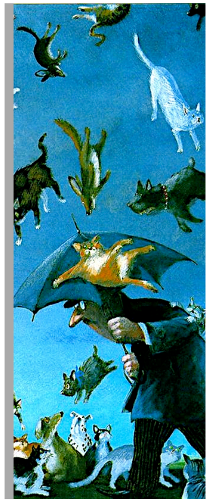
Statistical Association and Multiword Expressions

Stefan Evert
Institute of Cognitive Science
University of Osnabrück

stefan.evert@uos.de | purl.org/stefan.evert



RU Groningen | 10.05.2011

Collocations vs. multiword expressions

multiword expressions (phraseology) ⟷ **empirical collocations** (Firth / Sinclair)

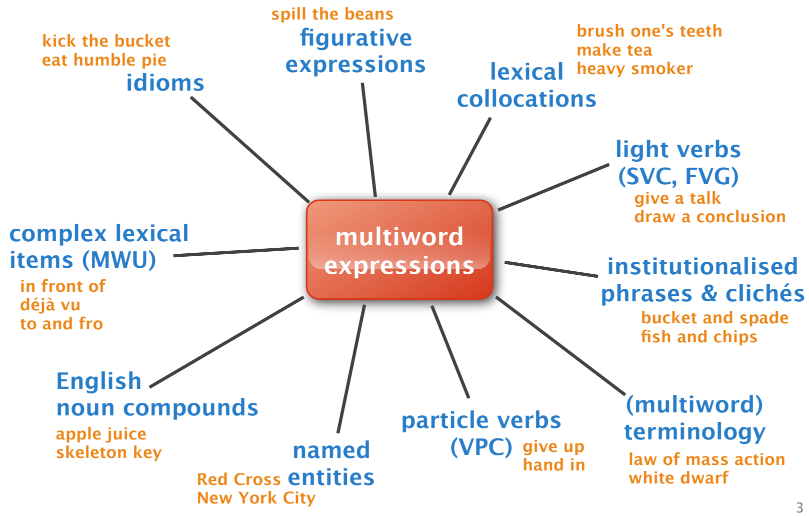
multiword expressions:

- based on native speaker judgments
- linguistic criteria/tests
- different subtypes
- still no precise def. & subtype classification
- "grey area" between idiomatic MWE and free combinations

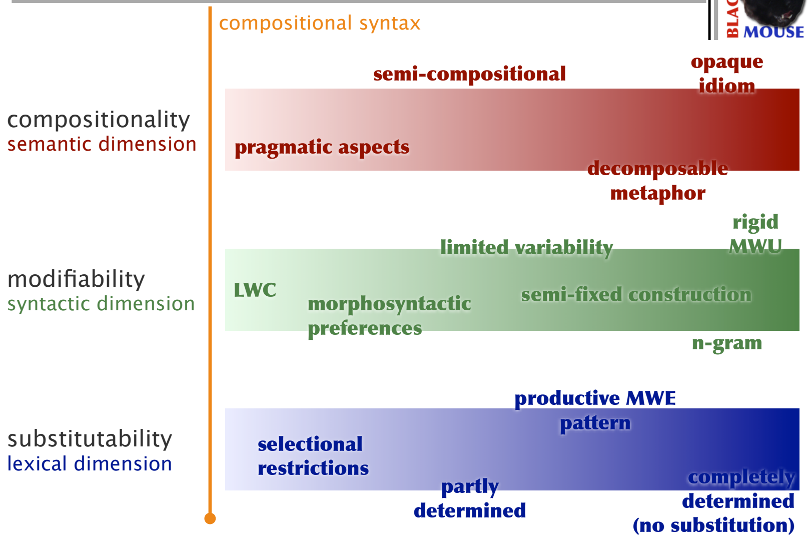
empirical collocations:

- quantitative empirical phenomenon
- based on corpus data
- linguistic status? words collocate for different "reasons"
- unclear: appropriate operationalisation ("collocation measure")

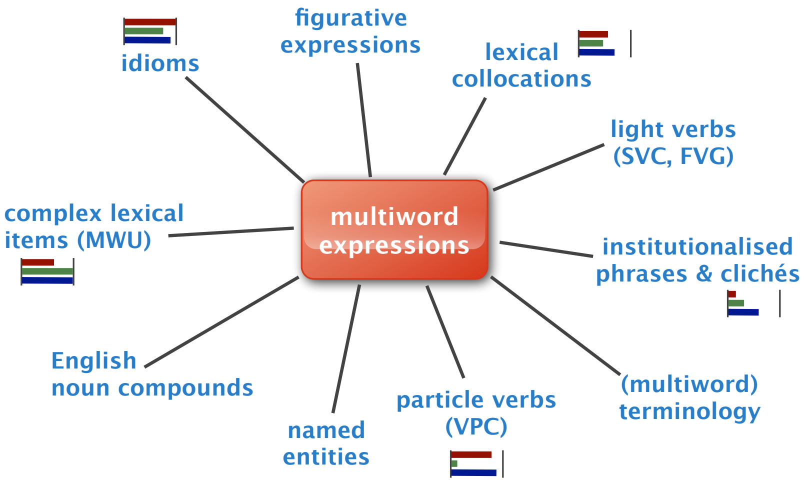
Types & examples of multiword expressions



Scales of MWE-ness



Types & examples of multiword expressions



Examples of collocations (BNC)

Produced with UCS toolkit | http://www.collocations.de/software.html

bucket

collocate	f	G^2
water	183	1064.079
spade	31	341.138
bucket	34	306.078
plastic	36	243.863
slop	15	213.303
mop	18	207.117
size	42	200.162
fill	38	195.749
record	42	174.693
throw	35	172.264
into	87	149.409
empty	18	148.807
with	191	147.546
ice	22	131.697
randomize	9	115.335
kick	19	113.238
of	488	81.765
single-record	5	81.107
large	36	80.852
shop	23	80.794
seat	20	78.645

rose

collocate	f	G^2
red	209	1113.501
shrub	68	572.793
hilaire	46	561.504
garden	133	548.936
cottage	75	418.288
bowl	66	389.237
petal	42	361.140
bush	65	324.711
net	63	321.964
white	104	319.160
pink	54	299.730
rose	66	285.331
mid-term	27	276.903
gun	60	271.541
axl	20	269.491
mary	64	265.474
wild	58	257.805
flower	63	251.148
per	74	231.465
miss	62	225.623
floyd	26	218.556

Examples of collocations (BNC)

Produced with UCS toolkit | http://www.collocations.de/software.html

bucket: nouns

collocate	f	G^2
water	183	1064.079
spade	31	341.138
bucket	34	306.078
plastic	36	243.863
slop	15	213.303
mop	18	207.117
size	42	200.162
record	42	174.693
ice	22	131.697
shop	23	80.794
seat	20	78.645
sand	13	68.814
brigade	10	67.080
shovel	7	64.335
coal	14	63.609
oats	7	62.659
rhino	7	60.813
champagne	10	59.556
density	10	59.132
algorithm	8	57.552
container	9	54.561

bucket: verbs

collocate	f	G^2
fill	38	195.749
throw	35	172.264
empty	18	148.807
randomize	9	115.335
kick	19	113.238
put	38	66.174
hold	31	62.765
tip	10	61.670
carry	25	59.554
fetch	9	52.665
chuck	7	50.638
store	10	48.327
pour	10	47.206
weep	7	43.396
douse	4	37.842
used	13	31.791
pack	7	29.582
use	33	28.469
slop	3	27.238
drop	10	26.855
clean	7	26.830

bucket: adjectives

collocate	f	G^2
single-record	5	81.107
large	36	80.852
cold	17	63.644
galvanized	4	51.373
full	22	49.746
steaming	4	32.883
leaky	3	29.520
empty	8	28.670
bottomless	3	28.397
galvanised	3	27.186
soggy	3	25.022
iced	3	24.535
small	20	24.033
clean	7	23.416
bowed	2	20.506
omnipresent	2	19.811
anglo-saxon	3	18.219
wooden	5	17.251
ice-cold	2	17.211
soapy	2	16.005
ten	10	15.864

Word sketch

http://beta.sketchengine.co.uk/

bucket British National Corpus freq = 1357

object_of	371	2.7
weep	7	7.61
empty	10	7.49
chuck	4	6.86
kick	14	6.6
fill	30	5.98
fetch	4	5.65
tip	3	5.18
pour	4	4.56
throw	11	4.33
drop	6	3.83

and/or	236	1.3
spade	28	10.06
mop	13	9.43
shovel	7	8.51
sponge	5	7.34
bin	4	6.17
bucket	4	5.76
container	5	5.73
cloth	5	5.4
brush	4	5.33
bowl	6	5.2

unary rels		
Sforto	5	5.3

particle	19	8.0
in	3	0.95
out	6	0.48

pp_of-p	248	3.5
whitewash	3	8.04
oats	4	7.55
water	127	6.36
champagne	3	5.19
sand	6	5.17
paint	4	4.89
coal	6	4.61
ice	3	4.11
blood	5	3.5
earth	3	3.2

pp_obj_in-p	102	3.0
store	8	5.58
drop	4	4.72
water	4	1.38

pp_with-p	27	2.3
champagne	3	5.31
capacity	4	3.39

adj_subject_of	24	1.5
full	13	3.74
large	4	1.78

pp_on-p	17	1.3
head	4	0.84

pp_obj_to-p	23	1.2
randomize	7	11.03

modifier	395	1.0
slop	11	9.61
galvanized	4	8.27
rhino	7	8.0
ten-record	3	7.95
full-track	3	7.94
leaky	3	7.7
bottomless	3	7.63
galvanised	3	7.5
plastic	29	7.32
mop	3	6.99

subject_of	57	0.8
stand	4	2.08
hold	10	2.07
contain	3	1.62

pp_obj_of-p	56	0.8
bottom	3	3.62
couple	4	2.71
use	3	0.76
number	4	0.36

pp_in-p	25	0.7
hand	5	0.87

modifies	158	0.5
algorithm	7	7.16
brigade	10	6.94
size	33	5.5
seat	20	5.18
shop	22	4.83
load	4	4.33
collection	10	4.08
hat	3	3.71
capacity	4	3.38
work	4	0.09

What are collocations?

Multiword Expressions (MWE)

idiom | compound | technical | lexical collocation | semantic relation | facts of life

bucket: nouns

collocate	f	G^2
water	183	1064.079
spade	31	341.138
bucket	34	306.078
plastic	36	243.863
slop	15	213.303
mop	18	207.117
size	42	200.162
record	42	174.693
ice	22	131.697
shop	23	80.794
seat	20	78.645
sand	13	68.814
brigade	10	67.080
shovel	7	64.335
coal	14	63.609
oats	7	62.659
rhino	7	60.813
champagne	10	59.556
density	10	59.132
algorithm	8	57.552
container	9	54.561

bucket: verbs

collocate	f	G^2
fill	38	195.749
throw	35	172.264
empty	18	148.807
randomize	9	115.335
kick	19	113.238
put	38	66.174
hold	31	62.765
tip	10	61.670
carry	25	59.554
fetch	9	52.665
chuck	7	50.638
store	10	48.327
pour	10	47.206
weep	7	43.396
douse	4	37.842
used	13	31.791
pack	7	29.582
use	33	28.469
slop	3	27.238
drop	10	26.855
clean	7	26.830

bucket: adjectives

collocate	f	G^2
single-record	5	81.107
large	36	80.852
cold	17	63.644
galvanized	4	51.373
full	22	49.746
steaming	4	32.883
leaky	3	29.520
empty	8	28.670
bottomless	3	28.397
galvanised	3	27.186
soggy	3	25.022
iced	3	24.535
small	20	24.033
clean	7	23.416
bowed	2	20.506
omnipresent	2	19.811
anglo-saxon	3	18.219
wooden	5	17.251
ice-cold	2	17.211
soapy	2	16.005
ten	10	15.864

Why collocations are important

- Primary tool for MWE identification
  - e.g. Evert/Krenn (2001, 2005) | MWE Workshops & Shared Task
- Language description: approximation of word meaning
  - Firth (1957) | Sinclair (1991) | computational lexicography
- Psycholinguistic relevance: priming & syntactic associates
  - priming effects | lexical priming (Hoey 2005) | link grammar etc.
- Collostructions, subcategorisation & selectional preferences
  - "collocations" between words & syntactic patterns
- Applications in NLP, e.g. long-distance adaptors for LM
- Basis of distributional semantic models (term-term matrix)

Key questions for MWE and collocations

- Linguistic definition of MWE and their subtypes
- Relation between (different subtypes of) MWE and (different quantitative notions of) empirical collocations
- Operationalisation of empirical collocations and appropriate quantitative measures

Co-occurrence and statistical association

Operationalising collocations

- Early "definitions"
  - recurrent, habitual word combinations (Firth 1957)
  - greater than chance co-occurrence (Sinclair 1966, 1970)
  - significant collocations (Kilgarriff & Tugwell 2002)
- Ingredient 1: co-occurrence
  - surface vs. textual vs. syntactic (Evert 2004, 2008)
  - contingency tables of joint & marginal frequencies
- Ingredient 2: statistical association
  - quantitative measure for tendency of events to co-occur
  - operationalises intuition of recurrent, “salient” combinations

Surface co-occurrence

Collocational span of 4 words (L4, R4), limited by sentence boundaries

A vast deal of coolness and a peculiar degree of judgement, are requisite in catching a hat. A man must not be precipitate, or he runs over it ; he must not rush into the opposite extreme, or he loses it altogether. [. . .] There was a fine gentle wind, and Mr. Pickwick’s hat *rolled* sportively before it. The wind puffed, and Mr. Pickwick puffed, and the hat *rolled* over and over, as merrily as a lively porpoise in a strong tide ; and on it might have *rolled*, far beyond Mr. Pickwick’s reach, had not its course been providentially stopped, just as that gentleman was on the point of resigning it to its fate.

f(hat, roll) = 2

Textual co-occurrence

Co-occurrence within sentences

A vast deal of coolness and a peculiar degree of judgement, are requisite in catching a hat.	hat	—
A man must not be precipitate, or he runs *over* it ;	—	over
he must not rush into the opposite extreme, or he loses it altogether.	—	—
There was a fine gentle wind, and Mr. Pickwick’s hat rolled sportively before it.	hat	—
The wind puffed, and Mr. Pickwick puffed, and the hat rolled *over* and *over* as merrily as a lively porpoise in a strong tide ;	hat	over

f(hat, over) = 1

Syntactic co-occurrence

Adjectival noun modification (prenominal adjectives)

In an *open barouche* [. . .] stood a *stout old gentleman*, in a *blue coat* and *bright buttons*, corduroy breeches and top-boots ; two *young ladies* in scarfs and feathers ; a *young gentleman* apparently enamoured of one of the *young ladies* in scarfs and feathers ; a lady of *doubtful age*, probably the aunt of the aforesaid ; and [. . .]

➜

open	barouche
stout	gentleman
old	gentleman
blue	coat
bright	button
young	lady
young	gentleman
young	lady
doubtful	age

f(young, gentleman) = 1

Observed frequency

- Collocations: "recurrent" combinations → simply use co-occurrence frequency as measure of salience?
- Example: most frequent adjacent bigrams from Brown corpus
- Frequent combinations don't seem to be very interesting collocations
- Mathematical reason:
  - f(is to) = 260
  - f(is) ≈ 10,000, f(to) ≈ 26,000
  - one would expect 260 co-occurrences if words were ordered randomly!

adjacent bigrams (Brown)

bigram	f	rank
of the	9702	1
in the	6018	2
to the	3478	3
on the	2459	4
and the	2242	5
for the	1845	6
to be	1715	7
at the	1654	8
with the	1530	9
it is	1482	10
of a	1469	11
in a	1413	12
from the	1410	13
that the	1378	14
by the	1347	15
it was	1338	16
he was	1110	17
as a	980	18
he had	933	19
...	...	...
is to	260	133

Observed & expected frequency

- Collocations: "recurrent" combinations → use co-occurrence frequency as measure of salience
- Example: most frequent adjacent bigrams from Brown corpus
- Frequent combinations don't seem to be very interesting collocations
- Mathematical reason:
  - f(is to) = 260
  - f(is) ≈ 10,000, f(to) ≈ 26,000
  - one would expect 260 co-occurrences if words were ordered randomly!

adjacent bigrams (Brown)

bigram	f	expected	rank
of the	9702	2186.75	1
in the	6018	1260.22	2
to the	3478	1613.01	3
on the	2459	384.84	4
and the	2242	1768.75	5
for the	1845	571.23	6
to be	1715	173.16	7
at the	1654	323.29	8
with the	1530	427.89	9
it is	1482	87.02	10
of a	1469	759.86	11
in a	1413	437.91	12
from the	1410	258.53	13
that the	1378	650.83	14
by the	1347	322.97	15
it was	1338	86.32	16
he was	1110	99.98	17
as a	980	155.42	18
he had	933	53.02	19
...	...	...	...
is to	260	266.61	133

Observed & expected contingency tables

	w_2	$\neg w_2$	
w_1	O_{11}	O_{12}	$= R_1$
$\neg w_1$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

observed

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

expected

- Contingency table = cross-classification of "items"
  - mathematical basis for concept of statistical association
- Statistics tells us how to calculate expected cell counts

Textual co-occurrence

Item = sentence (or other text segment)

A vast deal of coolness and a peculiar degree of judgement, are requisite in catching a hat.	hat	—
A man must not be precipitate, or he runs *over* it ;	—	over
he must not rush into the opposite extreme, or he loses it altogether.	—	—
There was a fine gentle wind, and Mr. Pickwick's hat rolled sportively before it.	hat	—
The wind puffed, and Mr. Pickwick puffed, and the hat rolled *over* and *over* as merrily as a lively porpoise in a strong tide ;	hat	over

f(hat, over) = 1
sample size N = 5

	over	¬over	
hat	O_{11}	O_{12}	R_1
¬hat	O_{21}	O_{22}	R_2
	C_1	C_2	N

Textual co-occurrence

Item = sentence (or other text segment)

A vast deal of coolness and a peculiar degree of judgement, are requisite in catching a hat.	hat	—
A man must not be precipitate, or he runs over it;	—	over
he must not rush into the opposite extreme, or he loses it altogether.	—	—
There was a fine gentle wind, and Mr. Pickwick's hat rolled sportively before it.	hat	—
The wind puffed, and Mr. Pickwick puffed, and the hat rolled over and over as merrily as a lively porpoise in a strong tide;	hat	over

	over	¬over	
hat	1	2	3
¬hat	1	1	2
	2	3	5

f(hat, over) = 1
sample size N = 5

Syntactic co-occurrence

Item = instance of adjective-noun modification

In an *open barouche* [...] stood a *stout old gentleman*, in a *blue coat* and *bright buttons*, corduroy breeches and top-boots; two *young ladies* in scarfs and feathers; a *young gentleman* apparently enamoured of one of the *young ladies* in scarfs and feathers; a lady of *doubtful age*, probably the aunt of the aforesaid; and [...]

➜

open	barouche
stout	gentleman
old	gentleman
blue	coat
bright	button
young	lady
young	gentleman
young	lady
doubtful	age

	•|gent.	•|¬gent	
young|•	1	2	3
¬young|•	2	4	6
	3	6	9

f(young, gentleman) = 1
sample size N = 9

Surface co-occurrence

Item = token

A vast deal of coolness and a peculiar degree of judgement, are requisite in catching a hat. A man must not be precipitate, or he runs over it; he must not rush into the opposite extreme, or he loses it altogether. [...] There was a fine gentle wind, and Mr. Pickwick's hat *rolled* sportively before it. The wind puffed, and Mr. Pickwick puffed, and the hat *rolled* over and over, as merrily as a lively porpoise in a strong tide; and on it might have *rolled*, far beyond Mr. Pickwick's reach, had not its course been providentially stopped, just as that gentleman was on the point of resigning it to its fate.

	roll	¬roll	
NEAR(hat)	2	18	20
¬NEAR(hat)	1	87	88
	3	105	108

f(hat, roll) = 2
sample size N = 108

Comparison

Data from BNC | RASP parser: http://www.informatics.susx.ac.uk/research/nlp/rasp/

"sell" – V+Obj		"sell" – (L0, R2)		"sell" – (L5, R5)		"sell" – sentence	
collocate	G^2	collocate	G^2	collocate	G^2	collocate	G^2
goods	1391.9	share	1312.8	goods	2177.7	price	3496.8
share	1231.6	goods	1309.6	product	1752.2	company	2661.1
product	946.0	product	1051.1	share	1749.0	market	2600.8
house	665.4	house	707.3	copy	1341.3	share	2593.6
property	602.1	ticket	611.0	shop	1232.0	goods	2435.3
land	478.0	property	460.9	ticket	1146.4	product	2363.9
ticket	453.3	land	388.4	property	903.0	shop	1922.8
asset	413.7	copy	383.5	company	869.1	sale	1460.9
copy	399.4	car	353.6	price	774.8	copy	1418.9
car	306.8	auction	276.3	house	728.4	dealer	1381.7
business	246.8	soul	236.8	dealer	632.8	property	1348.2
stock	224.1	liquor	223.7	car	595.5	business	1347.8
stake	205.5	asset	182.1	land	589.9	sales	1171.4
home	174.3	produce	166.9	asset	572.4	stock	1149.4
liquor	167.0	ware	156.5	market	524.0	ticket	1145.2
soul	166.0	bond	149.5	stock	501.7	profit	1109.3
bond	141.9	insurance	144.5	business	490.5	buyer	1076.7
produce	138.3	stake	131.2	auction	449.7	house	1048.2
company	110.8	stock	125.6	stake	362.0	auction	916.3
unit	105.5	advertising	112.6	liquor	335.4	owner	876.1
painting	105.2	cigarette	103.9	store	298.6	asset	873.6

Parsing accuracy

- How reliable is syntactic co-occurrence?
- Evert/Kermes (2003) evaluate adjective-noun identification
  - German prenominal adjectives
  - TIGER Treebank used as gold standard

candidates from	perfect tagging		TreeTagger tagging	
	precision	recall	precision	recall
adjacent pairs	98.47%	90.58%	94.81%	84.85%
window-based	97.14%	96.74%	93.85%	90.44%
YAC chunks	98.16%	97.94%	95.51%	91.67%

- Verb-object and verb-subject relations are much harder
  - Charniak-Johnson parser achieves 89.3% (direct object) and 96.5% (subject) on examples sentences from English Wiktionary
  - more difficult for languages with free word order (German)

Association measures (AM)

See Evert (2004, 2008) for details | http://www.collocations.de/

	w_2	$\neg w_2$	
w_1	O_{11}	O_{12}	$= R_1$
$\neg w_1$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

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

observed **expected**

Simple measures

Observed (O) vs. expected (E) co-occurrence frequency

$$\text{MI} = \log_2 \frac{O}{E} \qquad \text{MI}^k = \log_2 \frac{O^k}{E} \qquad \text{local-MI} = O \cdot \log_2 \frac{O}{E}$$

$$\text{z-score} = \frac{O - E}{\sqrt{E}} \qquad \text{t-score} = \frac{O - E}{\sqrt{O}} \qquad \text{simple-ll} = 2\left(O \cdot \log \frac{O}{E} - (O - E)\right)$$

Statistical measures

Comparison of full contingency tables (observed vs. expected)

$$\text{chi-squared} = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad \text{chi-squared}_{\text{corr}} = \frac{N\big(|O_{11}O_{22} - O_{12}O_{21}| - N/2\big)^2}{R_1 R_2 C_1 C_2}$$

$$\text{log-likelihood} = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}} \qquad \text{average-MI} = \sum_{ij} O_{ij} \cdot \log_2 \frac{O_{ij}}{E_{ij}}$$

$$\text{Dice} = \frac{2O_{11}}{R_1 + C_1} \qquad \text{odds-ratio} = \log \frac{\big(O_{11} + \frac{1}{2}\big)\big(O_{22} + \frac{1}{2}\big)}{\big(O_{12} + \frac{1}{2}\big)\big(O_{21} + \frac{1}{2}\big)}$$

$$\text{MI} = \log_2 \frac{O}{E} \qquad \text{MI}^k = \log_2 \frac{O^k}{E} \qquad \text{local-MI} = O \cdot \log_2 \frac{O}{E}$$

$$\text{z-score} = \frac{O - E}{\sqrt{E}} \qquad \text{t-score} = \frac{O - E}{\sqrt{O}} \qquad \text{simple-ll} = 2\left(O \cdot \log \frac{O}{E} - (O - E)\right)$$

Association measures (AM)

See Evert (2004, 2008) for details | http://www.collocations.de/

$$\mathrm{MI} = \log_2 \frac{O}{E} \qquad \mathrm{MI}^k = \log_2 \frac{O^k}{E} \qquad \text{local-MI} = O \cdot \log_2 \frac{O}{E}$$

recommended measures (Evert 2008)

$$\text{z-score} = \frac{O - E}{\sqrt{E}} \qquad \text{t-score} = \frac{O - E}{\sqrt{O}} \qquad \text{simple-ll} = 2\left(O \cdot \log \frac{O}{E} - (O - E)\right)$$

$$\text{chi-squared} = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad \text{chi-squared}_{\text{corr}} = \frac{N\left(|O_{11}O_{22} - O_{12}O_{21}| - N/2\right)^2}{R_1 R_2 C_1 C_2}$$

$$\text{log-likelihood} = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}} \qquad \text{average-MI} = \sum_{ij} O_{ij} \cdot \log_2 \frac{O_{ij}}{E_{ij}}$$

$$\text{Dice} = \frac{2O_{11}}{R_1 + C_1} \qquad \text{odds-ratio} = \log \frac{\left(O_{11} + \frac{1}{2}\right)\left(O_{22} + \frac{1}{2}\right)}{\left(O_{12} + \frac{1}{2}\right)\left(O_{21} + \frac{1}{2}\right)}$$

Comparison

Collocates of "bucket" in BNC (from Evert 2008)

collocate	f	f_2	simple-ll
water	184	37012	1083.18
a	590	2164246	449.30
spade	31	465	342.31
plastic	36	4375	247.65
size	42	14448	203.36
slop	17	166	202.30
mop	20	536	197.68
throw	38	11308	194.66
fill	37	10722	191.44
with	196	658584	171.78

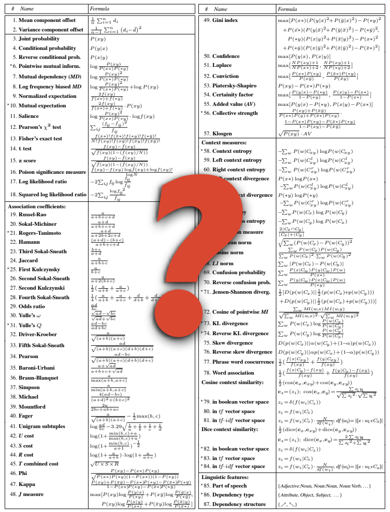
collocate	f	f_2	t-score
a	590	2164246	15.53
water	184	37012	13.30
and	479	2616723	10.14
with	196	658584	9.38
of	497	3040670	8.89
the	832	6041238	8.26
into	87	157565	7.67
size	42	14448	6.26
in	298	1937966	6.23
record	43	29404	6.12

collocate	f	f_2	MI
fourteen-record	4	4	13.31
ten-record	3	3	13.31
multi-record	2	2	13.31
two-record	2	2	13.31
a-row	1	1	13.31
anti-sweat	1	1	13.31
axe-blade	1	1	13.31
bastarding	1	1	13.31
dippermouth	1	1	13.31
Dok	1	1	13.31

collocate	$f \geq 5$	f_2	MI
single-record	5	8	12.63
randomize	10	57	10.80
slop	17	166	10.03
spade	31	465	9.41
mop	20	536	8.57
oats	7	286	7.96
shovel	8	358	7.83
rhino	7	326	7.77
synonym	7	363	7.62
bucket	18	1356	7.08

So many measures, so little time …

Pecina (2005) collects 57 association measures (and some other formulae)



Which measure?

How to choose an association measure

- Mathematical discussion
- Direct comparison
- Task-based evaluation
- Geometric interpretation
  - combine with insights from task-based evaluation

Significance of association

asymptotic hypothesis tests

$$\text{chi-squared} = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{N}{E_{22}} \cdot \frac{(O-E)^2}{E}$$

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

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

$$\text{Poisson} = \sum_{k=O}^{\infty} e^{-E} \frac{E^k}{k!}$$

exact hypothesis tests

simple hypothesis tests

$$\text{z-score} = \frac{O-E}{\sqrt{E}}$$

$$\text{simple-ll} = 2 \cdot \left(O \cdot \log \frac{O}{E} - (O - E) \right)$$

$$\text{t-score} = \frac{O-E}{\sqrt{O}}$$

$$\text{Poisson-likelihood} = e^{-E} \cdot \frac{(E)^O}{O!}$$

$$\text{Poisson-Stirling} = O \cdot \left(\log O - \log E - 1 \right)$$

likelihood measures

Degree of association / determination

$$\text{MI} = \log_2 \frac{O}{E}$$

$$\text{relative-risk} = \log \frac{O_{11} C_2}{O_{12} C_1}$$

$$\text{odds-ratio} = \log \frac{O_{11} O_{22}}{O_{12} O_{21}}$$

$$\text{gmean} = \frac{O_{11}}{\sqrt{R_1 C_1}} = \frac{O_{11}}{\sqrt{N E_{11}}}$$

measures of non-independence

$$p_F = \Pr(w_2 \mid w_1)$$
$$p_B = \Pr(w_1 \mid w_2)$$

$$\text{gmean} = \sqrt{p_F \cdot p_B} = \frac{O_{11}}{\sqrt{R_1 C_1}} = \frac{O}{\sqrt{NE}}$$

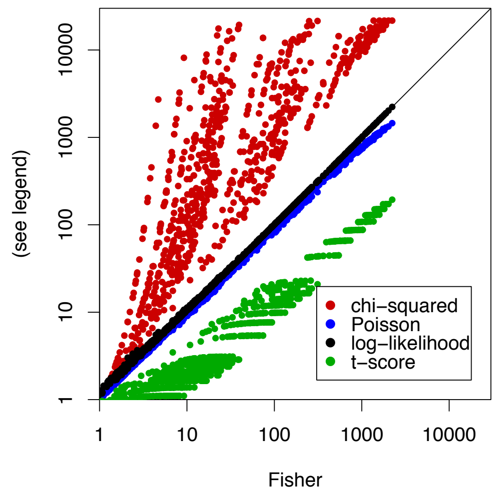
$$\text{Dice} = \left(\frac{1}{2p_F} + \frac{1}{2p_B} \right)^{-1} = \frac{2 O_{11}}{R_1 + C_1}$$

$$\text{MS} = \min\{p_F, p_B\} = \min\left\{ \frac{O_{11}}{R_1}, \frac{O_{11}}{C_1} \right\} = \frac{O_{11}}{\max\{R_1, C_1\}}$$

measures of (mutual) determination

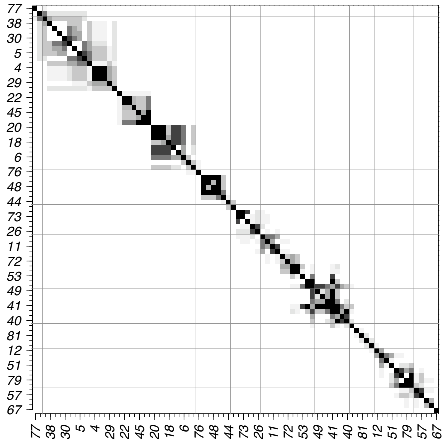
Direct comparison of association scores

Comparison of p-values on simulated data (see Evert 2004, 2008)



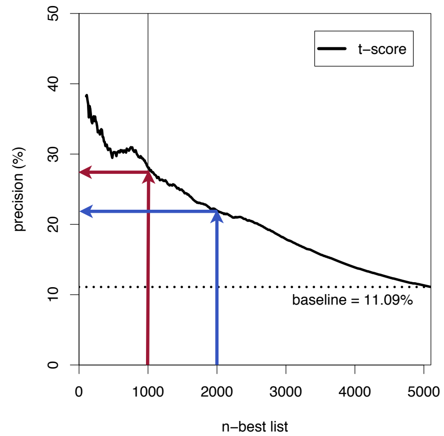
Direct comparison of AM scores

- ☆ Pecina & Schlesinger (2006) perform a systematic comparison
- ☆ Main result: several groups of highly correlated or even virtually identical AMs



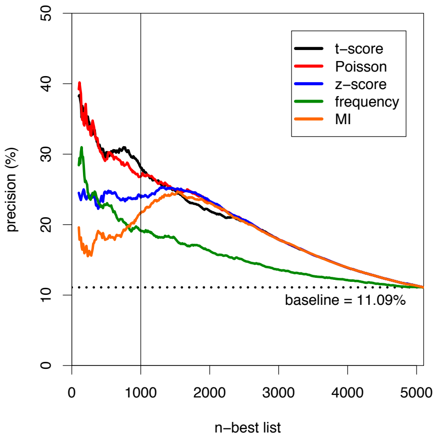
Empirical studies: MWE evaluation

- ☆ AM are used for ranking candidates in MWE extraction tasks
- ☆ Evaluation in terms of precision of n-best lists
- ☆ Gold standard
  - expert judgements of "usefulness" (for app.)
  - linguistically defined (subtypes of) MWE
  - always requires manual annotation of data!



Empirical studies: MWE evaluation

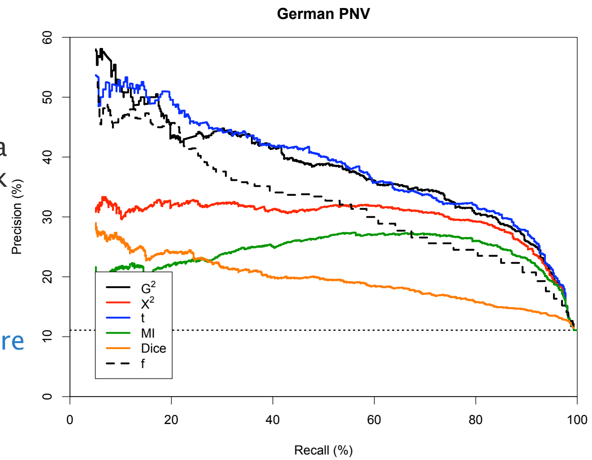
- ☆ German PP-verb pairs from FR corpus (f ≥ 30)
- ☆ MWE annotated by Brigitte Krenn (2000)
  - Funktionsvergefüge (FVG)
  - figurative expressions
- ☆ Data & guidelines: www.collocations.de



MWE 2008 Shared Task: DE-PNV

http://multiword.sf.net/mwe2008/

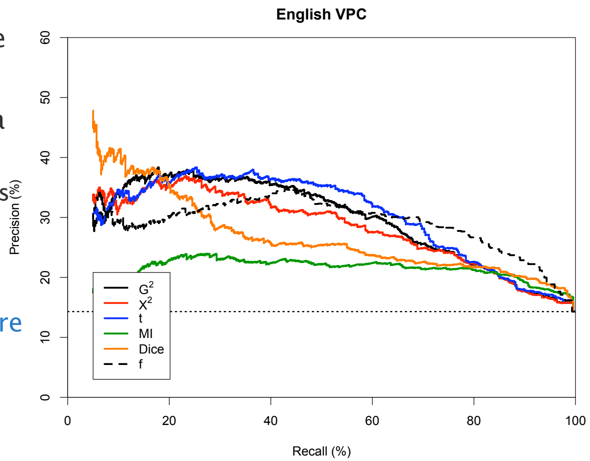
- ☆ Shared task on German V+PP
  - FVG
  - figurative
- ☆ Frequency data from FR, chunk parsed, f ≥ 30
- ☆ Baseline: 11.09%
- ☆ Best AM: t-score AP = 39.79%
- ☆ Frequency: AP = 33.88%



MWE 2008 Shared Task: EN-VPC

http://multiword.sf.net/mwe2008/

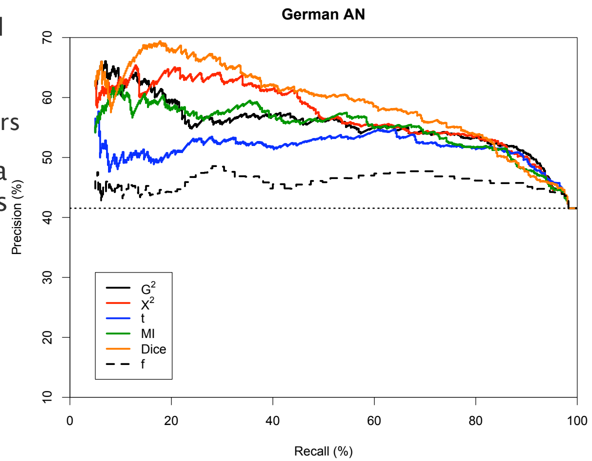
- Shared task on English particle verbs (VPC)
- Frequency data from full BNC
  - adjacent pairs
- Baseline: 14.29%
- Best AM: t-score AP = 29.94%
- Frequency: AP = 29.01%



MWE 2008 Shared Task: DE-AN

http://multiword.sf.net/mwe2008/

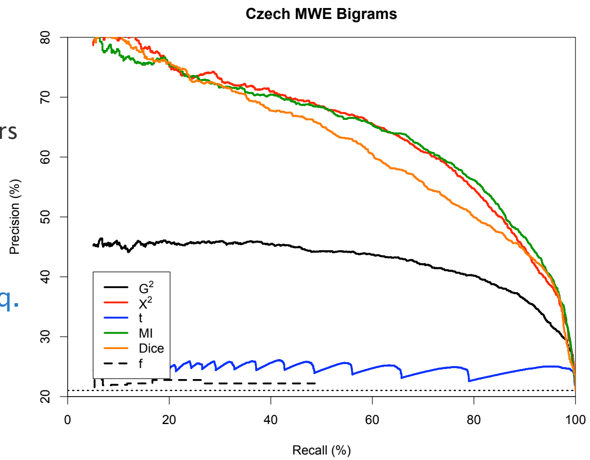
- Shared task on German Adj+N
  - evaluated by professional lexicographers
- Frequency data from FR corpus
- Baseline: 41.53%
- Best AM: Dice AP = 58.84%
- Frequency: AP = 46.90%



MWE 2008 Shared Task: CZ-MWE

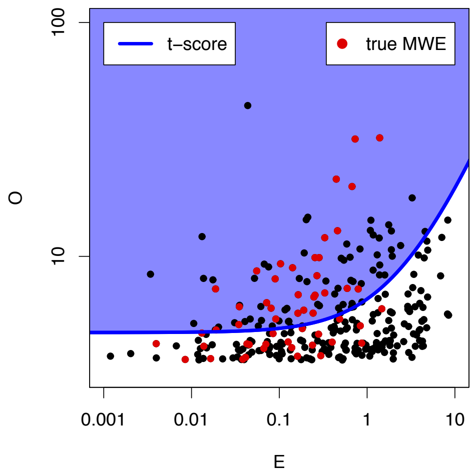
http://multiword.sf.net/mwe2008/

- Shared task on Czech MWE
  - evaluated by lexicographers
  - three judges
- Baseline: 21.03%
- Best AM: chi-sq. AP = 64.86%
- Frequency: AP = 21.70%



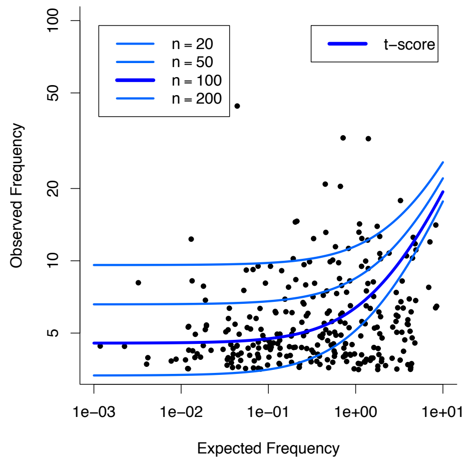
Geometric visualisation of AMs

See Evert (2004, 2008) for details

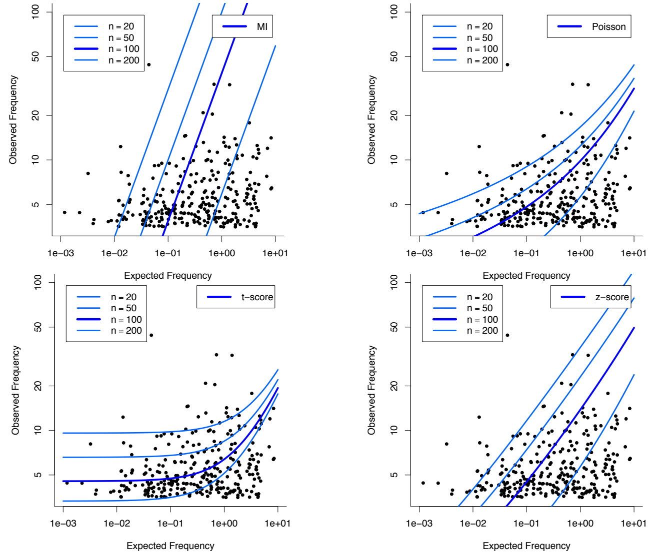


Geometric visualisation of AMs

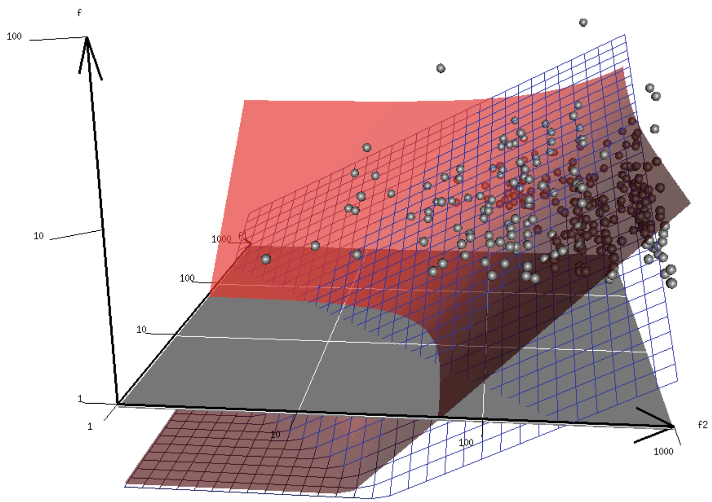
See Evert (2004, 2008) for details



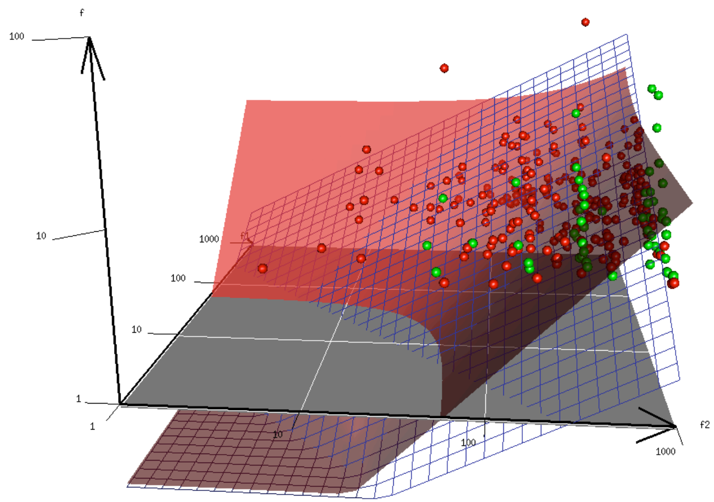
Geometric visualisation of AMs



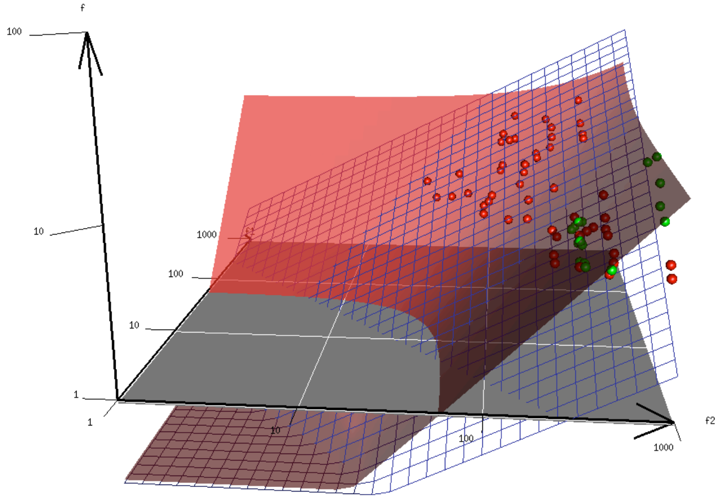
Geometric visualisation of AMs



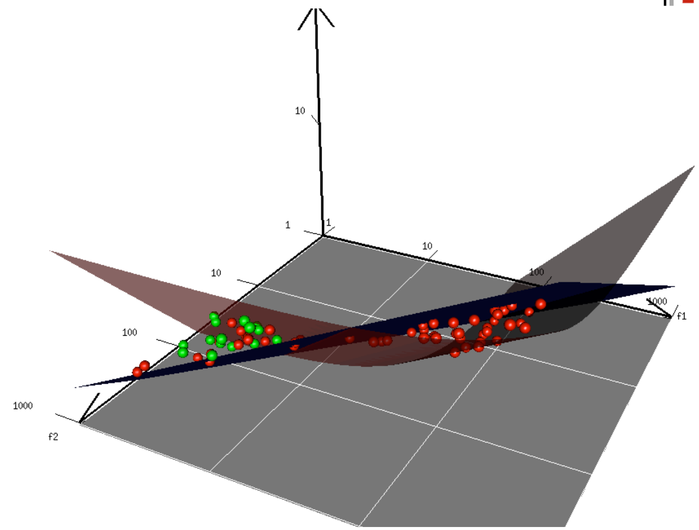
Evaluation & visualisation combined



Evaluation & visualisation combined



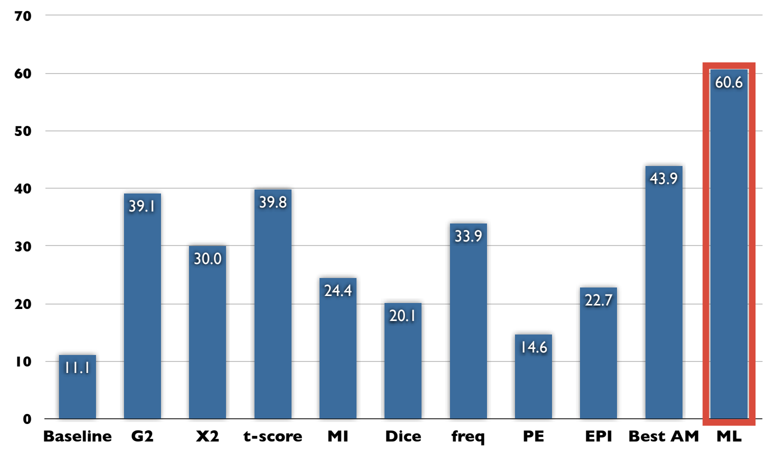
Evaluation & visualisation combined



Room for improvement

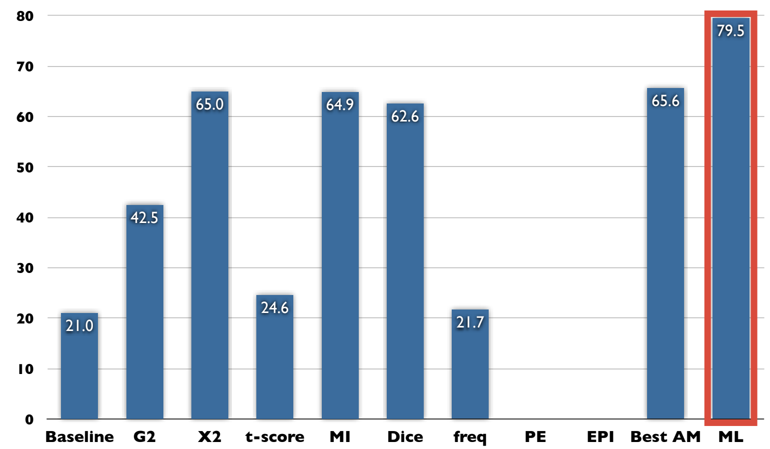
Machine learning (Pecina & Schlesinger 2006)

Results from MWE 2008 Shared Task: DE-PNV



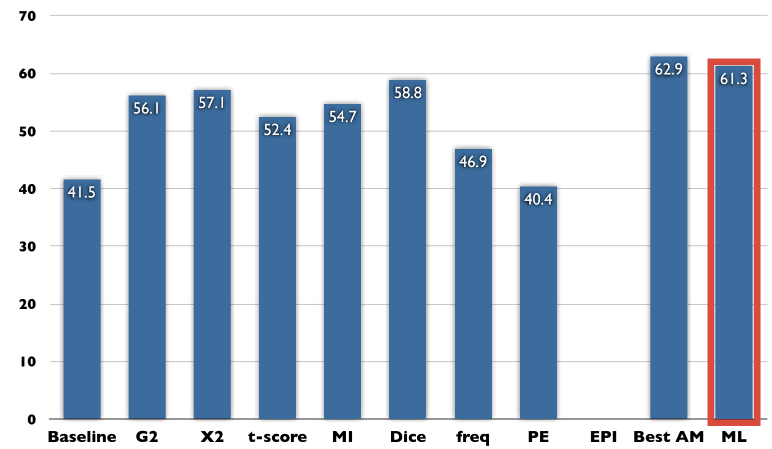
Machine learning (Pecina & Schlesinger 2006)

Results from MWE 2008 Shared Task: CZ-MWE



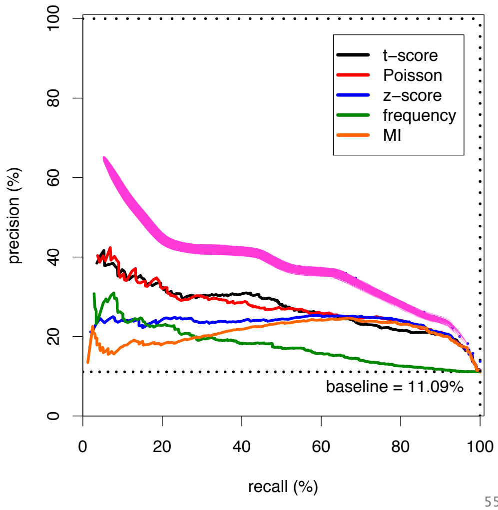
Machine learning (Pecina & Schlesinger 2006)

Results from MWE 2008 Shared Task: DE-AN

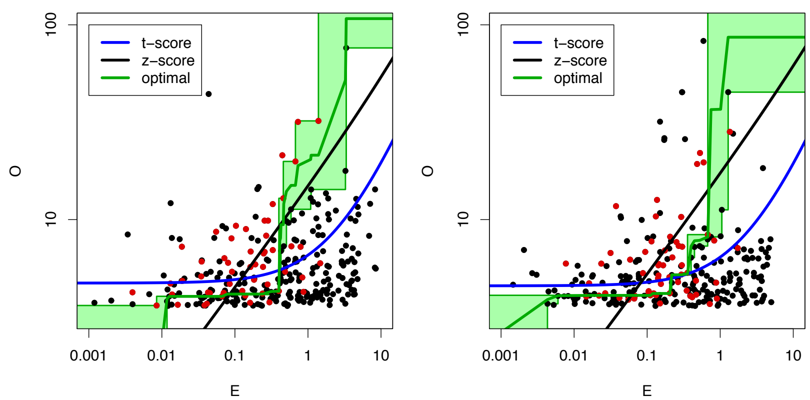


Upper limits: overtraining

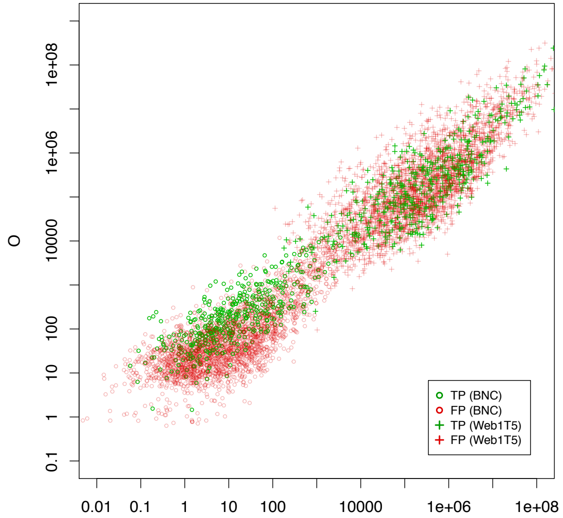
- ☆ What is the highest precision that a "sensible" AM can achieve in principle?
- ☆ Like a highly over-trained machine learning approach
- ☆ Restriction needed: simple AM



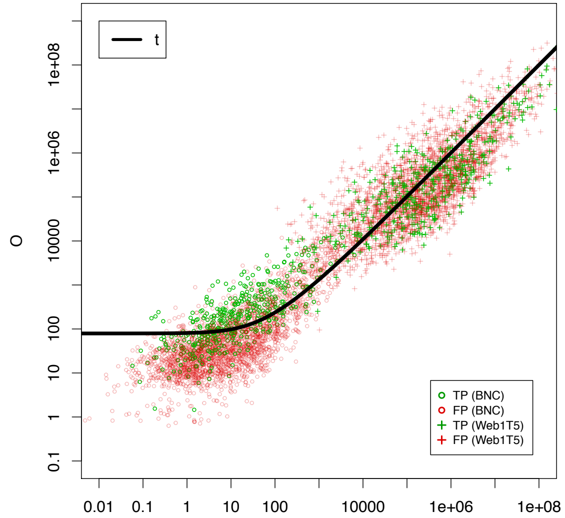
Upper limits: optimal simple AM



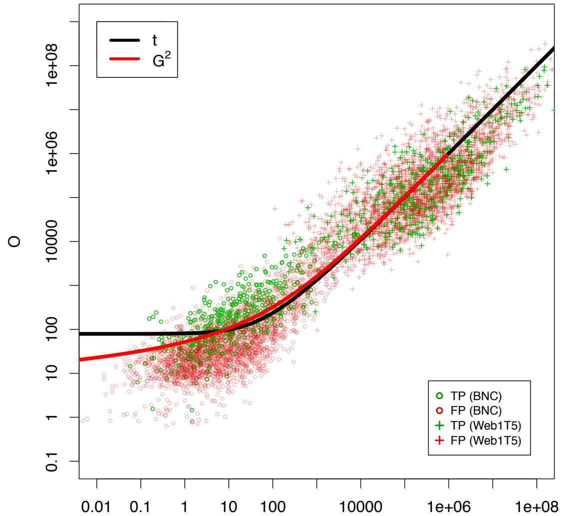
Do AMs scale up to the Web?



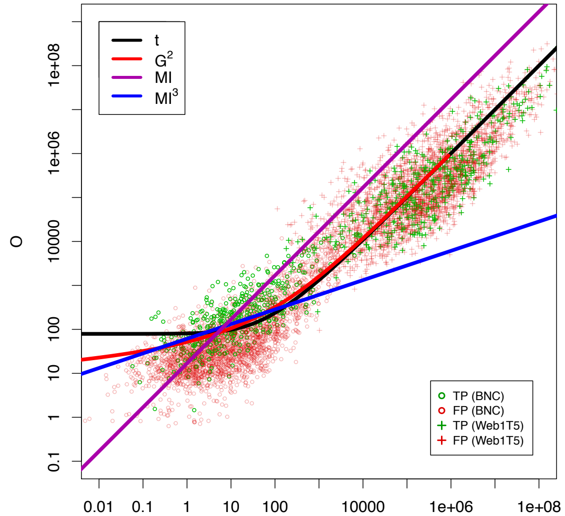
Do AMs scale up to the Web?



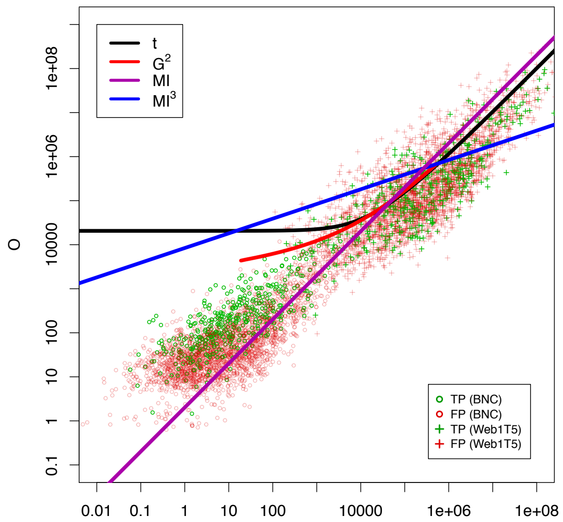
Do AMs scale up to the Web?



Do AMs scale up to the Web?



Do AMs scale up to the Web?



What else?

Some current research topics (my agenda)

- Optimised AMs for specific types of tasks and data sets
  - e.g. for identification of SVC vs. idioms
  - for very small or very large corpora, skewed frequency dist's
- Extension to combinations of three or more words
  - particularly important for MWE, but also empirical collocations
  - basis for higher-order distributional semantics (tensors)
- Asymmetric association measures (Michelbacher et al. 2011)
  - e.g. wellington boot, bated breath, high fidelity
  - virtually all statistical AM are symmetric
- Collocational patterns: productivity of collocations
  - integration of collocations with distributional similarity

(A)symmetry of association

- Collocations are often **asymmetric** (Kjellmer 1991)
  - e.g. wellington boot, bated breath, high fidelity
  - bated breath is “right-predictive”, high fidelity is “left-predictive”
  - effect may in part be due to frequency of collocates
- Well-known fact, but little research in linguistics & NLP
  - MWE and semantic relations are inherently symmetric
  - most sensible measures of 1st– and 2nd–order statistical association are also symmetric
  - including all association measures mentioned in this talk

Asymmetric association measures

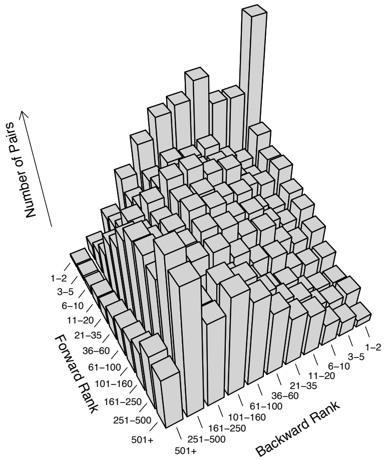
- Mathematically founded derivations lead to symmetric AM
  - how can asymmetry of association be accounted for?
- Michelbacher et al. (2007): forward vs. backward rank

“bated” (log-likelihood)		
collocate	**score**	**rank**
breath	339.10	1
with	99.75	2
waited	75.02	3
waiting	50.91	4
and	0.88	5
,	0.00	6
.	−0.11	7
the	−0.91	8

“breath” (log-likelihood)		
collocate	**score**	**rank**
deep	6787.38	1
took	3207.68	2
her	2812.36	3
his	2100.52	4
…	…	…
shuddering	399.37	19
bated	376.96	20
draw	343.53	21

Asymmetric association measures

- Michelbacher et al. (2007): forward vs. backward rank
- Asymmetric AM (AAM): score = difference between forward & backward rank
- Various AAM can be defined (one for each symmetric AM)
- Plot shows distribution of forward and backward ranks
  - based on log-likelihood AM
  - for symmetric A, largest bars would be on the diagonal



AAM evaluation results (work in progress)

- Free associations are often asymmetric
- Michelbacher et al. (2007) evaluate AAM on USF free A norms
- Results are inconlusive
  - presumably because free association norms are mostly based on paradigmatic relations
  - 1st-order statistical A is syntagmatic (is it?)

USF free associations			
cue	**target**	**fwd A**	**bwd A**
boys	girls	0.500	0.503
bad	good	0.750	0.758
dinner	supper	0.535	0.545
trout	fish	0.913	0.036
saddle	horse	0.879	0.103
crib	baby	0.842	0.032
exhausted	tired	0.895	0.075
bank	money	0.799	0.019
bouquet	flowers	0.828	0.053

AAM evaluation results (work in progress)

- Evaluation on new data set of free syntagmatic A
  - similar to free A norms, but asks for syntagmatic combination
- Michelbacher et al. (2011)
  - fwd/bwd ranks for different AM
  - compared to syntagmatic A

f	b	(w_1, w_2)	$R_f^{\rightarrow}$	$R_f^{\leftarrow}$	$R_{G^2}^{\rightarrow}$	$R_{G^2}^{\leftarrow}$	$R_t^{\rightarrow}$	$R_t^{\leftarrow}$
group A: rank measures and direction scores conform								
0.5891	0.2545	Academy Award	1	9	1	2	1	7
0.3328	0.0010	ancestral home	1	25	1	13	1	19
0.5551	0.1609	cable television	2	7	1	4	2	5
0.0127	0.0087	cut glass	1	75	1	46	1	58
0.6760	0.0010	felled tree	1	54	1	33	1	45
0.0683	0.0021	hunched shoulders	1	16	1	7	1	14
0.0875	0.0010	old-fashioned way	1	98	1	60	1	62
0.1667	0.0063	rightful place	1	26	1	6	1	15
0.1500	0.0496	rope ladder	1	4	1	4	1	4
0.0241	0.0010	shrewd idea	3	109	6	49	3	68
0.1719	0.0010	thick-set man	1	519	1	169	1	318
0.0641	0.0068	well-worn path	1	71	1	34	1	58
0.0127	0.0125	*impending retirement	9	18	8	14	9	18
0.0606	0.0563	*speech recognition	1	2	1	1	1	2
0.0010	0.0099	annual rent	29	2	20	1	28	2
0.0266	0.8208	Christmas decorations	11	1	8	1	11	1
0.0010	0.0101	female preferences	63	34	92	44	60	34
0.0010	0.0650	hard frost	39	1	21	1	35	1
0.0010	0.0312	legal wrangling	151	1	58	1	110	1
0.0081	0.1325	smoked mackerel	5	1	3	1	5	1
0.0010	0.0031	southern bypass	21	1	15	1	20	1
0.0046	0.0426	welcome diversion	17	3	15	1	16	3
0.0032	0.0425	*bond issuance	10	1	7	1	10	1

AAM evaluation results (work in progress)

- Some results good
  - previous slide
- Other results are less encouraging
  - AAM are unclear or contradict syntagmatic A
- wishful thinking
  - fwd/bwd rank 1 for all AM
  - right-predictive in human data

group B: rank measures and direction scores do not conform								
0.0955	0.1160	healthy food	6	19	6	20	5	15
0.1562	0.1543	missile silos	16	1	8	1	16	1
0.0010	0.0063	seasoned campaigners	1	9	1	6	1	9
group C: rank measures ambivalent								
0.5411	0.4620	epileptic seizure	2	3	2	1	2	3
0.0761	0.0335	dedicated follower	7	3	2	3	4	3
0.4340	0.0237	laboratory experiments	2	1	1	1	2	1
0.0683	0.1836	South East	1	2	3	2	1	2
group D: high mutual predictiveness								
0.2962	0.1337	bloody hell	1	1	1	1	1	1
0.1275	0.2833	*special needs	1	1	1	1	1	1
0.6810	0.2793	toxic waste	1	1	1	1	1	1
0.2613	0.1583	unleaded petrol	1	1	1	1	1	1
0.9521	0.0068	wishful thinking	1	1	1	1	1	1

Thank you!

References

Aston, Guy and Burnard, Lou (1998). *The BNC Handbook*. Edinburgh University Press, Edinburgh. See also the BNC homepage at http://www.natcorp.ox.ac.uk/.

Burger, Harald; Buhofer, Annelies; Sialm, Ambros (1982). *Handbuch der Phraseologie*. De Gruyter, Berlin, New York.

Evert, Stefan (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Published in 2005, URN urn:nbn:de:bsz:93-opus-23714. Also see online resources at http://www.collocations.de/.

Evert, Stefan (2008). Corpora and collocations. In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, chapter 58. Mouton de Gruyter, Berlin.

Evert, Stefan and Kermes, Hannah (2003). Experiments on candidate data for collocation extraction. In *Companion Volume to the Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 83–86.

Evert, Stefan and Krenn, Brigitte (2001). Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195, Toulouse, France.

Evert, Stefan and Krenn, Brigitte (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language*, **19**(4), 450–466.

Evert, Stefan and Krenn, Brigitte (2005a). Exploratory collocation extraction. Presentation at the *Phraseology 2005 Conference*, Louvain-la-Neuve, Belgium.

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

Harris, Zellig (1954). Distributional structure. *Word*, **10**(23), 146–162.

Hoey, Michael (2005). *Lexical Priming: A new theory of words and language*. Routledge, London.

Hoffmann, Sebastian; Evert, Stefan; Smith, Nicholas; Lee, David; Berglund Prytz, Ylva (2008). *Corpus Linguistics with BNCweb – a Practical Guide, volume 6 of English Corpus Linguistics*. Peter Lang, Frankfurt am Main.

References

Kilgarriff, Adam and Tugwell, David (2002). Sketching words. In M.-H. Corréard (ed.), *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*, pages 125–137. EURALEX.

Kilgarriff, Adam; Rychly, Pavel; Smrz, Pavel; Tugwell, David (2004). The sketch engine. In *Proceedings of the 11th EURALEX International Congress*, Lorient, France.

Kjellmer, Göran (1991). A mint of phrases. In K. Aijmer and B. Altenberg (eds.), *English Corpus Linguistics*, pages 111–127. Longman, London.

Krenn, Brigitte (2000). *The Usual Suspects: Data-Oriented Models for the Identification and Representation of Lexical Collocations*, Dissertation. DFKI & Universität des Saarlandes, Saarbrücken, Germany.

Landauer, Thomas K. and Dumais, Susan T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, **104**(2), 211–240.

Lea, Diana (ed.) (2002). *Oxford Collocations Dictionary for students of English*. Oxford University Press, Oxford, New York.

Lin, Dekang (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics* (COLING-ACL 1998), pages 768–774, Montreal, Canada.

Michelbacher, Lukas; Evert, Stefan; Schütze, Hinrich (2007). Asymmetric association measures. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (RANLP 2007), Borovets, Bulgaria.

Michelbacher, Lukas; Evert, Stefan; Schütze, Hinrich (2011). Asymmetry in corpus-derived and human word associations. *Corpus Linguistics and Linguistic Theory*, **7**(2), 245–276.

Pecina, Pavel (2005). *An extensive empirical study of collocation extraction methods*. In Proceedings of the ACL Student Research Workshop, pages 13–18, Ann Arbor, MI.

Pecina, Pavel and Schlesinger, Pavel (2006). Combining association measures for collocation extraction. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (COLING/ACL 2006), Poster Sessions, pages 651–658, Sydney, Australia. ACL.

Rapp, Reinhard (2004). A freely available automatically generated thesaurus of related words. In *Proceedings of the 4th International Conference on Language Resources and Evaluation* (LREC 2004), pages 395–398.

References

Sahlgren, Magnus (2006). *The Word Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Department of Linguistics, Stockholm University.

Schütze, Hinrich (1998). Automatic word sense discrimination. *Computational Linguistics*, **24**(1), 97–123.

Sinclair, John McH. (1966). *Beginning the study of lexis*. In C. E. Bazell, J. C. Catford, M. A. K. Halliday, and R. H. Robins (eds.), In Memory of J. R. Firth, pages 410–430. Longmans, London.

Sinclair, John; Jones, Susan; Daley, Robert; Krishnamurthy, Ramesh (1970/2004). *English Collocation Studies: The OSTI Report*. Continuum Books, London and New York. Originally written in 1970 (unpublished).

Sinclair, John (1991). *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

Sinclair, John (ed.) (1995). *Collins COBUILD English Dictionary*. Harper Collins, London. New edition, completely revised.