E-VIEW-alation: A large-scale evaluation study of association measures for collocation identification

Stefan Evert¹, Peter Uhrig¹, Sabine Bartsch², Thomas Proisl¹

¹FAU Erlangen-Nürnberg   ²TU Darmstadt
TL; DR

- Comprehensive evaluation study
  - lexicographic gold standard (collocations dictionaries)
  - focus: corpus, co-occurrence context, freq. threshold

- Novel approach to sharing results
  - interactive Web-based viewer
  - gives access to complete evaluation results
Lexicalised MWE in lexicography

- idioms (*kick the bucket, play cat and mouse*)
- figurative expressions (*spill the beans*)
- multiword units (*in front of, déjà vu, to and fro*)
- particle verbs (*give up, hand in*)
- lexical collocations (*brush teeth, heavy smoker*)
- light verb constructions (*give talk, draw conclusion*)
- compounds (*skeleton key, penalty kick*)
- named entities (*New York City, Red Cross*)
- clichés (*thigh-high boots, bucket and spade*)
Lexicalised MWE in lexicography

- idioms (*kick the bucket, play cat and mouse*)
- figurative expressions (*spill the beans*)
- multiword units (*in front of, déjà vu, to and fro*)
- particle verbs (*give up, hand in*)
- lexical collocations (*brush teeth, heavy smoker*)
- light verb constructions (*give talk, draw conclusion*)
- compounds (*skeleton key, penalty kick*)
- named entities (*New York City, Red Cross*)
- clichés (*thigh-high boots, bucket and spade*)
MWE identification

- Main cue: co-occurrence frequency, quantified by statistical association measures (AM, Sinclair 1966)
- Other criteria derived from properties of MWE (Manning & Schütze 1999, 184)
  - non-compositionality (→ distributional semantics)
  - non-modifiability (→ syntactic flexibility, fixed ordering)
  - non-substitutability (→ substitution tests)
- Recent work focuses on feature combination by machine learning and on specific subtypes of MWE
  - AMs still play central role, esp. for collocation identification
Evaluation

- Evaluation studies usually test a single new algorithm of focus on a specific subtype of MWE
- Our goal is a broad-scale comparative evaluation:
  - Which AM correlates best with collocativity?
  - What is an appropriate co-occurrence context?
  - Which source corpora provide the best results?
  - Does size matter? Or representativeness?
  - Are there interactions between these parameters?
  - Are crawled Web corpora and n-gram databases a viable substitute for expensive reference corpora?
Gold standard

- **BBI:** The BBI Combinatory Dictionary of English (Benson, Benson & Ilson 1986)
  - based on lexicographic native-speaker intuitions
  - pre-corpus era → no bias towards specific method/corpus

- **OCD2:** Oxford Collocations Dictionary for students of English, 2nd ed. (McIntosh, Francis & Poole 2009)
  - corpus-based, much more comprehensive
  - clearer notion of collocation (≈ our subtypes of MWE)
The Bartsch224 gold standard

- Set of 203 node words selected by Sabine Bartsch
  - original set contained approx. 224 node words
  - some obscure nodes with few collocates omitted

- Manually extracted all lexical words (nouns, verbs, adjectives, adverbs) from corresponding BBI entries
  - set of 2,845 node-collocate pairs
  - lemmatized, reduced to two-word collocations

- Automatic extraction from XML version of OCD2
  - also from other entries (our node word listed as collocate)
  - set of 18,545 node-collocate pairs
Gold standard example: BBI

Node: measure (noun or verb)

- cubic, dry, liquid, metric, tape, certain, good, make, take,

**measure** I  

1. a cubic; dry; liquid; metric ~ 2. a tape ~ 3. in a certain ~ (in large ~) 4. (misc.) for good ~ ('as smt. extra'); made to ~ ('custom-made'); to take smb.'s ~ ('to evaluate smb.') (see also measures)

**measure** II  

1. (d; tr.) to ~ against (to ~ one's accomplishments against smb. else's) 2. (P; intr.) the room ~s twenty feet by ten
Gold standard example: BBI

Node: measure (noun or verb)

- cubic, dry, liquid, metric, tape, certain, good, make, take, carry, coercive, compulsory, draconian, drastic, harsh, stern, stringent, emergency, extreme, radical, preventive, prophylactic, safety, security, stopgap, temporary

**measures n.** 1. to carry out, take ~ 2. coercive; compulsory; draconian; drastic, harsh, stern, stringent; emergency; extreme, radical; preventive, prophylactic; safety, security; stopgap, temporary ~ 3. ~ to + inf. (we took ~ to insure their safety) 4. ~ against (to take ~ against smuggling)
Gold standard example: OCD

Node: measure (noun or verb)

- cubic, dry, liquid, metric, tape, certain, good, make, take, carry, coercive, compulsory, draconian, drastic, harsh, stern, stringent, emergency, extreme, radical, preventive, prophylactic, safety, security, stopgap, temporary

- ability, able, accurate, accurately, achievement, activity, additional, adopt, aim, angle, appropriate, approve, austerity, autonomy, ballot, brandy, broad, calculate, carefully, change, circumference, composition, conservation, considerable, control, corrective, cost-cutting, crude, cup, defeat, defensive, density, derive, and 158 more
Parameters: association measure

- Mutual Information (MI, Church & Hanks 1990)
- t-score (t, Church et al. 1991)
- $\text{MI}^2, \text{MI}^3, \text{MI}^4$ (Daille 1994) + $\text{MI}_{\text{conf}}$ (Johnson 2001)
- chi-squared ($X^2$) and z-score ($z$) with Yates correction
- Dice (SketchEngine), Jaccard coefficient
- minimum sensitivity ($\text{MS}$, Pedersen & Bruce 1996)
- odds ratio ($\log \theta, \log \theta_{\text{disc}}$), relative risk ($\log r$)
- log-likelihood ($G^2$, Dunning 1993)
- $\Delta P$ (Gries 2013) in 4 variants (fwd, bwd, min, max)
- co-occurrence frequency ($f$)
Parameters: co-occurrence context

- syntactic co-occurrence: dependency relations
  - direct dependency (all types, both directions)
- surface co-occurrence: L1 / R1
- surface co-occurrence: L2 / R2
- surface co-occurrence: L3 / R3
- surface co-occurrence: L5 / R5
- surface co-occurrence: L10 / R10
- textual co-occurrence: sentence
## Parameters: corpus & annotation

<table>
<thead>
<tr>
<th>corpus</th>
<th>annotation</th>
<th>size</th>
</tr>
</thead>
<tbody>
<tr>
<td>British National Corpus (BNC, Aston &amp; Burnard 1998)</td>
<td>C&amp;C, Stanford</td>
<td>0.1 G</td>
</tr>
<tr>
<td>Darmstadt English Movie Subtitle Corpus (DESC)</td>
<td>C&amp;C, Stanford</td>
<td>0.1 G</td>
</tr>
<tr>
<td>Gigaword newspaper corpus (2nd ed.)</td>
<td>C&amp;C, Stanford</td>
<td>2.0 G</td>
</tr>
<tr>
<td>English wikipedia of 2009 (Wackypedia)</td>
<td>C&amp;C, Malt, Stfd</td>
<td>1.0 G</td>
</tr>
<tr>
<td>Subcorpus WP500 (500 words per article)</td>
<td>C&amp;C, Malt, Stfd</td>
<td>0.2 G</td>
</tr>
<tr>
<td>Web corpus ukWaC (Baroni et al. 2009)</td>
<td>C&amp;C, Malt</td>
<td>2.0 G</td>
</tr>
<tr>
<td>Web corpus WebBase (Han et al. 2013)</td>
<td>C&amp;C</td>
<td>3.0 G</td>
</tr>
<tr>
<td>Web corpus UKCOW 2012 (Schäfer et al. 2012)</td>
<td>C&amp;C</td>
<td>4.0 G</td>
</tr>
<tr>
<td>Web corpus ENcow 2014</td>
<td>C&amp;C, Malt</td>
<td>10.0 G</td>
</tr>
<tr>
<td>All Web corpora + Wackypedia (JOINT)</td>
<td>C&amp;C</td>
<td>16.0 G</td>
</tr>
</tbody>
</table>
## Parameters: corpus & annotation

<table>
<thead>
<tr>
<th>corpus</th>
<th>annotation</th>
<th>size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Web 1T 5-Grams (Web1T5, Brants &amp; Franz 2006)</td>
<td></td>
<td>1000 G</td>
</tr>
<tr>
<td>Google Books N-Grams 2012 (BooksEN, Lin et al. 2012)*</td>
<td>parsed</td>
<td>500G</td>
</tr>
<tr>
<td>Google Books N-Grams 2012 GB (BooksGB)*</td>
<td>parsed</td>
<td>50 G</td>
</tr>
</tbody>
</table>

* Google Books data sets only include n-gram counts from contemporary books published in 1980 and later (evaluation on full 20th century yields very similar results)
Evaluation methodology

- Precision / recall of n-best lists for each node
  - strategy used by Uhrig & Proisl (2012)
  - task: determine most salient collocates for given node
  - results averaged over all 203 nodes

- Precision vs. recall for list of all candidate pairs
  - strategy used by Bartsch & Evert (2013), following Evert & Krenn (2001, 2005)
  - task: determine most salient collocational pairs
  
  we present results for this approach
Evaluation: global ranking

<table>
<thead>
<tr>
<th>node</th>
<th>collocate</th>
<th>$G^2$</th>
<th>BBI?</th>
</tr>
</thead>
<tbody>
<tr>
<td>minister</td>
<td>prime</td>
<td>111653.34</td>
<td>++</td>
</tr>
<tr>
<td>prime</td>
<td>minister</td>
<td>103587.58</td>
<td>—</td>
</tr>
<tr>
<td>authority</td>
<td>local</td>
<td>64395.65</td>
<td>++</td>
</tr>
<tr>
<td>take</td>
<td>place</td>
<td>43787.49</td>
<td>—</td>
</tr>
<tr>
<td>place</td>
<td>take</td>
<td>42551.75</td>
<td>++</td>
</tr>
<tr>
<td>set</td>
<td>up</td>
<td>37871.00</td>
<td>—</td>
</tr>
<tr>
<td>state</td>
<td>secretary</td>
<td>37588.70</td>
<td>—</td>
</tr>
<tr>
<td>door</td>
<td>open</td>
<td>37287.35</td>
<td>++</td>
</tr>
<tr>
<td>open</td>
<td>door</td>
<td>37193.15</td>
<td>—</td>
</tr>
<tr>
<td>head</td>
<td>shake</td>
<td>32301.90</td>
<td>++</td>
</tr>
</tbody>
</table>

most strongly associated node-collocate pairs (n = 10)

$P = \frac{5 \text{ TP}}{10 \text{ cand.}} = 50\%$

$R = \frac{5}{2845} \text{ TP} = 0.2\%$

TP = true positive
(according to BBI dictionary)
Evaluation: precision vs. recall | BBI

BNC [100M] | L5/R5 span | gold: BBI

coverage: 96.4%

precision (%) vs. recall | BBI

n = 1193
R = 10.0%
P = 23.9%
Evaluation: precision vs. recall | BBI

BNC [100M] | L5/R5 span | gold: BBI

coverage: 96.4%

\[ G^2 = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}} \]

log-likelihood (Dunning 1993)
Evaluation: precision vs. recall | BBI

BNC [100M] | L5/R5 span | gold: BBI

coverage: 96.4%

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

Mutual Information (Church & Hanks 1990)
Evaluation: precision vs. recall | BBI

BNC [100M] | L5/R5 span | gold: BBI

coverage: 96.4%

- $G^2$
- MI
- t

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

t-score
(Church et al. 1991)
Evaluation: precision vs. recall | BBI

BNC [100M] | L5/R5 span | gold: BBI

coverage: 96.4%

Dice coefficient (Sketch Engine)

\[
\text{Dice} = \frac{2O}{R_1 + C_1}
\]
Evaluation: precision vs. recall | BBI

**BNC [100M] | L5/R5 span | gold: BBI**

- **Coverage:** 96.4%
- **Baseline:** 0.30%
- **AP50:** 17.4%

**Average Precision up to 50% recall**
Evaluation: precision vs. recall | BBI

BNC [100M] | L5/R5 span | gold: BBI

coverage: 96.4%

- $G^2$
- MI
- t
- Dice

AP50 = 17.4%
AP50 = 18.8%

Average Precision up to 50% recall
Evaluation: precision vs. recall | BBI

BNC [100M] | L5/R5 span | gold: BBI

coverage: 96.4%

precision (%) vs. recall (%)

- \( G^2 \)
- \( MI \)
- \( t \)
- \( MI^2 \)

\[ MI^2 = \log_2 \frac{O^2}{E} \]

heuristic measure (Daille 1994)

Baseline = 0.30%
Evaluation: precision vs. recall | BBI

BNC [100M] | L5/R5 span | gold: BBI

coverage: 96.4%

- G²
- MI
- t
- MI²
- X²

\[ X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \]

standard chi-squared test
Evaluation: precision vs. recall | BBI

BNC [100M] | L5/R5 span | gold: BBI

coverage: 96.4%

Evaluation: precision vs. recall | BBI

collision
c-occurrence
c-frequency
for comparison

co-occurrence
frequency
for comparison

baseline = 0.30%
Evaluation: precision vs. recall | BBI

BNC [100M] l L5/R5 span l gold: BBI

coverage: 96.4%

\[ \Delta P_{2|1} = \frac{O_{11}}{R_1} - \frac{O_{21}}{R_2} \]

asymmetric \( \Delta P \)

(Allan 1980; Gries 2013)
Evaluation: precision vs. recall | BBI

BNC [100M] | L5/R5 span | gold: BBI

coverage: 96.4%

- G^2
- MI
- t
- MI^2
- X^2
- f
- ΔP_{21}
- Dice

precision (%) vs. recall (%)
Factor: context size | BBI

BNC [100M] | L3/R3 span | gold: BBI

coverage: 95.4%

G^2, MI, t, MI^2, X^2, f, ∆P_{21}, Dice

precision (%) vs recall (%)

baseline = 0.41%
Factor: context size | BBI

BNC [100M] | L2/R2 span | gold: BBI

coverage: 93.4%

precision (%) vs. recall (%)

baseline = 0.57%
Factor: context size | BBI

BNC [100M] | L1/R1 span | gold: BBI

coverage: 85.1%

precision (%) vs. recall (%) graph:
- Baseline = 1.08%
- Various measures including $G^2$, MI, $t$, $MI^2$, $X^2$, $f$, $\Delta P_{21}$, and Dice.
Factor: context size | BBI

BNC [100M] | syntactic | gold: BBI

- coverage: 91.6%

- baseline = 0.70%

- precision (%) vs. recall (%) graph with various lines representing different measures:
  - $G^2$
  - MI
  - t
  - $MI^2$
  - $X^2$
  - f
  - $\Delta P_{21}$
  - Dice

eLex 2017, Leiden, 20 Sep 2017
Factor: corpus | syntactic | BBI

Subtitles (DESC v2) [100M] | syntactic | gold: BBI

coverage: 80.9%

precision (%) vs recall (%)

G², MI, t, MI², X², f, ΔP₂₁, Dice

baseline = 1.13%
Factor: corpus | syntactic| BBI

Wackypedia [1G] I syntactic I gold: BBI

coverage: 97.0%

baseline = 0.32%

precison (%) vs recall (%)
Factor: corpus | syntactic | BBI

ukWaC [2G] | syntactic | gold: BBI

coverage: 98.7%

baseline = 0.23%

precision (%) vs. recall (%)
Factor: corpus | syntactic | BBI

UKCOW [4G] I syntactic I gold: BBI

coverage: 99.3%

precision (%) vs. recall (%)

baseline = 0.17%

G², MI, t², MI², X², f, ΔP₂₁, Dice
Joint Web [16G] | syntactic | gold: BBI

coverage: 99.8%

baseline = 0.11%

precision (%) vs. recall (%) for different metrics:
- G^2
- MI
- t
- MI^2
- X^2
- f
- ΔP_21
- Dice
Results: coverage

Coverage of BBI gold standard | L2/R2

Coverage (%)

BNC                  DESC                  Gigaword                WP500                  Wiki                  UKWAC                  WEBBASE                UKCOW                  ENCow                  JOINT                  WEB1T5                BooksGBmod                BooksENmod

0 20 40 60 80 100

\[ f \geq 1 \]
\[ f \geq 5 \]
\[ f \geq 50 \]
\[ f \geq 1000 \]
Results: coverage

Coverage of OCD gold standard | L2/R2

<table>
<thead>
<tr>
<th>Coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

- BNC
- DESC
- Gigaword
- WP500
- Wiki
- UKWAC
- WEBBASE
- UKCOW
- ENCOW
- JOINT
- WEB1T5
- BooksGBmod
- BooksENmod

Symbols:
- $f \geq 1$
- $f \geq 5$
- $f \geq 50$
- $f \geq 1000$
Results: context size

AP 50 | BNC | f ≥ 1

syntactic  L1/R1 span  L2/R2 span  L3/R3 span  L5/R5 span  L10/R10 span  sentence

AP50 (%)  

0 10 20 30 40 50 60 70

ocd  bbi
Results: context size

AP 50 | ENCow | f ≥ 1

- **AP50 (%)**
- **syntactic**
- **L1/R1 span**
- **L2/R2 span**
- **L3/R3 span**
- **L5/R5 span**
- **L10/R10 span**
- **sentence**

- **ocd**
- **bbi**

---
eLex 2017, Leiden, 20 Sep 2017
Results: context size

AP 50 | ENCOW | f ≥ 50

Syntactic L1/R1 span L2/R2 span L3/R3 span L5/R5 span L10/R10 span sentence

AP50 (%) 0 10 20 30 40 50 60 70

ocd
bbi
Results: corpus | L2/R2 span

AP 50 | L2/R2 span | f ≥ 1

<table>
<thead>
<tr>
<th>Corpus</th>
<th>OCD</th>
<th>BBI</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DESC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gigaword</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WP500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wiki</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UKWAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WEBBASE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UKCOW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENCOW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JOINT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WEB1T5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BooksGBmod</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BooksENmod</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

eLex 2017, Leiden, 20 Sep 2017
Results: frequency threshold

AP 50 | syntactic | gold: BBI

- $f \geq 1$
- $f \geq 3$
- $f \geq 5$
- $f \geq 10$
- $f \geq 20$
- $f \geq 50$
- $f \geq 100$
- $f \geq 500$
- $f \geq 1000$
- $f \geq 5000$
- $f \geq 10000$

AP50 (%)

BNC
Wiki
ENCOW
BooksENmod
Results: interactions

AP 50 \mid L2/R2 \text{ span} \mid f \geq 1 \mid \text{gold: BBI}

- G^2
- MI
- MI^2
- \Delta P_{2|1}
- Dice

- BNC
- DESC
- Gigaword
- WP500
- Wiki
- UKWAC
- WEBASE
- UKCOW
- ENcow
- JOINT
- WEB1T5
- BooksGBmod
- BooksENmod

eLex 2017, Leiden, 20 Sep 2017
Results: interactions

AP 50 | BNC | f ≥ 1 | gold: BBI

<table>
<thead>
<tr>
<th>syntactic</th>
<th>L1/R1 span</th>
<th>L2/R2 span</th>
<th>L3/R3 span</th>
<th>L5/R5 span</th>
<th>L10/R10 span</th>
<th>sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>G^2</td>
<td>MI</td>
<td>t</td>
<td>ΔP_{21}</td>
<td>MI^2</td>
<td>X^2</td>
<td></td>
</tr>
</tbody>
</table>

dice
Results: interactions

AP 50 \mid BNC \mid f \geq 5 \mid gold: BBI

<table>
<thead>
<tr>
<th>syntactic</th>
<th>L1/R1 span</th>
<th>L2/R2 span</th>
<th>L3/R3 span</th>
<th>L5/R5 span</th>
<th>L10/R10 span</th>
<th>sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP50 (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results: interactions

AP 50 | ENCOW | syntactic | gold: BBI

- $f \geq 1$
- $f \geq 3$
- $f \geq 5$
- $f \geq 10$
- $f \geq 20$
- $f \geq 50$
- $f \geq 100$
- $f \geq 500$
- $f \geq 1000$
- $f \geq 5000$
- $f \geq 10000$

AP50 (%)

G$^2$, MI, t, MI$^2$, $X^2$, $f$, $\Delta P_{2|1}$, Dice

eLex 2017, Leiden, 20 Sep 2017
Lost in results

- No space for complete results in talk or paper
  - you have to take our word that they look similar
- No room for the other 12 association measures
- AP50 may not be the most appropriate criterion
  - perhaps 30% recall sufficient, perhaps 80% needed
  - hides details of trade-off between precision and recall
  - do parameters affect the shape of P/R curves?
- Too much data for supplementary materials
  - 2.3 GiB of co-occurrence data (compressed)
  - gold standard cannot be redistributed
E-VIEW-alation

- Interactive Web-based viewer for P/R plots
- Gives user full control over evaluation parameters
  - make your own animations like those in the presentation

http://www.collocations.de/eviewalation/

👉 to be released as open-source software
Conclusions

- Small co-occurrence contexts are better
- Size matters, but also corpus quality
  - very large Web corpora outperform BNC
- Frequency threshold does not improve results
  - possibly due to focus on small number of nodes
- Virtually no interactions between parameters
  - corpus vs. context size vs. AM
  - most findings hold across both gold standards (except AM)
- Share all results with E-VIEW-alation!
Questions?

THANK YOU!
Appendix: Technicalities

- Corpora indexed with IMS CWB (Evert & Hardie 2011)
- Data extraction with UCS toolkit (Evert 2004)
- Evaluation & plots with UCS/R
- Precision-recall data exported as JSON files
- E-VIEW-alation
  - client: Vega 2.6 + JQuery UI
  - server: Perl CGI script serves requested JSON data
Appendix: Per-node evaluation

by node | British National Corpus [100M] | L5/R5 span (f >= 5) | gold: Bartsch224–BE

G²
MI
t
MI²
X²
f
ΔP₂₁

recall (%)
n–best list
Appendix: Per-node evaluation

by node | British National Corpus [100M] | L5/R5 span (f >= 5) | gold: Bartsch224–BE

Appendix: Per-node evaluation

by node | British National Corpus [100M] | L5/R5 span (f >= 5) | gold: Bartsch224–BE

Appendix: Per-node evaluation

by node | British National Corpus [100M] | L5/R5 span (f >= 5) | gold: Bartsch224–BE

Appendix: Per-node evaluation

by node | British National Corpus [100M] | L5/R5 span (f >= 5) | gold: Bartsch224–BE

Appendix: Per-node evaluation

by node | British National Corpus [100M] | L5/R5 span (f >= 5) | gold: Bartsch224–BE

Appendix: Per-node evaluation

by node | British National Corpus [100M] | L5/R5 span (f >= 5) | gold: Bartsch224–BE

Appendix: Per-node evaluation

by node | British National Corpus [100M] | L5/R5 span (f >= 5) | gold: Bartsch224–BE

Appendix: Per-node evaluation

by node | British National Corpus [100M] | L5/R5 span (f >= 5) | gold: Bartsch224–BE

Appendix: Per-node evaluation

by node | British National Corpus [100M] | L5/R5 span (f >= 5) | gold: Bartsch224–BE
Appendix: Manual validation

discrepancies between BBI / OCD2 and corpus data

<table>
<thead>
<tr>
<th>Bartsch 224, ENCOW, Malt dependencies, X2 'shake' (1000 candidates) [bartsch]</th>
<th>label for entry #3287924 set to FP</th>
<th>[export]</th>
<th>back to main page</th>
</tr>
</thead>
<tbody>
<tr>
<td>11477613 shake furiously 41 6971.790 --- OCD</td>
<td>TP ▼</td>
<td>Set</td>
<td></td>
</tr>
<tr>
<td>17066993 shake quake 42 6818.999 --- ---</td>
<td>TP ▼</td>
<td>Set</td>
<td></td>
</tr>
<tr>
<td>15079184 shake milk 43 6617.603 --- OCD (under collocate)</td>
<td>TP ▼</td>
<td>Set</td>
<td></td>
</tr>
<tr>
<td>17993610 shake snort 44 6486.647 --- ---</td>
<td>TP ▼</td>
<td>Set</td>
<td></td>
</tr>
<tr>
<td>17878103 shake scuttle 45 6073.180 --- ---</td>
<td>TP ▼ LSP</td>
<td>Set</td>
<td></td>
</tr>
<tr>
<td>16052416 shake pan 46 5520.637 --- ---</td>
<td>TP ▼</td>
<td>Set</td>
<td></td>
</tr>
<tr>
<td>7184709 shake cry 47 5308.692 --- ---</td>
<td>FP ▼</td>
<td>Set</td>
<td></td>
</tr>
<tr>
<td>1753930 shake ass 48 5103.859 --- ---</td>
<td>TP ▼</td>
<td>Set</td>
<td></td>
</tr>
<tr>
<td>11878568 shake grab 49 4952.691 --- ---</td>
<td>--- ▼</td>
<td>Set</td>
<td></td>
</tr>
<tr>
<td>10226053 shake explosion 50 4783.420 --- OCD (under collocate)</td>
<td>TP ▼</td>
<td>Set</td>
<td></td>
</tr>
<tr>
<td>17994048 shake swarm 51 4777.921 --- ---</td>
<td>TP ▼ LSP</td>
<td>Set</td>
<td></td>
</tr>
<tr>
<td>3287924 shake body 52 4619.348 --- OCD (under collocate)</td>
<td>FP ▼</td>
<td>Set</td>
<td></td>
</tr>
</tbody>
</table>

variable ordering: A..B 68.2% / B..A 31.8%  non-contiguous: 90.9% adjacent

The GTC’s chassis, the stiffest of any convertible in the world, provides a firm foundation for suspension control, minimising scuttle shake and contributing to the GTC’s refined handling.

They do handle well, have only driven the cab which to be honest can suffer from scuttle shake on poorly surfaced roads but it still handles way better than my previous Saab 9-3 Aero ( later shape on Hirsch springs ), god only knows what a well fertled Coupe on coilovers, new bushes and uprate ARB ’s is like , suspect not much modern stuff is any better.

Ride quality is generally acceptable for this class of car but there are signs of scuttle shake if you hit mid-corner bumps at speed.

Only the very occasional sound of movement between the tiny rear quarter window and the side window rubber betrays any hint of convertible scuttle shake .

The latest round of negotiations ends in just 4 days - but outcries in each of our countries could shake the confidence of negotiators and scuttle
### Appendix: Manual validation
discrepancies between BBI and OCD2

<table>
<thead>
<tr>
<th>Bartsch 224, ENCOW, Malt dependencies, X2 'argue' (1000 candidates)</th>
<th>[export] back to main page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1535292</strong></td>
<td><strong>argue</strong> convincingly</td>
</tr>
<tr>
<td><strong>1536640</strong></td>
<td><strong>argue</strong> forcefully</td>
</tr>
<tr>
<td><strong>1539055</strong></td>
<td><strong>argue</strong> plausibly</td>
</tr>
<tr>
<td><strong>1534717</strong></td>
<td><strong>argue</strong> case</td>
</tr>
<tr>
<td><strong>1539037</strong></td>
<td><strong>argue</strong> plaintiff</td>
</tr>
<tr>
<td><strong>1540747</strong></td>
<td><strong>argue</strong> strongly</td>
</tr>
<tr>
<td><strong>1535589</strong></td>
<td><strong>argue</strong> defendant</td>
</tr>
<tr>
<td><strong>1534170</strong></td>
<td><strong>argue</strong> author</td>
</tr>
<tr>
<td><strong>1538856</strong></td>
<td><strong>argue</strong> passionately</td>
</tr>
<tr>
<td><strong>1536059</strong></td>
<td><strong>argue</strong> economist</td>
</tr>
<tr>
<td><strong>1540725</strong></td>
<td><strong>argue</strong> strenuously</td>
</tr>
<tr>
<td><strong>989811</strong></td>
<td><strong>argue</strong> also</td>
</tr>
</tbody>
</table>

**variable ordering: A..B 74.5% / B..A 25.5% non-contiguous: 79.2% adjacent**

Although no one has argued that amendments other than the Tenth and Fourteenth have impliedly amended the AC, one might **plausibly argue** that each has necessarily done so.

It cannot **plausibly be argued**, in my opinion, that the Human Rights Act erodes the sovereignty of Parliament or amounts to a usurpation of power by the judges.

Rather than viewing them as epistemically morally-politically pernicious forms of hasty generalization by contrast, say, with Gadamerian pre-judgements or putatively more benign practices of categorization, Fricker **argues plausibly** for a "neutral" sense of stereotype which catches their frequent reliability as part of a "hearer's rational resources" in making credibility judgements.

But, it can be **plausibly argued** that we are getting there.

Given any statement, we can **argue plausibly** that it is about Maine.


References

References


- Gries, Stefan Th. (2013). 50-something years of work on collocations: What is or should be next ... . *International Journal of Corpus Linguistics, 18*(1), 137–165.


References

- Pedersen, Ted and Bruce, Rebecca (1996). What to infer from a description. Technical Report 96-CSE-04, Southern Methodist University, Dallas, TX.