Information Retrieval

Collocations

Ulf Leser

Content of this Lecture

- Collocations
- Statistical methods for finding collocations
- Case study

- Most material from
  - [MS99], Chapter 5: “Collocations”
  - Schweppe & Broß, FU Berlin, WS 2007/2008
  - Heyer, G., Quasthoff, U. and Wittig, T. (2006). "Text Mining: Wissensrohstoff Text", W3L Verlag.

Co-occurrence

- Two terms co-occur if they appear together in a sentence
  - Also possible: Same paragraph, not more than X words apart, ...
- Simple method for finding relationships between terms
  - If two terms (genes, people, companies etc.) appear in the same sentence, they very like have a relationship to each other
  - The type of relationship very likely is the verb of the sentence
  - The more often we find a specific co-occurrence in a corpus, the stronger the evidence that there is a relationship
  - Almost 100% recall (why not 100%?)
  - Precision depends a lot on the task, anything from 10% to 95%
  - Often used as baseline for relationship extraction

Special Co-Occurrences

- In human languages, some words go together very well
  - Best practice, stiff breeze, Big Blue, Big Apple, ...
  - Strong breeze? Stiff wind? Big green? Big strawberry?
  - Dark night – white night (OK - Dostojewksi) – yellow night?
- How do we know? Google phrase search
  - “big apple”: 4M hits, “big strawberry”: 120K hits
  - “stiff breeze”: 450K, “stiff wind”: 220K
    - But: “wind”: 1000M; “breeze”: 200M; “stiff”: 145M
    - We would expect many more “stiff wind” than “stiff breeze”
  - “Dark/white/yellow night”: 3.2M / 1.2M / 259K

Examples

- Starker Tobak – schwacher Tobak?
- Sinn machen – Sinn haben – Sinn ergeben?
- Es regnet in Strömen – es regnet in Bächen - es regnet in Flüssen?
- Mittleres Management – vorderes Management?
- In der Regel, im allgemeinen, unter anderem, ...
- Take a decision – make a decision?
- Red wine, white wine, blue wine?

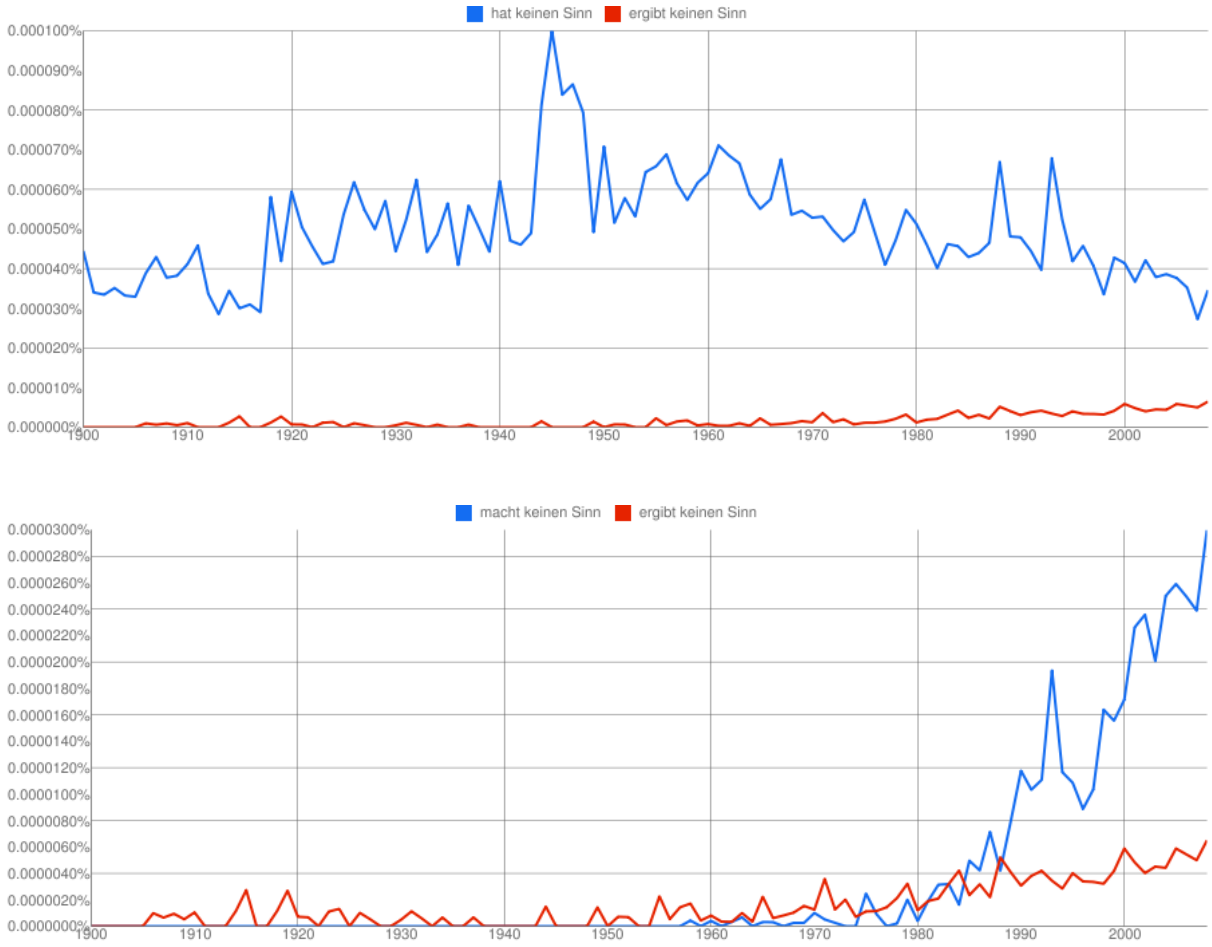
Characterization

- Collocations: Co-occurrences with its own meaning
- Characteristics: Colocations ...
  - ... are accepted combinations of terms
    - “schwacher Tobak” is a semantically correct statement that everybody understands, but it is never used
  - ... have a special meaning or co-notation, close to a “Sprichwort”
    - “Ganzer Kerl”
    - “Eine Leiche im Keller haben” – “To have a skeleton in the closet”
    - “To be hands in gloves with somebody” – “unter einer Decke stecken”
  - ... represent a single, fused concept in our mind
  - ... are very important for speaking a language properly
    - And difficult to be acquired by non-native speakers
  - ... are a constantly changing characteristics of a spoken language

Example

- What is more common – since when?
  - Hat keinen Sinn
  - Ergibt keinen Sinn
  - Macht keinen Sinn

Example



Source: Google's n-gram viewer, many books from 1900-2008: http://ngrams.googlelabs.com/

NLP

- Definitions from NLP research
- “A collocation is an expression consisting of two or more words that correspond to some conventional way of saying things” [MS99]
- “Collocations of a word are statements of the habitual or customary places of that word” [Firth, 1957]

Types of Collocations

- Collocations include
  - Proper names (New York)
  - Fixed verb – noun constructions (take a decision)
  - Terminological expressions (data model, text mining)
  - Associative collocations (Hospital doctor, university member)
  - ...
- Legal text is full of formalized collocations with mystic meaning
  - “Abschlussarbeiten werden in der Regel von zwei Prüfern begutachtet.” (?)

Application: Understanding Semantic Differences

- Frequent co-occurrences of “strong” and of “powerful”
- Lists are disjoint
- Hint to subtle semantic differences
- Listing accepted collocations is one of the best “explanations” for such differences
  - Distributional semantics
  - Current trend: Word embeddings

TERM1	TERM2	Freq
strong	US	46123
strong	I	39807
strong	Christian	28577
strong	European	18188
strong	And	15555
strong	Q4	13300
strong	R	12955
strong	American	12283
strong	IT	12021
strong	AI	10991
powerful	Speaker	48618
powerful	Web	36161
powerful	DVD	30215
powerful	Windows	23368
powerful	HRMS	21987
powerful	Business	20400
powerful	Internet	20321
powerful	PC	20233
powerful	God	19555
powerful	FTP	19513

Content of this Lecture

- Collocations
- Statistical methods for finding collocations
  - Bi-Gram frequencies
  - Word distance
  - Hypothesis testing
- Case study

Statistical Approach I: Counting Frequencies

- Obviously, we should find “white wine” much more often in a corpus than “black wine”
- First approach: Count bi-grams
- Google LDC corpus
  - Tokens: 1,024,908,267,229
  - Sentences: 95,119,665,584
  - Unigrams: 13,588,391
  - Bigrams: 314,843,401
- Just appearing together frequently is not a reliable indication for a colocation
  - Beware: Collocations need not be continuous

w1	w2	cnt(w1,w2)
Contact	Us	198887927
United	States	173331792
Privacy	Policy	161052207
New	York	153457830
Site	Map	111486987
...		
Shopping	Cart	33662959
It	Now	32652561
Web	Site	32482311
OF	THE	32013260
In	Stock	30534425

Three Tricks for Getting Rid of Boring Bi-grams

- Look at Part-of-Speech tags
  - In collocations, the combinations of POS tags are fairly restricted
- Look at distribution of distances
  - Collocations have preferred distances in sentences
- Consider frequency of constituent words
  - “of the” not surprising, because both words are very frequent
  - “Privacy policy” is surprising, because both words are rather rare
  - We need to quantify “surprisingness”

POS Tagging

- Simple tag set
  - The/**D** koala/**N** put/**V** the/**D** keys/**N** on/**P** the/**D** table/**N**
- Including morphological information
  - The/**D** koala/**Ns** put/**V-past-3rd** the/**D** keys/**N-p** on/**P** ...
- Using Penn tag set
  - The/**DT** koala/**NN** put/**VBN** the/**DT** keys/**NNS** on/**P** ...

The	koala	put	the	keys	on	the	table
D	N	V	D	N	P	D	N
D	N-sing	V-past-3rd	D	N-plu	P	D	N-sing
DT	NN	VBN	DT	NNS	P	DT	NN

Brown Tag Set

- Has 87 tags in total
  - Table: Most important tags
- Definition of classes is not at all fixed
  - London-Lund Corpus of Spoken English: 197 tags
  - Lancaster-Oslo/ Bergen: 135 tags
  - U-Penn: 45 tags

DT	Determiner
IN	Preposition
JJ	Adjective
NN	Noun, singular
NNP	Proper Noun
NNS	Noun, plural
PERIOD	„.“, „?“, „!“
PN	Personal pronoun
RB	Adverb
TO	„to“
VB	Verb, base form
VBZ	Verb, 3d singular present
VBD	Verb, past tense
WDT	Wh – determiner
...	...

Using POS Tags

- Allow as collocations only a small set of POS-tag pairs [Justeson, Katz, 1995]
  - ADJ NN (linear function)
  - NN NN (Regression coefficient)
  - ADJ ADJ NN (Gaussian random variable)
  - NN ADJ NN (mean squared error)
  - ...
- Result: The combination of (bi-gram) frequency and POS filtering works quite well

Contact	Us	JJ	PN	198887927
Privacy	Policy	NN	NN	161052207
Site	Map	NN	NN	111486987
...				
It	Now	PR	RB	32652561
Web	Site	NN	NN	32482311
OF	THE	IN	DT	32013260
In	Stock	IN	NN	30534425

Problems with Bi-Grams

- Bi-Gram counting is restricted to consecutive collocations
- What about more distant collocations?
  - You knock on a door; you don't "beat a door" or "hit a door"
  - Thus, "knock" and "door" are a collocation
  - But they never appear directly after each other
  - "Knock the door, please"
  - "She knocked on his door"
  - "They knocked at the door"
  - "She knocked on Peters door"
  - "She knocked on the black, large and metal door"

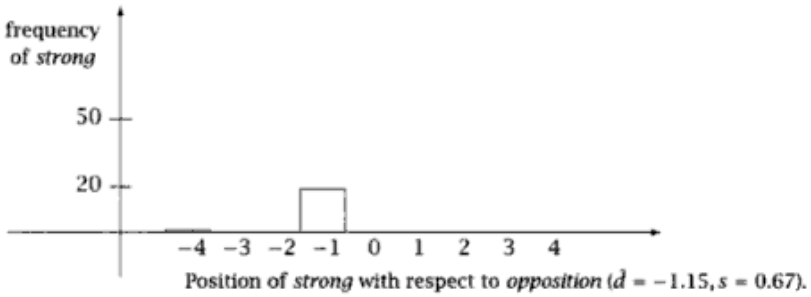
Relaxed-Bi-Gram Definition

- Option 1: Relax bi-gram definition
  - Slide a window of size t over the text
  - Within t, count all pairs of words (in whatever distance and order)
  - Example (t=4)
    - (she knocked), (she on), (she his), (knocked on), (knocked his), (knocked door), (on his), (his door), ...
    - Counts: (knock door 3), (she on 3), (on door 2), ...
    - But we will not find (hit door)
  - A bit arbitrary; which t should we chose?
- Option 2: Analyze distances between words
  - Often, words in a collocation have a somewhat constant distance
  - Characteristic distance depends on the specific collocation

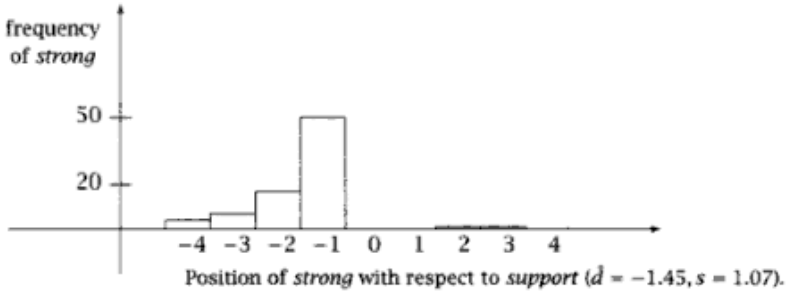
Word Distances

- Idea: Count for a given pair of words
  - All distances of both in the same sentence
  - Compute mean and variance
- What do we expect?
  - Collocations should have a small mean and a small variance
  - Small mean: Collocations usually are local (<5 words)
  - Small variance: Expression must be fairly stable (by definition)
- Example
  - “She knocked on his door”, “They knocked at the door”, “She knocked on Peters door”, “She knocked on the black, large and metal door”
  - ∅(knock, door) = 12/4 = 3; var(knock, door) = 3
  - A counter-example: s4 is “too strange” (and certainly very rare)

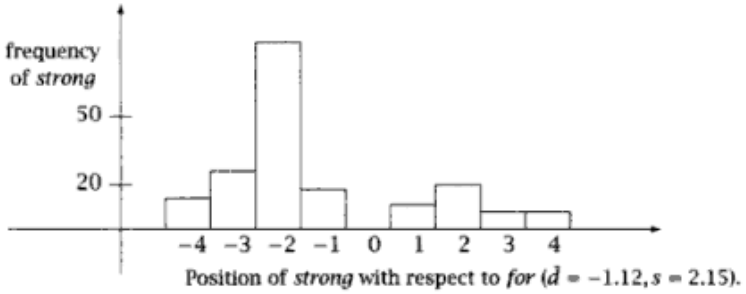
Frequency Histograms



Clear collocation



Collocation



No collocation

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

Source: [MS99]

Variance and Mean

Mean s and variance d

s	$\hat{d}$	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 $\hat{d}$ of the distances between 12 word pairs.

- Small mean, small variance: Collocation
- Small mean, large variance: No collocation
- Large mean, even with small variance: No collocation
- Small mean, medium variance: In between

Comparison

- Counting bi-grams only works for bi-grams
  - Combined with POS-pair filtering, results are acceptable
- Using sliding window and/or mean/variance vastly increases search space, but also improves accuracy
  - Sliding window: Many more pairs of words
  - Word distance: We either need to know what we are looking for, or we need to test all word pairs in each sentence
- Many variations
  - Count bi-grams with gaps
  - Let gap length vary slightly

Content of this Lecture

- Collocations
- Statistical methods for finding collocations
  - Bi-Gram frequencies
  - Word distance
  - Hypothesis testing
- Case study

Surprising Collocations

- Recall the problem of boring bi-grams
- The core of the problem
  - Pairs of frequent words are frequent just by chance
  - Frequently finding pairs of frequent words is not surprising
- How can we measure the "surprisingness" of a bi-gram?
  - Given the frequencies of the words and the size of a corpus?
  - Beware: If the corpus is "large enough", many words become somewhat frequent
- Solution: Statistical test

w1	w2	cnt(w1, w2)
Contact	Us	198887927
United	States	173331792
Privacy	Policy	161052207
New	York	153457830
Site	Map	111486987
...		
Shopping	Cart	33662959
It	Now	32652561
Web	Site	32482311
OF	THE	32013260
In	Stock	30534425

Statistical Tests

- Statistical test: Assess the probability that a certain value has been generated by chance or not
- Approach: Probability of null hypothesis
  - Null hypothesis H_0: w_1, w_2 statistically independent: $c=p(w_1)*p(w_2)$
  - Compute probability p of the observed count assuming H_0
  - Refute H_0, if p is too small, e.g. p <= 0,05 (=5%)
  - Application: If w1,w2 are not statistically independent, assume a co-location

Example

- Before looking at co-locations, we test something simpler
- We measure the height of persons and reason about their mean value
- Example
  - Assume H_0: Mean height in a given population is d=158
  - In a sample N=100, we observe d'= 160, variance s'=2,6
  - Given this sample, how likely is it that H_0 is true?
  - Depends on the expected distribution of values given the mean
  - Underdetermined: Need an additional assumption
  - We know that height is not equally distributed (in range [0;250])
  - Height is normally distributed

Naive Approach: Bootstrapping (“Just Try”)

- Generate normally distributed values according to H0 very often and see how often this yields the observed mean
  - H0: Height is normal distributed with mean d=158
  - Underdetermined: We also need variance s
  - Trick: Assume that variance in sample and reality is equal
- Operations
  - Generate 100 values drawing from normal distr. with d, s
  - Compute mean d’’ of sample
  - Repeat 10.000 times (or more)
  - How often was d’’=d’?
- Problem: Very slow
- We need a test that is independent of d and s

Frequently Used: t-Test

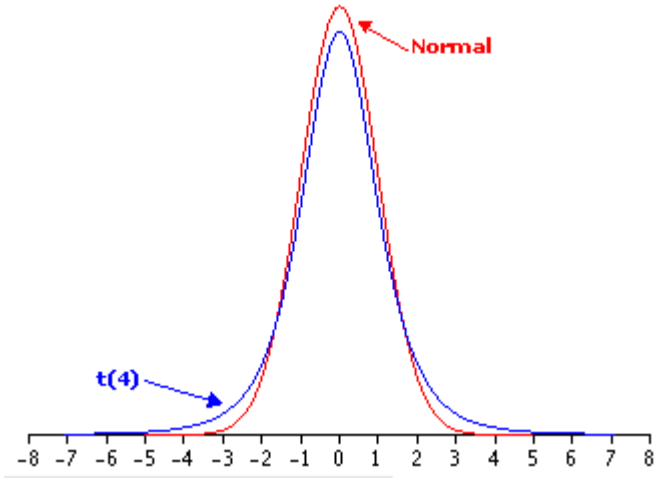
- Statistical test for samples of a normal distribution
  - d: distribution mean, s: distribution variance (often unknown)
  - d': sample mean, s': sample variance
  - N: sample size
- Without further knowledge, we use s' as estimate for s
- We compute the t-value, a measure for the deviation in the mean (d'-d) given the variance s

$$t = \sqrt{N}\,\frac{d'-d}{\sqrt{s}}$$

  - Large s: Differences are less significant
  - Large N: Differences become more and more significant

Meaning of a t-Value

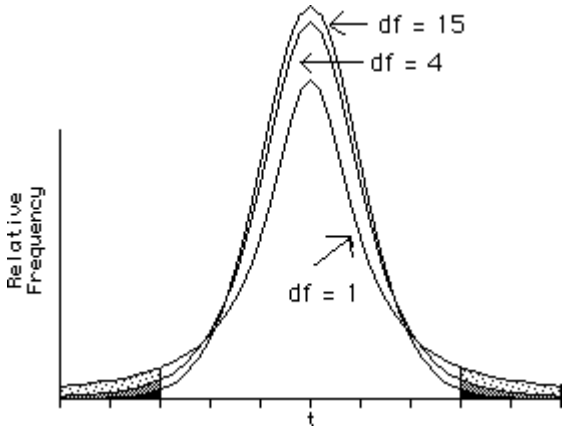
- Assume a normally distributed set X of values
- Compute mean d and variance s
- Now do the following very often
  - Sample N values at random from X
  - Compute sample mean d' and variance s'
  - Compute t-value
- Gives a distribution of t-values: The t-distribution
  - Similar, but not identical to normal distribution
  - Depends on N



Source: http://davidmlane.com/hyperstat/A48339.html

Application

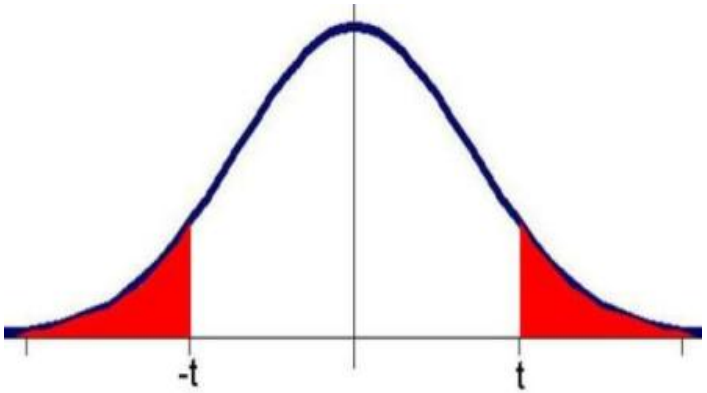
- We can assess the probability of a given t-value by looking at a pre-computed distribution of t-Values
  - Dependent on N: Degree-of-freedom
- Table gives the probability that a given t-Value has emerged by chance (given N)



One Sided	75%	80%	85%	90%	95%	97.5%	99%	99.5%	99.75%	99.9%	99.95%
Two Sided	50%	60%	70%	80%	90%	95%	98%	99%	99.5%	99.8%	99.9%
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015

Source: Wikipedia

One Sided, two sided



- p-Value of t-value t: Probability of a value from the t-distribution being absolutely larger than abs(t)
- Together
  - Compute t-value, lookup p-value: The probability of H_0 being wrong
  - Refute H_0 if p is too large (compared to your favorite threshold)
- One sided: Prob. of the mean being greater than expected
- Two sided: Prob. of the mean being away from expected

Example

- H_0: mean height of some population is 158
- In a sample of N=100, we observe d'= 160, s=s'=2.6
- We get a value of t ~ 12,4 > 3.17
- For N=100, 3.17 corresponds to a significance level of p=0.002
  - Smallest p for which precomputed t is smaller than observed t
- Thus, H_0 can be rejected with >99,8% confidence

$$t = \sqrt{N}\,\frac{d'-d}{\sqrt{s}}$$

App for Co-Locations: Preparatory Work

- H_0: w_1, w_2 are statistically independent
- We expect $p_{ind}=p(w_1,w_2) = p(w_1)*p(w_2)$
- This count p_{ind} is normally distributed
  - Consider the experiment of drawing very often N bi-grams randomly, where (w_1,w_2) appears with a relative frequency of p_{ind}, and each appearance of (w_1, w_2) is counted as 1, all others are counted as 0
  - This is a Bernoulli trial, creating a normal distribution of counts
  - The mean of this distribution is $p_{ind}*N$, its variance is $s=p_{ind}*(1-p_{ind})$
  - Since p_{ind} will be very small, we may assume $s=p_{ind}*(1-p_{ind}) \sim p_{ind}$

Application

- We apply the t-test to collocations
- Set N= Number of bi-grams in corpus
- Set d=p_{ind}, the expected relative frequency (given H_0)
- Set s=p_{ind}
- Set d'=count(w_1, w_2)/N
- Set s'=s
  - Again, assuming equal variance in sample and distribution
- Compute t-Value, set your threshold, refute/accept H_0

Example

- Consider the term “new company”
  - Assume it appears 8 times in a corpus of N=14,307,668 bi-grams
  - Assume count(new)=12,828, count(company)=4,675
- Under H_0: d=count(new)*count(company)/N^2 ~ 2.93E-7
- The observed relative frequency is d’ = 8/N ~ 5.59E-7
- t-value

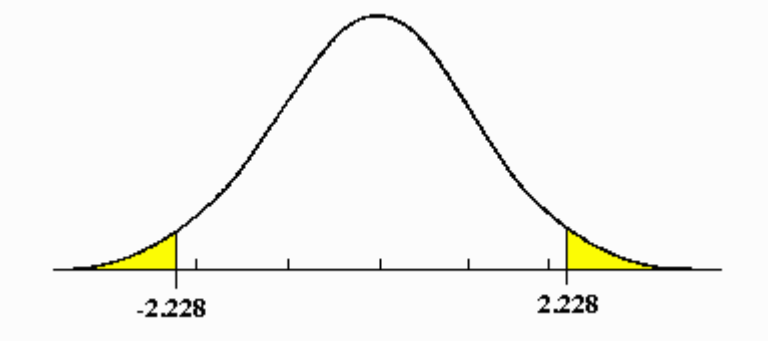
$$t = \frac{d' - d}{\sqrt{s/N}} = \frac{5.59E^{-7} - 2.93E^{-7}}{\sqrt{5.59E^{-7}/1.43E^{7}}} \sim 1.35$$

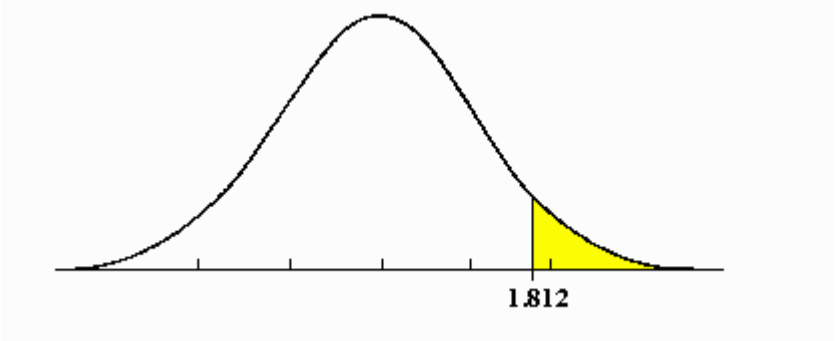
- p-value around 0.1
- H_0 should rather not be refuted
  - “new company” is not a collocation but probably occurs in this corpus that often by chance

One Sided	75%	80%	85%	90%	95%	97.5%	99%	99.5%	99.75%	99.9%	99.95%
Two Sided	50%	60%	70%	80%	90%	95%	98%	99%	99.5%	99.8%	99.9%
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

One more Detail

- Two-sided t-test: Probability that the absolute of a given t-Value is created by chance (given N)
- Single-sided t-test: Probability that any t-value larger than the given t-Value is created by chance (given N)
- We need to apply the single sided test: We are not looking for “negative collocations” ~ words co-occurring must less often than expected by chance
- Computation: Simply divide p-value by 2





Discussion

- [MS99]: Out of 831 bi-grams which occurred >20 times, H_0 is rejected for 824 (p=0.05)
- Thus, 824 pairs (~all) should be considered as collocations
- Many pairs of words are surprisingly (and significantly) frequent
- This is a property of language, because only very few pairs actually occur (and those rather often)
- Independence assumption is no good candidate for H_0
  - This assumption will be refuted too often
- t-test still useful for ranking potential collocations

t	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	w^1	w^2
4.4721	42	20	20	Ayatollah	Ruhollah
4.4721	41	27	20	Bette	Midler
4.4720	30	117	20	Agatha	Christie
4.4720	77	59	20	videocassette	recorder
4.4720	24	320	20	unsalted	butter
2.3714	14907	9017	20	first	made
2.2446	13484	10570	20	over	many
1.3685	14734	13478	20	into	them
1.2176	14093	14776	20	like	people
0.8036	15019	15629	20	time	last

ble 5.6 Finding collocations: The t test applied to 10 bigrams that occur w
quency 20.

Multiple Testing

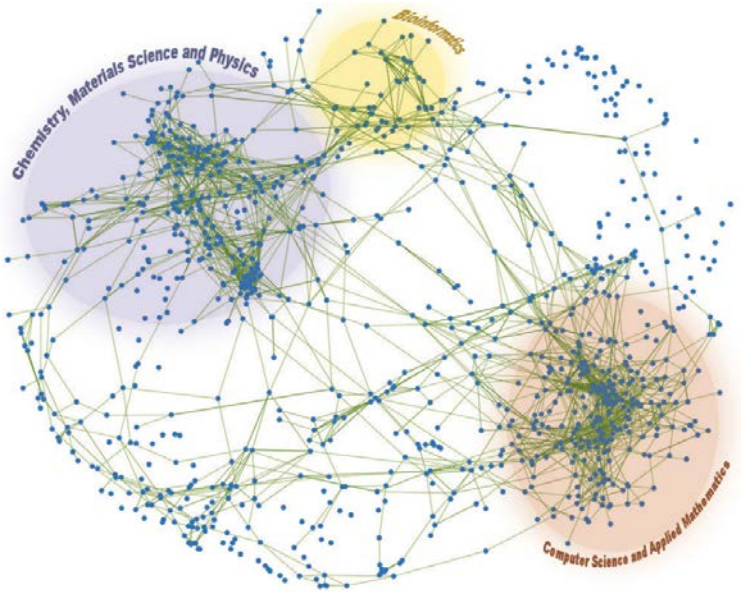
- Given threshold p=0.05; are all word pairs in a corpus of 100M different bigrams with p-value smaller k collocations?
  - Every single test as an error probability of up to 0.05
  - We performed 100M such tests
  - Thus, approximately 100M*p=5M of the tests went wrong
  - Many collocations are false positives, i.e., stem from bigram frequencies that probabily occurred by chance only
- We need multiple testing correction
  - Whenever many tests are performed, results of statistical tests must be corrected
  - The more urgent, the more liberal the threshold is chosen
  - Simplest method: Divide threshold by N

Testing Collocations Empirically

- How can we empirically test whether a word pair should be considered as a collocation?
- Stimulus-Response Test
  - Give a set of persons one of the words
  - Let them, very quickly, write down words that come into their mind first when they hear the first word
- Good methods for collocations perform surprisingly well
  - Ranking by t-value yields similar top-K collocations as stimulus-response tests
- But: One is usually interested in finding new (rare) collocations, i.e., those that do not come to mind first
  - To learn about language use, language evolution, etc.

Co-Occurrence Graphs

- Co-occurrences can be visualized nicely
  - Layout: Bring (Euclidian) distances close to semantic distances
- Clusters in the graph usually form semantically close topics
- Applications
  - Learn about a domain
  - Disambiguation of senses
  - Detection of synonyms
- Properties
  - Small world
  - Distribution of the degrees of the nodes is Zipf
- Also true for "human" assoc-graphs



Source: Luis Rocha, U Indiana

Selbsttest

- What is a colocation? Give examples
- Name three ways to find colocations
- A t-test produces a p-value of 0,12 for a certain result of an experiment. What does this mean?
- What are assumptions of a t-test (which are not tested)
- Why is the independence assumption inherent in our application of t-tests to colocation analysis wrong?

Content of this Lecture

- Definition of collocations
- Statistical methods for finding collocations
- Case study: Learning a Terminology and an Ontology
  - Defining a Phenotype Terminology
  - Learning a Phenotype Ontology

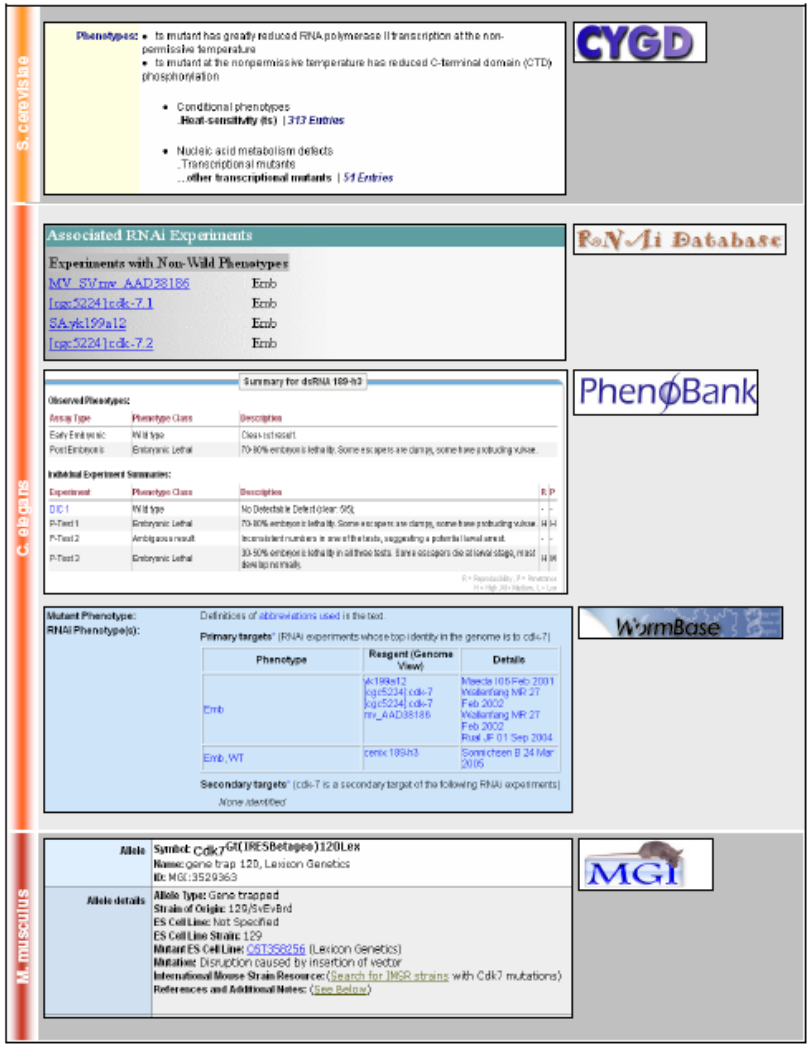
From Phenotype to Function

(Groth et al. 2008, Böhm et al. 2009, Groth at al. 2010a, Groth et al. 2010b)

Phenotypes

- Observable characteristic of an organism
  - Description of a disease
  - Response to a drug
  - ...
- A “phenotype” usually is a derivation from the norm
- Small-scale experiments to measure phenotypes since long
- Systematic experimental approaches only for some years
  - Systematic perturbation of genotypes or the environment => effect on phenotype
  - Natural mutations, breeding, knock-out, RNAi

Describing Phenotypes



- Many data sources
- Different species, different experiments, different vocabulary, different format, different ...
- Technical integration is a challenge
- Semantic integration is much more of a challenge
- Least common denominator: Text

Motivation

- Building ontologies manually is costly
- Ontology bootstrapping
  Automatically building a first draft of an ontology by analyzing a domain-specific corpus

- Four steps
  - Concept discovery – concepts of the ontology
  - Concept matching – occurrences of concepts
  - Relationship extraction – relationships between concepts
  - Ontology extraction – a “good” subset of all relationships

From Text to Ontology

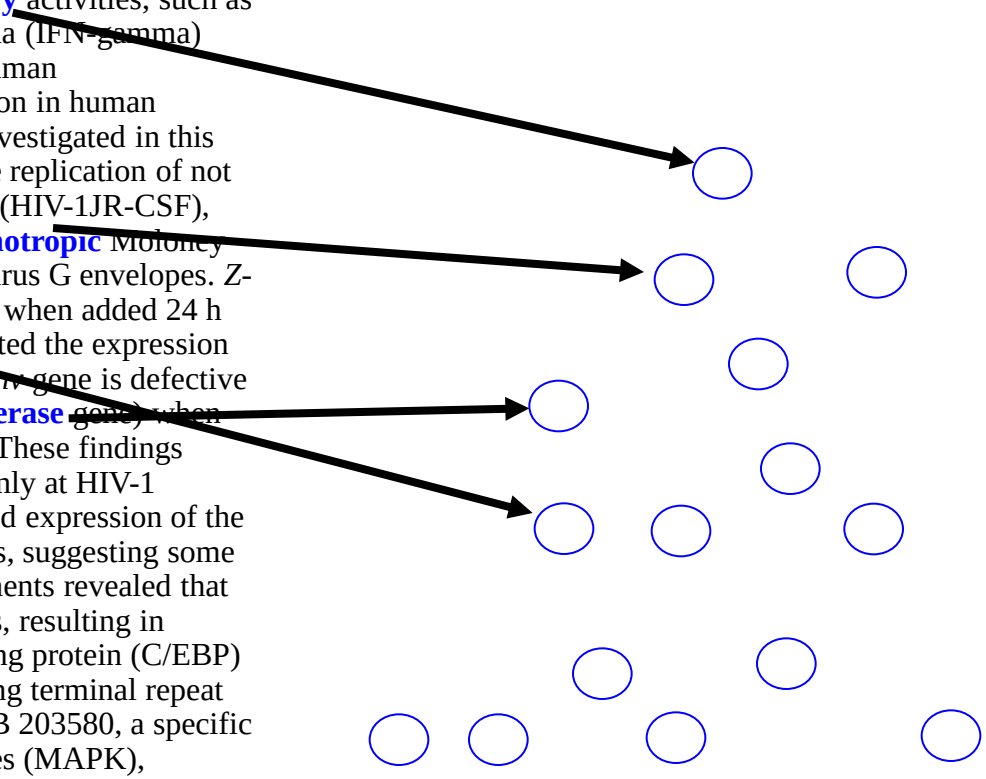
Z-100 is an *arabinomannan* extracted from Mycobacterium tuberculosis that has various immunomodulatory activities, such as the induction of interleukin 12, interferon gamma (IFN-gamma) and beta-chemokines. The effects of *Z-100* on human immunodeficiency virus type 1 (HIV-1) replication in human monocyte-derived macrophages (MDMs) are investigated in this paper. In MDMs, *Z-100* markedly suppressed the replication of not only macrophage-tropic (M-tropic) HIV-1 strain (HIV-1JR-CSF), but also HIV-1 pseudotypes that possessed amphotropic Moloney murine leukemia virus or vesicular stomatitis virus G envelopes. *Z-100* was found to inhibit HIV-1 expression, even when added 24 h after infection. In addition, it substantially inhibited the expression of the pNL43lucDeltaenv vector (in which the *env* gene is defective and the *nef* gene is replaced with the *firefly luciferase* gene) when this vector was transfected directly into MDMs. These findings suggest that *Z-100* inhibits virus replication, mainly at HIV-1 transcription. However, *Z-100* also downregulated expression of the cell surface receptors CD4 and CCR5 in MDMs, suggesting some inhibitory effect on HIV-1 entry. Further experiments revealed that *Z-100* induced IFN-beta production in these cells, resulting in induction of the 16-kDa CCAAT/enhancer binding protein (C/EBP) beta transcription factor that represses HIV-1 long terminal repeat transcription. These effects were alleviated by SB 203580, a specific inhibitor of p38 mitogen-activated protein kinases (MAPK), indicating that the p38 MAPK signalling pathway was involved in *Z-100*-induced repression of HIV-1 replication in MDMs. These findings suggest that *Z-100* might be a useful immunomodulator for control of HIV-1 infection.

1. Define Set of Terms

Z-100 is an *arabinomannan* extracted from Mycobacterium tuberculosis that has various **immunomodulatory** activities, such as the **induction** of interleukin 12, interferon gamma (IFN-gamma) and beta-chemokines. The effects of *Z-100* on human immunodeficiency virus type 1 (HIV-1) replication in human monocyte-derived **macrophages** (MDMs) are investigated in this paper. In MDMs, *Z-100* markedly suppressed the replication of not only macrophage-tropic (M-tropic) HIV-1 strain (HIV-1JR-CSF), but also HIV-1 pseudotypes that possessed **amphotropic** Moloney **murine leukemia** virus or vesicular stomatitis virus G envelopes. *Z-100* was found to inhibit HIV-1 expression, even when added 24 h after infection. In addition, it substantially inhibited the expression of the pNL43lucDeltaenv **vector** (in which the *env* gene is defective and the *nef* gene is replaced with the *firefly* **luciferase** gene) when this vector was transfected directly into MDMs. These findings suggest that *Z-100* inhibits virus replication, mainly at HIV-1 transcription. However, *Z-100* also downregulated expression of the cell **surface receptors** CD4 and CCR5 in MDMs, suggesting some inhibitory effect on HIV-1 entry. Further experiments revealed that *Z-100* induced IFN-beta production in these cells, resulting in induction of the 16-kDa CCAAT/enhancer binding protein (C/EBP) beta transcription factor that **represses** HIV-1 long terminal repeat transcription. These effects were alleviated by SB 203580, a specific inhibitor of p38 mitogen-activated protein kinases (MAPK), indicating that the p38 MAPK signalling **pathway** was involved in *Z-100*-induced repression of HIV-1 replication in MDMs. These findings suggest that *Z-100* might be a useful **immunomodulator** for control of HIV-1 infection.

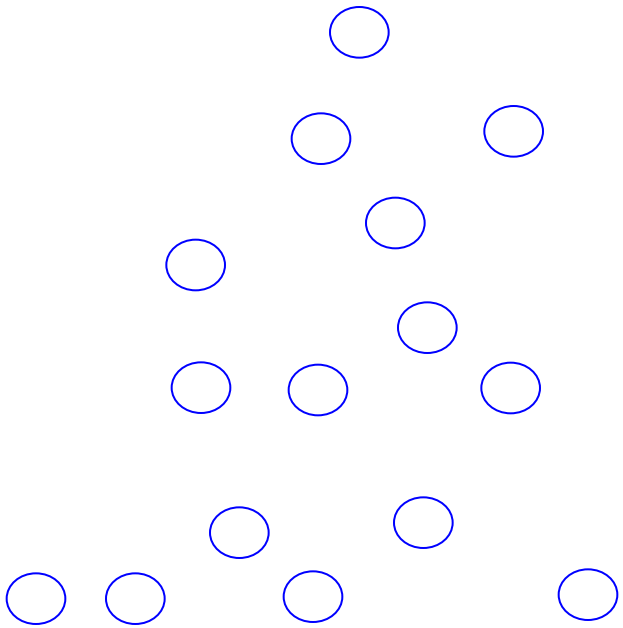
1. Define Set of Terms

Z-100 is an *arabinomannan* extracted from Mycobacterium tuberculosis that has various **immunomodulatory** activities, such as the **induction** of interleukin 12, interferon gamma (IFN-gamma) and beta-chemokines. The effects of *Z-100* on human immunodeficiency virus type 1 (HIV-1) replication in human monocyte-derived **macrophages** (MDMs) are investigated in this paper. In MDMs, *Z-100* markedly suppressed the replication of not only macrophage-tropic (M-tropic) HIV-1 strain (HIV-1JR-CSF), but also HIV-1 pseudotypes that possessed **amphotropic** Moloney **murine leukemia** virus or vesicular stomatitis virus G envelopes. *Z-100* was found to inhibit HIV-1 expression, even when added 24 h after infection. In addition, it substantially inhibited the expression of the pNL43lucDeltaenv **vector** (in which the *env* gene is defective and the *nef* gene is replaced with the *firefly* **luciferase** gene) when this vector was transfected directly into MDMs. These findings suggest that *Z-100* inhibits virus replication, mainly at HIV-1 transcription. However, *Z-100* also downregulated expression of the cell **surface receptors** CD4 and CCR5 in MDMs, suggesting some inhibitory effect on HIV-1 entry. Further experiments revealed that *Z-100* induced IFN-beta production in these cells, resulting in induction of the 16-kDa CCAAT/enhancer binding protein (C/EBP) beta transcription factor that **represses** HIV-1 long terminal repeat transcription. These effects were alleviated by SB 203580, a specific inhibitor of p38 mitogen-activated protein kinases (MAPK), indicating that the p38 MAPK signalling **pathway** was involved in *Z-100*-induced repression of HIV-1 replication in MDMs. These findings suggest that *Z-100* might be a useful **immunomodulator** for control of HIV-1 infection.



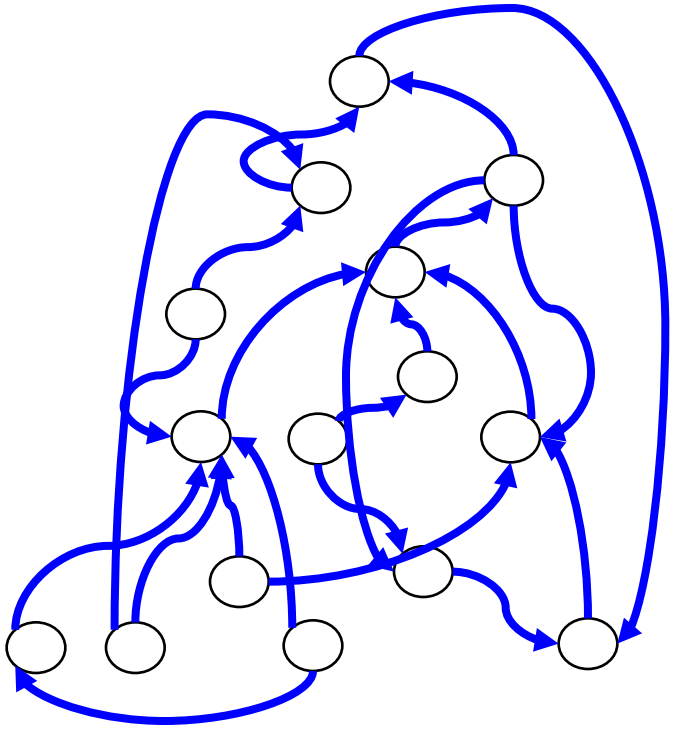
2. Find all Occurrences of those Terms

Z-100 is an *arabinomannan* extracted from Mycobacterium tuberculosis that has various **immunomodulatory** activities, such as the **induction** of interleukin 12, interferon gamma (IFN-gamma) and beta-chemokines. The effects of *Z-100* on human **immunodeficiency** virus type 1 (HIV-1) replication in human monocyte-derived **macrophages** (MDMs) are investigated in this paper. In MDMs, *Z-100* markedly **suppressed** the replication of not only **macrophage-tropic** (M-tropic) HIV-1 strain (HIV-1JR-CSF), but also HIV-1 pseudotypes that possessed **amphotropic** Moloney **murine leukemia** virus or vesicular stomatitis virus G envelopes. *Z-100* was found to **inhibit** HIV-1 expression, even when added 24 h after infection. In addition, it substantially **inhibited** the expression of the pNL43lucDeltaenv **vector** (in which the *env* gene is defective and the *nef* gene is replaced with the *firefly* **luciferase** gene) when this vector was transfected directly into MDMs. These findings suggest that *Z-100* inhibits virus replication, mainly at HIV-1 transcription. However, *Z-100* also downregulated expression of the cell **surface receptors** CD4 and CCR5 in MDMs, suggesting some **inhibitory** effect on HIV-1 entry. Further experiments revealed that *Z-100* induced IFN-beta production in these cells, resulting in induction of the 16kDa **CCAAT/enhancer** binding protein (C/EBP) beta transcription factor that **represses** HIV-1 long terminal repeat transcription. These effects were alleviated by SB 203580, a specific **inhibitor** of p38 mitogen-activated protein kinases (MAPK), indicating that the p38 MAPK signalling **pathway** was involved in *Z-100*-induced **repression** of HIV-1 **replication** in MDMs. These findings suggest that *Z-100* might be a useful **immunomodulator** for control of HIV-1 infection.



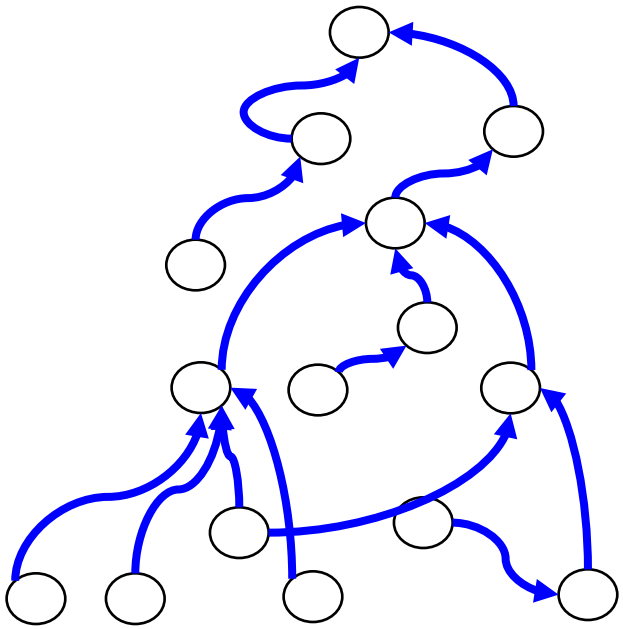
3. Find Relationships between Terms

Z-100 is an *arabinomannan* extracted from Mycobacterium tuberculosis that has various **immunomodulatory** activities, such as the **induction** of interleukin 12, interferon gamma (IFN-gamma) and beta-chemokines. The effects of *Z-100* on human **immunodeficiency** virus type 1 (HIV-1) replication in human monocyte-derived **macrophages** (MDMs) are investigated in this paper. In MDMs, *Z-100* markedly **suppressed** the replication of not only **macrophage-tropic** (M-tropic) HIV-1 strain (HIV-1JR-CSF), but also HIV-1 pseudotypes that possessed **amphotropic** Moloney **murine leukemia** virus or vesicular stomatitis virus G envelopes. *Z-100* was found to **inhibit** HIV-1 expression, even when added 24 h after infection. In addition, it substantially **inhibited** the expression of the pNL43lucDeltaenv **vector** (in which the *env* gene is defective and the *nef* gene is replaced with the *firefly* **luciferase** gene) when this vector was transfected directly into MDMs. These findings suggest that *Z-100* inhibits virus replication, mainly at HIV-1 transcription. However, *Z-100* also downregulated expression of the cell **surface receptors** CD4 and CCR5 in MDMs, suggesting some **inhibitory** effect on HIV-1 entry. Further experiments revealed that *Z-100* induced IFN-beta production in these cells, resulting in induction of the 16kDa **CCAAT/enhancer** binding protein (C/EBP) beta transcription factor that **represses** HIV-1 long terminal repeat transcription. These effects were alleviated by SB 203580, a specific **inhibitor** of p38 mitogen-activated protein kinases (MAPK), indicating that the p38 MAPK signalling **pathway** was involved in *Z-100*-induced **repression** of HIV-1 **replication** in MDMs. These findings suggest that *Z-100* might be a useful **immunomodulator** for control of HIV-1 infection.



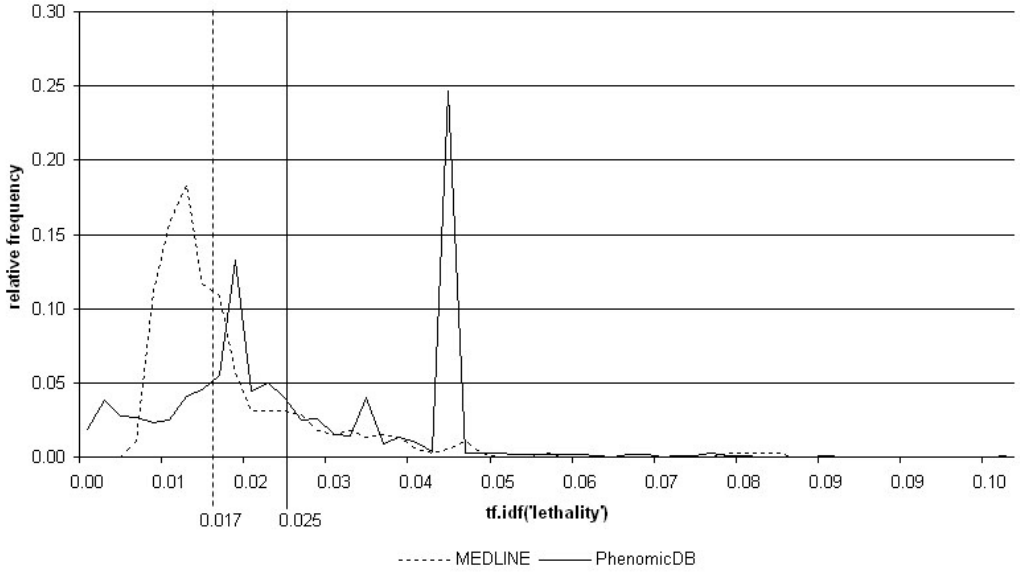
4. Extract a Nice and Consistent Ontology

Z-100 is an *arabinomannan* extracted from Mycobacterium tuberculosis that has various **immunomodulatory** activities, such as the **induction** of interleukin 12, interferon gamma (IFN-gamma) and beta-chemokines. The effects of *Z-100* on human **immunodeficiency** virus type 1 (HIV-1) replication in human monocyte-derived **macrophages** (MDMs) are investigated in this paper. In MDMs, *Z-100* markedly **suppressed** the replication of not only **macrophage-tropic** (M-tropic) HIV-1 strain (HIV-1JR-CSF), but also HIV-1 pseudotypes that possessed **amphotropic** Moloney **murine leukemia** virus or vesicular stomatitis virus G envelopes. *Z-100* was found to **inhibit** HIV-1 expression, even when added 24 h after infection. In addition, it substantially **inhibited** the expression of the pNL43lucDeltaenv **vector** (in which the *env* gene is defective and the *nef* gene is replaced with the *firefly* **luciferase** gene) when this vector was transfected directly into MDMs. These findings suggest that *Z-100* inhibits virus replication, mainly at HIV-1 transcription. However, *Z-100* also downregulated expression of the cell **surface receptors** CD4 and CCR5 in MDMs, suggesting some **inhibitory** effect on HIV-1 entry. Further experiments revealed that *Z-100* induced IFN-beta production in these cells, resulting in induction of the 16kDa **CCAAT/enhancer** binding protein (C/EBP) beta transcription factor that **represses** HIV-1 long terminal repeat transcription. These effects were alleviated by SB 203580, a specific **inhibitor** of p38 mitogen-activated protein kinases (MAPK), indicating that the p38 MAPK signalling **pathway** was involved in *Z-100*-induced **repression** of HIV-1 **replication** in MDMs. These findings suggest that *Z-100* might be a useful **immunomodulator** for control of HIV-1 infection.



Step 1: What is a „Phenotypic“ Term?

- Build a “phenotype” corpus and a “normal” corpus
- Look at each term occurring in both corpora
- Compute TF*IDF values of each term in each doc
- Compare the distributions of TF*IDF values across documents of each term in both corpora (A, B)



Comparing Distributions

- When are two distributions significantly different?
  - No t-test: We look at the entire distributions, not just the means
- Alternative: Two sample Wilcoxon Rank Sum Test
  - Non-parametric test – does not assume any value distributions
  - Decides with which probability two distributions are equal
  - To this end, it sorts all values (of both distributions) and computes the sum of the ranks of each corpus
  - If both samples are from the same distribution, these sums follow a pre-computable distribution
  - We can lookup the probability of the computed sum to be generated from this distribution
  - This defines a p-value for H_0: A=B

Multi-Token Terms

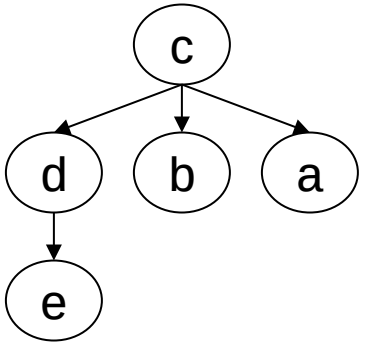
- The previous method only works for single-token terms
- Finding multi-token (composed) concepts (here for n=2)
  - Count frequencies of both terms
  - Count frequency of combined concept
  - (Very debatable) filter: Only consider composed terms consisting only of phenotypic terms
  - Test for statistical independence
  - Test defines a ranking of composed terms
  - We used the first 3.000 composed terms

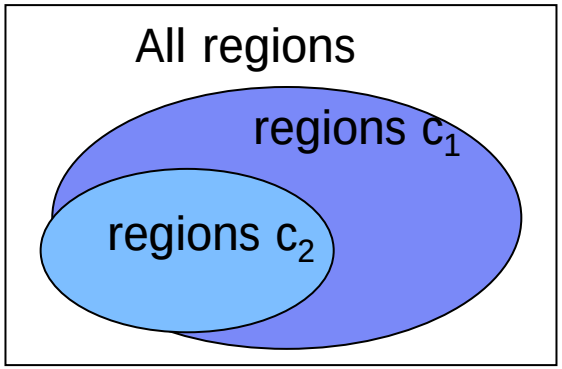
Phenotypic Concepts

- “significant defects”
- “spindel elongation”
- “mutant phenotype growth”
- But: Occurrences in text
  - „We observed a *significant* genomic *defect* in ...“
  - „*Elongation* of the *spindel* correlated with ...“
  - „*Mutant growth* was normal compared to ...“
- Interspersed token, missing token, re-order, spelling variations, ...
- Avg. concept length in MPO is 3.5, ~5% single token

3. Relationship Extraction

- Goal: Infer that
  - cancer ISA disease
  - early abort ISA abort
- Various proposals in the literature
  - Subsumption, Hearst-Pattern, ...
- Subsumption
  - For every pair of concepts c_1, c_2, compute $p(c_1|c_2)$
    - How often do we see an occurrence of c_1 in the neighborhood of an occurrence of c_2?
  - $p(c_1|c_2)>t$ => c_2 is a specialization of c_1



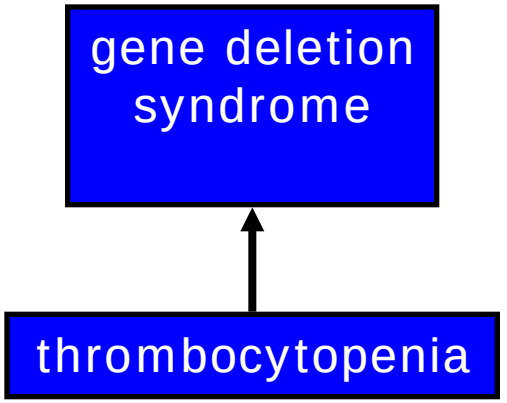


Example

Phenotype description

A number sign (#) is used with this entry because it represents a contiguous gene deletion syndrome. See 274000 for another contiguous gene deletion syndrome, thrombocytopenia-absent radius (TAR) syndrome, that maps to a nonoverlapping region of chromosome 1q21.1.

Gene deletion syndromes also play an important role …

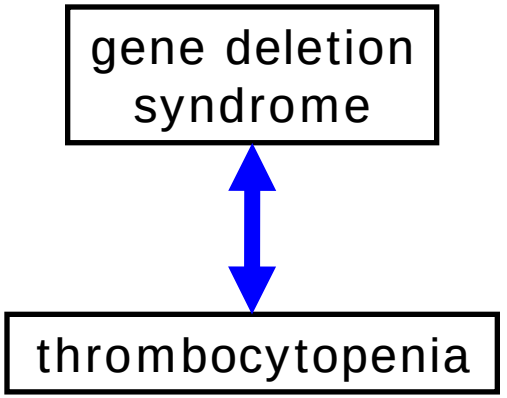


Example - Problem

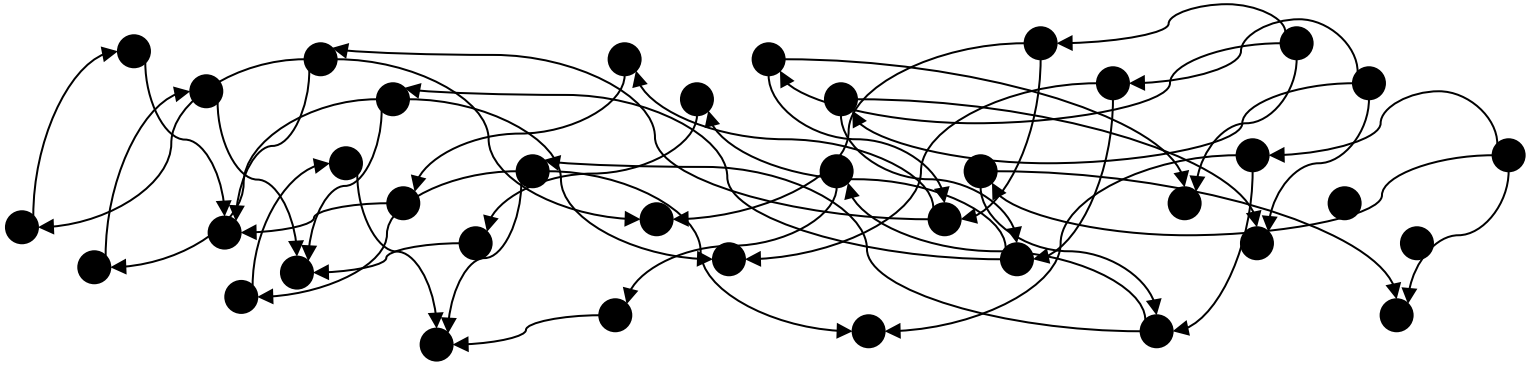
Phenotype description

A number sign (#) is used with this entry because it represents a contiguous gene deletion syndrome. See 274000 for another contiguous gene deletion syndrome, thrombocytopenia-absent radius (TAR) syndrome, that maps to a nonoverlapping region of chromosome 1q21.1.

Gene deletion syndromes also play an important role in various genetic diseases, including thrombocytopenia



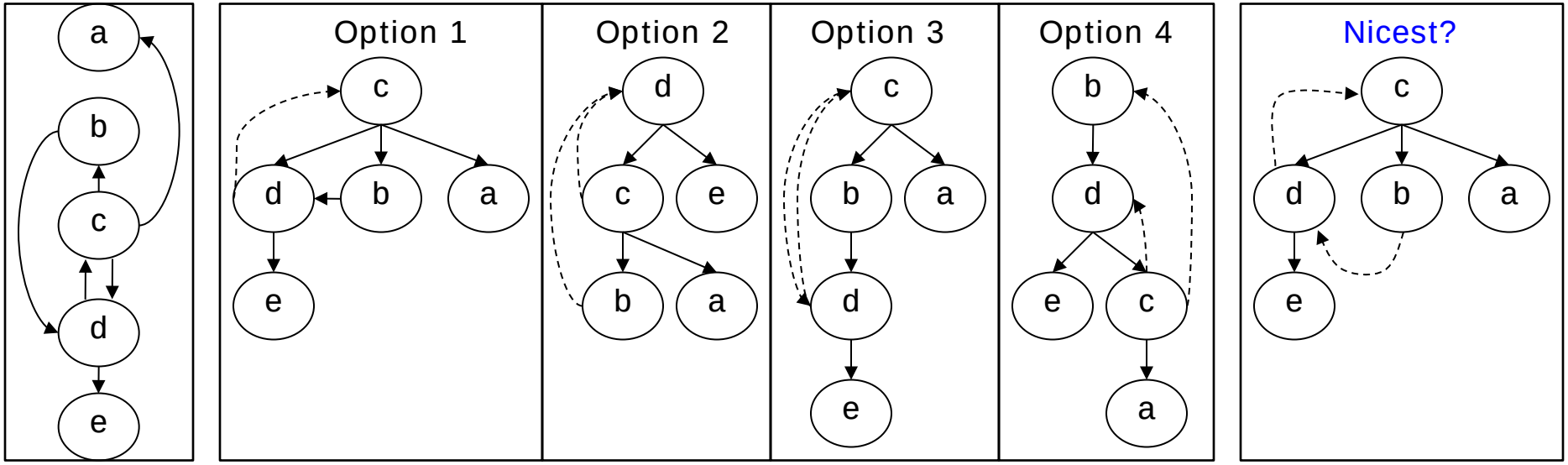
Application to 300K Texts and 12K Concepts



- No tree-like backbone structure
- Cyclic relationships: A ISA B ISA A
- Semantically suspicious, redundant, “not nice” parts
  - Parents that are brothers
  - Chains of single-child specializations
  - Parents with hundreds of children
  - ...
- Incomprehensible

4. Ontology Extraction Problem

- Given a directed, weighted Concept Graph G=(V,E)
  - Edge weights: strength of evidence
- Find a subgraph (Ontology Graph) G‘ that is
  - Consistent (= cycle-free)
  - Maximal confidence (= maximal total edge weight)
  - Nice (= adheres to some topological properties)



Evaluation Compared to MPO

- Mammalian Phenotype Ontology
  - 11700 concepts, 6828 relations, 172134 transitive relations
- Greedy Edge Inclusion (GEI)
  - 4,400 True Positives
  - Precision 0.45
- Hierarchical Greedy Expansion (HGE)
  - 1,200 True Positives
  - Precision 0.51
- Weighted Dominating Set Approach (wDSP)
  - 1,900 True Positives
  - Precision 0.54