, David (1991). What every computer scientist should know about floating point arithmetic. *ACM Computing Surveys*, **23**(1), 5–48.

Kearfott, R. Baker (1996). Interval computations: Introduction, uses, and resources. *Euromath Bulletin*, **2**(1), 95–112.

Manning, Christopher D. and Schütze, Hinrich (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Motulsky, Harvey (1995). *Intuitive Biostatistics*. Oxford University Press, New York.

Pedersen, Ted (1996). Fishing for exactness. In *Proceedings of the South-Central SAS Users Group Conference*, Austin, TX.