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

- 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 (Firth / Sinclair)
- quantitative empirical phenomenon
- based on corpus data
- linguistic status?

Types & examples of multiword expressions

- idioms: kick the bucket, eat humble pie
- lexical collocations: brush one’s teeth, make tea
- light verbs: SVC, FVG, give a talk, draw a conclusion
- institutionalised phrases & clichés: bucket and spade, fish and chips
- (multiword) terminology: heavy smoker, law of mass action
- complex lexical items (MWU): déjà vu, to and fro
- English noun compounds: apple juice, skeleton key
- particle verbs (VPC): give up hand in
- named entities: Red Cross, New York City

Scales of MWE-ness

- compositional syntax
- compositionality
- semantic dimension
- pragmatic aspects
- decomposable metaphor
- modifiability
- syntactic dimension
- LWC
- limited variability
- rigid MWE
- substitutability
- lexical dimension
- selectional restrictions
- partly determined
- productive MWE
- completely determined
- (no substitution)
Types & examples of multiword expressions

- idioms
- figurative expressions
- lexical collocations
- light verbs (SVC, FVG)
- complex lexical items (MWU)
- English noun compounds
- named entities
- particle verbs (VPC)
- (multitword) terminology

Illustration:

- multiword expressions
  - figurative expressions
  - lexical collocations
  - light verbs (SVC, FVG)
  - complex lexical items (MWU)
  - English noun compounds
  - named entities
  - particle verbs (VPC)
  - (multitword) terminology

Examples of collocations (BNC)

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

<table>
<thead>
<tr>
<th>bucket: nouns</th>
<th>bucket: verbs</th>
<th>bucket: adjectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>water (183) 1064.079</td>
<td>spoke (31) 241.318</td>
<td>bucket (14) 106.078</td>
</tr>
<tr>
<td>plastic (36) 243.863</td>
<td>stop (15) 213.003</td>
<td>mop (18) 207.317</td>
</tr>
<tr>
<td>size (42) 200.162</td>
<td>fill (15) 197.749</td>
<td>record (42) 174.693</td>
</tr>
<tr>
<td>with (19) 147.546</td>
<td>ice (22) 131.697</td>
<td>throw (35) 172.264</td>
</tr>
<tr>
<td>snow (66) 189.237</td>
<td>push (65) 324.711</td>
<td>white (104) 318.160</td>
</tr>
<tr>
<td>red (209) 1153.501</td>
<td>black (46) 562.504</td>
<td>pink (54) 290.730</td>
</tr>
<tr>
<td>blue (75) 416.280</td>
<td>white (66) 285.311</td>
<td>green (27) 276.903</td>
</tr>
<tr>
<td>green (60) 271.541</td>
<td>blue (20) 268.493</td>
<td>yellow (64) 263.474</td>
</tr>
<tr>
<td>orange (63) 253.144</td>
<td>yellow (74) 253.465</td>
<td>brown (62) 225.623</td>
</tr>
<tr>
<td>red (26) 218.526</td>
<td>yellow (20) 217.526</td>
<td>brown (20) 217.526</td>
</tr>
</tbody>
</table>

Examples of collocations (BNC)

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

<table>
<thead>
<tr>
<th>bucket: nouns</th>
<th>bucket: verbs</th>
<th>bucket: adjectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>water (183) 1064.079</td>
<td>spoke (31) 241.318</td>
<td>bucket (14) 106.078</td>
</tr>
<tr>
<td>plastic (36) 243.863</td>
<td>stop (15) 213.003</td>
<td>mop (18) 207.317</td>
</tr>
<tr>
<td>size (42) 200.162</td>
<td>fill (15) 197.749</td>
<td>record (42) 174.693</td>
</tr>
<tr>
<td>with (19) 147.546</td>
<td>ice (22) 131.697</td>
<td>throw (35) 172.264</td>
</tr>
<tr>
<td>snow (66) 189.237</td>
<td>push (65) 324.711</td>
<td>white (104) 318.160</td>
</tr>
<tr>
<td>red (209) 1153.501</td>
<td>black (46) 562.504</td>
<td>pink (54) 290.730</td>
</tr>
<tr>
<td>blue (75) 416.280</td>
<td>white (66) 285.311</td>
<td>green (27) 276.903</td>
</tr>
<tr>
<td>green (60) 271.541</td>
<td>blue (20) 268.493</td>
<td>yellow (64) 263.474</td>
</tr>
<tr>
<td>orange (63) 253.144</td>
<td>yellow (74) 253.465</td>
<td>brown (62) 225.623</td>
</tr>
<tr>
<td>red (26) 218.526</td>
<td>yellow (20) 217.526</td>
<td>brown (20) 217.526</td>
</tr>
</tbody>
</table>

Word sketch

http://beta.sketchengine.co.uk/
### What are collocations?

<table>
<thead>
<tr>
<th>Bucket: Nouns</th>
<th>Bucket: Verbs</th>
<th>Bucket: Adjectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>water: 10.04</td>
<td>throw: 12.26</td>
<td>large: 8.08</td>
</tr>
<tr>
<td>spade: 14.33</td>
<td>empty: 14.88</td>
<td>cold: 6.84</td>
</tr>
<tr>
<td>bucket: 10.06</td>
<td>know: 13.13</td>
<td>full: 2.24</td>
</tr>
<tr>
<td>plate: 9.06</td>
<td>put: 16.87</td>
<td>stealing: 3.28</td>
</tr>
<tr>
<td>mop: 10.77</td>
<td>hold: 16.79</td>
<td>leaky: 2.93</td>
</tr>
<tr>
<td>record: 17.48</td>
<td>tip: 16.17</td>
<td>empty: 2.88</td>
</tr>
<tr>
<td>ice: 21.35</td>
<td>carry: 15.93</td>
<td>bottleless: 2.87</td>
</tr>
<tr>
<td>shop: 30.79</td>
<td>fetch: 15.84</td>
<td>flashy: 3.13</td>
</tr>
<tr>
<td>seat: 20.69</td>
<td>chuck: 15.68</td>
<td>empty: 2.87</td>
</tr>
<tr>
<td>sand: 66.81</td>
<td>store: 10.27</td>
<td>lead: 3.25</td>
</tr>
<tr>
<td>house: 64.35</td>
<td>pour: 10.27</td>
<td>small: 20.37</td>
</tr>
<tr>
<td>coal: 63.69</td>
<td>weep: 10.27</td>
<td>clean: 2.34</td>
</tr>
<tr>
<td>core: 62.69</td>
<td>used: 10.27</td>
<td>broad: 2.19</td>
</tr>
<tr>
<td>rhino: 10.81</td>
<td>pack: 10.27</td>
<td>present: 2.02</td>
</tr>
<tr>
<td>champagne: 50.95</td>
<td>use: 10.27</td>
<td>round: 3.00</td>
</tr>
<tr>
<td>density: 50.94</td>
<td>stop: 10.27</td>
<td>cold: 2.16</td>
</tr>
<tr>
<td>algorithm: 57.02</td>
<td>drop: 10.27</td>
<td>ice-cold: 2.16</td>
</tr>
<tr>
<td>container: 54.56</td>
<td>clean: 10.27</td>
<td>ten: 1.00</td>
</tr>
</tbody>
</table>

### 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**
Operationalising collocations

Early "definitions"
- recurrent, habitual word combinations (Firth 1957)
- greater than chance co-occurrence (Sinclair 1966, 1970)
- significant collocations (Kilgariff & 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

Textual co-occurrence

Co-occurrence within sentences

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.

\[ f(\text{hat, roll}) = 2 \]

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 [...] 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.

\[ f(\text{young, gentleman}) = 1 \]
**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:
- \( f(is\ to) = 260 \)
- \( f(is) \approx 10,000, f(to) \approx 26,000 \)
- one would expect 260 co-occurrences if words were ordered randomly!

**Observed & expected contingency tables**

<table>
<thead>
<tr>
<th>( w_1 )</th>
<th>( \neg w_1 )</th>
<th>( w_2 )</th>
<th>( \neg w_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( O_{11} )</td>
<td>( O_{12} )</td>
<td>( E_{11} = \frac{R_1 C_1}{N} )</td>
<td>( E_{12} = \frac{R_1 C_2}{N} )</td>
</tr>
<tr>
<td>( O_{21} )</td>
<td>( O_{22} )</td>
<td>( \neg w_1 )</td>
<td>( \neg w_2 )</td>
</tr>
</tbody>
</table>

\( w_1 \) = \( C_1 \) = \( C_2 \) = \( N \)

**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.

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 over and over as merrily as a lively porpoise in a strong tide;

The wind puffed, and Mr. Pickwick puffed, and the hat rolled over and over with the same velocity as a mouse.

\( f(hat, over) = 1 \)

Sample size \( N = 5 \)
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.
Parsing accuracy

How reliable is syntactic co-occurrence?


- German prenominal adjectives
- TIGER Treebank used as gold standard

<table>
<thead>
<tr>
<th>candidates from</th>
<th>perfect tagging</th>
<th>TreeTagger tagging</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>precision</td>
<td>recall</td>
</tr>
<tr>
<td>adjacent pairs</td>
<td>98.47%</td>
<td>90.58%</td>
</tr>
<tr>
<td>window-based</td>
<td>97.14%</td>
<td>96.74%</td>
</tr>
<tr>
<td>YAC chunks</td>
<td>98.16%</td>
<td>97.94%</td>
</tr>
</tbody>
</table>

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)

Simple measures

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

\[
\text{MI} = \log_2 \frac{O}{E} \quad \text{MI}^k = \log_2 \frac{O^k}{E} \quad \text{local-MI} = O \cdot \log_2 \frac{O}{E}
\]

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

Association measures (AM)


\[
\begin{align*}
R_1 &= O_{11} - O_{12} = C_1 = N - O_{21} - O_{22} = C_2 \quad \\
R_2 &= O_{21} - O_{22} = C_1 = N - O_{11} - O_{12} = C_2 \\
E_{11} &= \frac{R_1}{N} C_1 \quad E_{12} = \frac{R_1}{N} C_2 \\
E_{21} &= \frac{R_2}{N} C_1 \quad E_{22} = \frac{R_2}{N} C_2
\end{align*}
\]

Expected

\[
\begin{align*}
O_{ij} &= \frac{R_1}{N} C_1 \quad E_{ij} = \frac{R_1}{N} C_2 \\
w_i &= O_{1i} + O_{2i} = C_i \quad w_j = O_{i1} + O_{i2} = C_i
\end{align*}
\]

Simple measures

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

\[
\begin{align*}
\text{MI} = \log_2 \frac{O}{E} \quad \text{MI}^k = \log_2 \frac{O^k}{E} \quad \text{local-MI} = O \cdot \log_2 \frac{O}{E} \\
z\text{-score} = \frac{O - E}{\sqrt{E}} \quad t\text{-score} = \frac{O - E}{\sqrt{O}} \quad \text{simple-ll} = 2 \left( O \cdot \log \frac{O}{E} - (O - E) \right)
\end{align*}
\]

Statistical measures

Comparison of full contingency tables (observed vs. expected)

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

\[
\text{log-likelihood} = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}} \quad \text{average-MI} = \sum_{ij} O_{ij} \cdot \log_2 \frac{O_{ij}}{E_{ij}}
\]

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

\[
\begin{align*}
\text{MI}^2 &= \log_2 \frac{O}{E}^2 \quad \text{MI}^k &= \log_2 \frac{O^k}{E}^k \quad \text{local-MI} = O \cdot \log_2 \frac{O}{E} \\
z\text{-score} = \frac{O - E}{\sqrt{E}} \quad t\text{-score} = \frac{O - E}{\sqrt{O}} \quad \text{simple-ll} = 2 \left( O \cdot \log \frac{O}{E} - (O - E) \right)
\end{align*}
\]
Association measures (AM)


\[ MI = \log_2 \frac{O}{E} \]

\[ MI^k = \log_2 \frac{O^k}{E^k} \]

\[ z-score = \frac{O - E}{\sqrt{E}} \]
\[ t-score = \frac{O - E}{\sqrt{O^2}} \]

\[ chi-squared = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \]

\[ chi-squared_{corr} = \frac{N((O_{11}O_{22} - O_{12}O_{21})^2)}{R_1E_2C_1C_2} \]

\[ log-likelihood = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}} \]

\[ local-MI = O \cdot \log_2 \frac{O}{E} \]

\[ simple-II = 2 \left( O \cdot \log_2 \frac{O}{E} - (O - E) \right) \]

\[ MI = \log_2 \frac{O}{E} \]

\[ MI^k = \log_2 \frac{O^k}{E^k} \]

\[ z-score = \frac{O - E}{\sqrt{E}} \]
\[ t-score = \frac{O - E}{\sqrt{O^2}} \]

\[ chi-squared = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \]

\[ chi-squared_{corr} = \frac{N((O_{11}O_{22} - O_{12}O_{21})^2)}{R_1E_2C_1C_2} \]

\[ log-likelihood = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}} \]

\[ local-MI = O \cdot \log_2 \frac{O}{E} \]

\[ simple-II = 2 \left( O \cdot \log_2 \frac{O}{E} - (O - E) \right) \]

\[ Log-likelihood = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}} \]

\[ simple-II = 2 \left( O \cdot \log_2 \frac{O}{E} - (O - E) \right) \]

\[ odds-ratio = \log \left( \frac{O_{11} + \frac{1}{2}}{O_{12} + \frac{1}{2}} \right) \]

\[ Dice = \frac{2O_{11}}{R_1 + C_1} \]

recommended measures (Evert 2008)

So many measures, so little time …

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

Comparison

Collocates of “bucket” in BNC (from Evert 2008)

<table>
<thead>
<tr>
<th>Collocate</th>
<th>f</th>
<th>f2</th>
<th>simple-II</th>
</tr>
</thead>
<tbody>
<tr>
<td>water</td>
<td>184</td>
<td>590</td>
<td>1083.18</td>
</tr>
<tr>
<td>seed</td>
<td>216-42</td>
<td>449.30</td>
<td></td>
</tr>
<tr>
<td>plastic</td>
<td>31</td>
<td>465</td>
<td>342.31</td>
</tr>
<tr>
<td>size</td>
<td>36</td>
<td>4375</td>
<td>247.65</td>
</tr>
<tr>
<td>mop</td>
<td>17</td>
<td>166</td>
<td>202.30</td>
</tr>
<tr>
<td>throw</td>
<td>20</td>
<td>536</td>
<td>197.68</td>
</tr>
<tr>
<td>fill</td>
<td>37</td>
<td>1072</td>
<td>191.44</td>
</tr>
<tr>
<td>with</td>
<td>106</td>
<td>658584</td>
<td>171.78</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>collocate</th>
<th>f</th>
<th>f2</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>fourteen-record</td>
<td>4</td>
<td>4</td>
<td>13.31</td>
</tr>
<tr>
<td>ten-record</td>
<td>3</td>
<td>3</td>
<td>13.31</td>
</tr>
<tr>
<td>multi-record</td>
<td>2</td>
<td>2</td>
<td>13.31</td>
</tr>
<tr>
<td>two-record</td>
<td>2</td>
<td>2</td>
<td>13.31</td>
</tr>
<tr>
<td>a-row</td>
<td>1</td>
<td>1</td>
<td>13.31</td>
</tr>
<tr>
<td>anti-sweat</td>
<td>1</td>
<td>1</td>
<td>13.31</td>
</tr>
<tr>
<td>axe-blade</td>
<td>1</td>
<td>1</td>
<td>13.31</td>
</tr>
<tr>
<td>bastard</td>
<td>1</td>
<td>1</td>
<td>13.31</td>
</tr>
<tr>
<td>dippermouth</td>
<td>1</td>
<td>1</td>
<td>13.31</td>
</tr>
<tr>
<td>dink</td>
<td>1</td>
<td>1</td>
<td>13.31</td>
</tr>
</tbody>
</table>

Which measure?
How to choose an association measure

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

**Degree of association / determination**

- MI = \( \log_2 \frac{O}{E} \)
- relative-risk = \( \log \frac{O_{11}C_2}{O_{12}C_1} \)
- odds-ratio = \( \log \frac{O_{11}O_{22}}{O_{12}O_{21}} \)
- gmean = \( \frac{O_{11}}{\sqrt{R_1C_1}} = \frac{O_{11}}{\sqrt{NE_{11}}} \)

- \( p_F = \Pr(w_2 | w_1) \)
- \( p_B = \Pr(w_1 | w_2) \)
- gmean = \( \sqrt{p_F \cdot p_B} = \frac{O_{11}}{\sqrt{R_1C_1}} = \frac{O}{\sqrt{NE}} \)
- Dice = \( \left( \frac{1}{2p_F} + \frac{1}{2p_B} \right)^{-1} = \frac{2O_{11}}{R_1 + C_1} \)
- MS = \( \min \{ p_F, p_B \} = \min \left\{ \frac{O_{11}}{R_1}, \frac{O_{11}}{C_1} \right\} \)

measures of non-independence

measures of (mutual) determination

**Significance of association**

**asymptotic hypothesis tests**

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

**simple hypothesis tests**

- z-score = \( \frac{O - E}{\sqrt{E}} \)
- log-likelihood = \( 2 \sum_{i,j} O_{ij} \log \frac{O_{ij}}{E_{ij}} \)
- simple-ll = \( 2 \cdot (O \cdot \log \frac{O}{E} - (O - E)) \)
- t-score = \( \frac{O - E}{\sqrt{O}} \)

- Poisson-likelihood = \( e^{-E} \cdot \frac{(E)^O}{O!} \)
- Poisson-Stirling = \( O \cdot (\log O - \log E - 1) \)

**exact hypothesis tests**

**likelihood measures**

**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

MWE 2008 Shared Task: DE–PNV

- 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: EN–VPC

- 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

- 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

- 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

Evaluation & visualisation combined
Evaluation & visualisation combined

Room for improvement

Machine learning (Pecina & Schlesinger 2006)

Results from MWE 2008 Shared Task: DE-PNV

<table>
<thead>
<tr>
<th>Method</th>
<th>MI</th>
<th>Dice</th>
<th>t-score</th>
<th>freq</th>
<th>PE</th>
<th>EPI</th>
<th>Best AM</th>
<th>ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
</tr>
<tr>
<td>G2</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
</tr>
<tr>
<td>X2</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
</tr>
<tr>
<td>t-score</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
</tr>
<tr>
<td>MI</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
</tr>
<tr>
<td>Dice</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
</tr>
<tr>
<td>freq</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
</tr>
<tr>
<td>PE</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
</tr>
<tr>
<td>EPI</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
</tr>
<tr>
<td>Best AM</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
</tr>
<tr>
<td>ML</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
<td>11.1</td>
</tr>
</tbody>
</table>
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?

![Graph 1](image1)

![Graph 2](image2)

![Graph 3](image3)

![Graph 4](image4)
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

<table>
<thead>
<tr>
<th>“bated” (log-likelihood)</th>
<th>“breath” (log-likelihood)</th>
</tr>
</thead>
<tbody>
<tr>
<td>collocate</td>
<td>score</td>
</tr>
<tr>
<td>breath</td>
<td>339.10</td>
</tr>
<tr>
<td>with</td>
<td>99.75</td>
</tr>
<tr>
<td>waited</td>
<td>75.02</td>
</tr>
<tr>
<td>waiting</td>
<td>50.91</td>
</tr>
<tr>
<td>and</td>
<td>0.88</td>
</tr>
<tr>
<td>.</td>
<td>0.00</td>
</tr>
<tr>
<td>.</td>
<td>-0.11</td>
</tr>
<tr>
<td>the</td>
<td>-0.91</td>
</tr>
</tbody>
</table>

AAM evaluation results (work in progress)

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

<table>
<thead>
<tr>
<th>USF free associations</th>
</tr>
</thead>
<tbody>
<tr>
<td>cue</td>
</tr>
<tr>
<td>boys</td>
</tr>
<tr>
<td>bad</td>
</tr>
<tr>
<td>dinner</td>
</tr>
<tr>
<td>trout</td>
</tr>
<tr>
<td>saddle</td>
</tr>
<tr>
<td>crip</td>
</tr>
<tr>
<td>exhausted</td>
</tr>
<tr>
<td>bank</td>
</tr>
<tr>
<td>bouquet</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Evaluation on new data set of free syntagmatic A</th>
</tr>
</thead>
<tbody>
<tr>
<td>similar to free A norms, but asks for syntagmatic combination</td>
</tr>
<tr>
<td>Michelbacher et al. (2011)</td>
</tr>
</tbody>
</table>

- fwd/bwd ranks for different AM
- compared to syntagmatic A

<table>
<thead>
<tr>
<th>f</th>
<th>h</th>
<th>(x1,y1)</th>
<th>R ↑</th>
<th>R ↓</th>
<th>R G</th>
<th>R t</th>
<th>R χ</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3991</td>
<td>0.2545</td>
<td>0.9999</td>
<td>0.0010</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>0.3228</td>
<td>0.0101</td>
<td>1.0010</td>
<td>0.0000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>0.3551</td>
<td>0.1609</td>
<td>0.7500</td>
<td>0.0005</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>0.1272</td>
<td>0.0067</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>0.6750</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>0.0683</td>
<td>0.0021</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>0.0875</td>
<td>0.0010</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>62</td>
</tr>
<tr>
<td>0.1667</td>
<td>0.0063</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>0.1500</td>
<td>0.0496</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>0.0241</td>
<td>0.0010</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>68</td>
</tr>
<tr>
<td>0.1719</td>
<td>0.0010</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>318</td>
</tr>
<tr>
<td>0.0641</td>
<td>0.0068</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>58</td>
</tr>
<tr>
<td>0.0127</td>
<td>0.0125</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>0.0606</td>
<td>0.0563</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>0.0010</td>
<td>0.0099</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>0.0266</td>
<td>0.0208</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>0.0010</td>
<td>0.0101</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>0.0010</td>
<td>0.0650</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.0010</td>
<td>0.0012</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.0010</td>
<td>0.1325</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>0.0010</td>
<td>0.0131</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>0.0046</td>
<td>0.0426</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>0.0032</td>
<td>0.0425</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
</tr>
</tbody>
</table>
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

🌟 References


