An NLP Approach to the Evaluation of Web Corpora

Stefan Evert

Corpus Linguistics Group, Department Germanistik & Komparatistik
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

stefan.evert@fau.de

Leuven, 17 February 2015

Collaborators

parts of this presentation are based on the following studies

- Biemann, Chris; Bildhauer, Felix; Evert, Stefan; Goldhahn, Dirk; Quasthoff, Uwe; Schäfer, Roland; Simon, Johannes; Swiezinski, Leonard; Zesch, Torsten (2013). Scalable construction of high-quality Web corpora. *Journal for Language Technology and Computational Linguistics (JLCL)*, 28(2), 23–59.


- Bartsch, Sabine; Evert, Stefan; Proisl, Thomas; Uhrig, Peter (2015). (Association) measure for measure: comparing collocation dictionaries with co-occurrence data for a better understanding of the notion of collocation. Presentation at ICAME 36 Conference, Trier, Germany.

Introduction

The Web as Corpus

Why Web corpora?

because more data are better data (Church and Mercer 1993)

- Properties of the Web
  - Internet English, distribution of Web genres, hyperlink graph
  - Web corpus = random sample of the (public) WWW

- Computer-mediated communication (CMC)
  - Twitter, Facebook, chatroom logs, discussion groups, . . .
  - many Web genres share aspects of interactive CMC
  - Web corpus = targeted collection of CMC genres

- As replacement for linguistic reference corpora
  - main goal of the early WaC(ky) community
  - cheaper, larger and more up-to-date than traditional corpora
  - Web corpus should be similar to reference corpus

- Scaling up NLP training data (Banko and Brill 2001)
  - 1964: 1 million words (Brown Corpus)
  - 1995: 100 million words (British National Corpus)
  - 2003: 1,000+ million words (English Gigaword, WaCky)
  - 2006: 1,000,000 million words (Google Web 1T 5-Grams)

Is bigger always better?

- From small, clean and well designed . . .
  - British National Corpus (BNC)
  - movie subtitles, newspapers, . . .

- . . . to large and messy . . .
  - WaCky, WebBase, COW, TenTen, GloWbE, Aranea, . . .
  - sampling frame unclear, lack of metadata
  - boilerplate, duplicates, non-standard language

- . . . to huge n-gram databases
  - largest corpora only available as n-gram databases, e.g.
    - Google’s 1-trillion-word Web corpus (Web 1T 5-Grams)
    - tend to be even messier, often w/o linguistic annotation
    - lack of context, incomplete because of frequency threshold
The Google Web 1T 5-Gram database
Brants and Franz (2006)

<table>
<thead>
<tr>
<th>word 1</th>
<th>word 2</th>
<th>word 3</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>supplement</td>
<td>depend</td>
<td>on</td>
<td>193</td>
</tr>
<tr>
<td>supplement</td>
<td>depending</td>
<td>on</td>
<td>174</td>
</tr>
<tr>
<td>supplement</td>
<td>depends</td>
<td>entirely</td>
<td>94</td>
</tr>
<tr>
<td>supplement</td>
<td>depends</td>
<td>on</td>
<td>338</td>
</tr>
<tr>
<td>supplement</td>
<td>derived</td>
<td>from</td>
<td>2668</td>
</tr>
<tr>
<td>supplement</td>
<td>des</td>
<td>coups</td>
<td>77</td>
</tr>
<tr>
<td>supplement</td>
<td>described</td>
<td>in</td>
<td>200</td>
</tr>
</tbody>
</table>

This looks very much like a relational database table
So why not just put the data into an off-the-shelf RDBMS?
▶ built-in indexing for quick access
▶ powerful query language SQL

Web1T5-Easy database encoding procedure
Pre-processing (normalisation, filtering, ...)
⇓
Numeric ID coding & database insertion [1d 23h]
⇓
Collapse duplicate rows (from normalisation) [6d 7h]
⇓
Indexing of each n-gram position [3d 2h]
⇓
Statistical analysis for query optimisation [not useful]
⇓
Build database of co-occurrence frequencies [ca. 3d]
⇓

211 GiB
Carried out in spring 2009 on quad-core Opteron 2.6 GHz with 16 GiB RAM
— should be faster on state-of-the-art server with latest version of SQLite.

Use numeric ID coding as in IR / large-corpus query engines
More efficient to store, index and sort in RDBMS
Frequency-sorted lexicon is beneficial for variable-length coding of integer IDs (used by SQLite)
Querying the database

It’s easy to search the database for patterns like

```
association ... %al Y
```

with a “simple” SQL query:

```
SELECT w3, w4, SUM(f) AS freq FROM ngrams
WHERE w1 IN (SELECT id FROM vocab WHERE w='association')
AND w3 IN (SELECT id FROM vocab WHERE w LIKE '%al')
GROUP BY w3, w4 ORDER BY freq DESC;
```

Web1T5-Easy implements a more user-friendly query language:

```
association ? %al *
```

Web1T5-Easy query performance

<table>
<thead>
<tr>
<th>Web1T5-Easy query</th>
<th>cold cache</th>
<th>warm cache</th>
</tr>
</thead>
<tbody>
<tr>
<td>corpus linguistics</td>
<td>0.11s</td>
<td>0.01s</td>
</tr>
<tr>
<td>web as corpus</td>
<td>1.29s</td>
<td>0.44s</td>
</tr>
<tr>
<td>time of *</td>
<td>2.71s</td>
<td>1.09s</td>
</tr>
<tr>
<td>%ly good fun</td>
<td>181.03s</td>
<td>24.37s</td>
</tr>
<tr>
<td>[sit,sits,sat,sitting] * ? chair</td>
<td>1.16s</td>
<td>0.31s</td>
</tr>
<tr>
<td>* linguistics (association ranking)</td>
<td>11.42s</td>
<td>0.05s</td>
</tr>
<tr>
<td>university of * (association ranking)</td>
<td>1.48s</td>
<td>0.48s</td>
</tr>
</tbody>
</table>

(64-bit Linux server with 2.6 GHz AMD Opteron CPUs, 16 GiB RAM and fast local hard disk; based on timing information from the public Web interface.)

Evaluating the “quality” of Web corpora

- Statistical properties
  - type-token distributions, n-gram frequencies, other markers
  - representativeness (as sample of the Web)
  - genre distribution (traditional vs. Web genres)
- Corpus comparison
  - between Web corpora (reliability)
  - between Web corpus and reference corpus
  - compared to within-corpus variation
- Training data for NLP application
  - larger amount of training data is often beneficial
  - confounding factors (NLP algorithm, training regime, . . .)
- Linguistic evaluation of Web corpora
  - as substitute for / extension of reference corpus
  - need linguistic tasks that can be judged quantitatively and that make immediate use of corpus frequency data
Linguistic evaluation of Web corpora

- **Frequency comparison**
  - “good” Web corpora should agree with reference corpus on core phenomena
  - e.g. Basic English vocabulary, compound nouns, ...

- **Identification of multiword expressions (MWE)**
  - well-know NLP task based on co-occurrence statistics
  - some gold standard data sets available
  - e.g. “phrasal verbs”, lexical collocations, ...

- **Distributional semantic models (DSM)**
  - hypothesis: semantic similarity ~ distributional similarity
  - distribution quantified by co-occurrences with other words
  - DSMs can be evaluated in various shared tasks

### Research questions

- Are English Web corpora a substitute for the BNC?
- What are the differences between Web corpora?
- Does size matter more than content?
- How useful are n-gram databases?
  - esp. detrimental effects of frequency thresholds
- How important is (automatic) linguistic annotation?
- Do Web corpora offer better coverage?

### Corpora in the evaluation

- **Ref**: British National Corpus (Aston and Burnard 1998) C&C 0.1 G
- **English Movie Subtitles (DESC v2)** C&C 0.1 G
- **Gigaword newspaper corpus (2nd edition)** 2.0 G
- **English Wackypedia Malt** 1.0 G
- **Wackypedia subset (WP500)** Malt 0.2 G
- **ukWaC (Baroni et al. 2009)** Malt 2.0 G
- **WebBase (Han et al. 2013)** 3.0 G
- **UKCOW 2012 (Schäfer and Bildhauer 2012)** 4.0 G
- **Joint Web corpus** 10.0 G
- **Web 1T 5-Grams (Brants and Franz 2006)** 1000.0 G
- **LCC n-gram database** 1.0 G
- **ENCOW 2014 Malt** 10.0 G
- **Google Books 2012 EN (Lin et al. 2012)** Malt 900.0 G
- **Google Books 2012 GB Malt** 100.0 G
Comparison of frequency counts

- Scatterplots of (log) frequencies in BNC vs. other corpora
  - Pearson correlation $r$ from regression $f_{\text{WebC}} \sim \beta \cdot f_{\text{BNC}}$ etc.
  - only consider items that occur in both corpora
  (as low coverage is not penalized directly)

- Test data sets
  - Basic English words (lemmatized)
  - inflected forms of Basic English words
  - binary compound nouns extracted from WordNet 3.0

- Morphological query expansion for unannotated n-grams

<table>
<thead>
<tr>
<th>Query</th>
<th>$f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>hear sound</td>
<td>36,304</td>
</tr>
<tr>
<td>[hear, hears, heard, hearing]</td>
<td>95,453</td>
</tr>
</tbody>
</table>

(dashed lines indicate acceptable frequency difference within one order of magnitude)

S. Evert (stefan.evert@fau.de)
MWE Identification & Collocations

Collocations

- Collocation: frequent co-occurrence within short span of up to 5 words (Firth 1957; Sinclair 1966, 1991)
  - plays important role in lexicography, corpus linguistics, language description, word sense disambiguation, ...
  - key feature for MWE identification
  - collocation database is also a sparse representation of a distributional semantic model (term-term matrix)

- Web1T5 only provides exact co-occurrence frequencies for immediately adjacent bigrams (e.g. *day* and *day*)
- Approximate counts for distance n from n + 1-gram table
  - day ? ? * and * ? ? day
  - quasi-collocations

Quasi-collocations database

- Web1T5-Easy: pre-compiled database of quasi-collocations
  - brute-force, multi-pass algorithm
  - runtime approx. 3 days on server with 16 GiB RAM
- Flexible collocational span L4, ..., L1 / R1, ..., R4
  - separate count for each collocate and position
  - co-occurrence frequency in user-defined span and association scores are calculated on the fly
  - benefits from tight integration of Perl & SQLite
- Standard association measures: $X^2$, $G^2$, $t$, MI, Dice

Quasi-collocations demo

Collocates of “corpus” (f=5137372)
Evaluation on English VPC extraction task
(Baldwin 2008)

- English verb-particle constructions (VPC) consisting of head verb + one obligatory prepositional particle
  - hand in, back off, wake up, set aside, carry on, ...
- Data set of 3,078 candidate VPC types
  - extracted from written part of BNC with combination of tagger-, chunker-, and parser-based methods
- Manually annotated as compositional / non-compositional
  - baseline: 14.3% non-compositional VPC (440 / 3078)
  - compositional: carry around, fly away, refer back, ...
- Evaluation: candidate ranking based on each corpus
  - surface co-occurrence (L0,R3) + POS filter (except Web1T5)
  - standard association measures: $G^2$, $t$, MI, Dice, $X^2$, $f$
- Evaluation results:
  - Average Precision: $G^2$: 31.06%, Dice: 30.13%, $X^2$: 32.12%
  - n-best list ($n = 12$)
    - $P = \frac{5}{12} = 41.7\%$
    - $R = \frac{5}{440} = 1.1\%$

---

Evaluation on English VPC extraction task
(Baldwin 2008)
Why does Web1T5 perform so badly in this task despite its size? Possible explanations include:

- Co-occurrence counts are underestimated for larger windows because of frequency threshold in n-gram database
  → quasi-collocations as poor approximation
- No part-of-speech annotation available to filter candidates

Identification of lexical collocations as habitual, recurrent word combinations (Firth 1957)

- essential for advanced learners (idiomatic English)
- different from lexicalised MWE
- semi-compositional or fully compositional
- no clear-cut linguistic criteria or tests available

Gold standard: BBI combinatory dict. (Benson et al. 1986)

- BBI was compiled manually based on lexicographer intuitions
- lexical collocations automatically extracted from scanned BBI, manually checked for 224 selected node words (Bartsch224)
Evaluation on BBI collocation identification task
(Benson et al. 1986)

- Candidate extraction and ranking
  - for each of the 224 nodes, extract all co-occurrences with the 7,711 words that occur as collocates somewhere in the BBI
  - full list of candidates ranked according to standard association measures (not per individual node) ➔ precision-recall graphs
  - composite: AP50 = average precision up to 50% recall, selecting best measure for each data set
- Dictionary-based evaluation problematic (Evert 2004, 139f)
  - provides lower bound on n-best precision
  - coverage of native speaker intuitions by corpus data
  - may be biased against recent corpora, Web texts, etc.
- Manual validation of ranked collocates for selected nodes
  - work in progress, using custom Web-based annotation tool

Evaluation on BBI: collocational span (Benson et al. 1986)

Additional graphs showing precision and recall for different corpora:
- British National Corpus [100M] syntactic dependency | gold: BBI
  - coverage: 91.7%
  - baseline = 1.31%

  - coverage: 98.9%
  - baseline = 0.26%
Evaluation results

MWE identification

Replication on Oxford Collocations Dictionary (2nd ed.)
(McIntosh et al. 2009)

- Criticism against BBI
  - lexicographer intuitions may be incomplete
  - outdated (published 1986, native speakers primed in 1950s)
  - biased against recent corpora
- Additional gold standard: OCD2 (McIntosh et al. 2009)
  - corpus-based + manual validation (1st ed. was based on BNC)
  - up to date, covers much broader range of collocates

Additional gold standard: OCD2 (McIntosh et al. 2009)

- automatic extraction from XML version of the dictionary
- collocates explicitly marked → no manual validation necessary
- Bartsch224 nodes accepted as headwords and as collocates
  (OCD follows Hausmann’s (1989) base-collocate distinction)

Candidate extraction and ranking
- same candidate data as for BBI experiments (lazy researcher)
- incomplete coverage of OCD2 gold standard: 4,636 out of 18,515 collocations discarded (= 25.0%)

Evaluation on OCD2: collocational span
(McIntosh et al. 2009)

![Graph of evaluation results](image1)

Evaluation on OCD2: different corpora (L3/R3)
(McIntosh et al. 2009)

![Graph of evaluation results](image2)
Distributional Semantics

Distributional semantics

- A computer can (sometimes) do the same, with sufficient amounts of corpus data and full collocational profiles

<table>
<thead>
<tr>
<th></th>
<th>get</th>
<th>see</th>
<th>use</th>
<th>hear</th>
<th>eat</th>
<th>kill</th>
</tr>
</thead>
<tbody>
<tr>
<td>w1</td>
<td>51</td>
<td>20</td>
<td>84</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>w2</td>
<td>52</td>
<td>58</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>26</td>
</tr>
<tr>
<td>w3</td>
<td>115</td>
<td>83</td>
<td>10</td>
<td>42</td>
<td>33</td>
<td>17</td>
</tr>
<tr>
<td>w4</td>
<td>59</td>
<td>39</td>
<td>23</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>w5</td>
<td>98</td>
<td>14</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>w6</td>
<td>12</td>
<td>17</td>
<td>3</td>
<td>2</td>
<td>9</td>
<td>27</td>
</tr>
<tr>
<td>banana</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>18</td>
<td>0</td>
</tr>
</tbody>
</table>

- sim(???, knife) = 0.770

Distributional semantics

- Distributional hypothesis (Harris 1954): meaning of a word can be inferred from its distribution across contexts
  “You shall know a word by the company it keeps!” — (Firth 1957)

- Reality check: What is the mystery word?
  - He handed her her glass of XXXXX.
  - Nigel staggered to his feet, face flushed from too much XXXXX.
  - Malbec, one of the lesser-known XXXXX grapes, responds well to Australia’s sunshine.
  - I dined off bread and cheese and this excellent XXXXX.
  - The drinks were delicious: blood-red XXXXX as well as light, sweet Rhenish.

- XXXXX = claret
  - all examples from BNC (carefully selected & slightly edited)
A computer can (sometimes) do the same, with sufficient amounts of corpus data and full collocational profiles.

- **knife**
  - get: 51
  - see: 20
  - use: 84
  - hear: 0
  - eat: 3
  - kill: 0

- **cat**
  - get: 52
  - see: 58
  - use: 4
  - hear: 4
  - eat: 6
  - kill: 26

- **???**
  - get: 115
  - see: 83
  - use: 10
  - hear: 42
  - eat: 33
  - kill: 17

- **boat**
  - get: 59
  - see: 39
  - use: 23
  - hear: 4
  - eat: 0
  - kill: 0

- **cup**
  - get: 98
  - see: 14
  - use: 6
  - hear: 2
  - eat: 1
  - kill: 0

- **pig**
  - get: 12
  - see: 17
  - use: 3
  - hear: 2
  - eat: 9
  - kill: 27

- **banana**
  - get: 11
  - see: 2
  - use: 2
  - hear: 0
  - eat: 18
  - kill: 0

\[
sim(???, \text{pig}) = 0.939
\]

- **knife**
  - get: 51
  - see: 20
  - use: 84
  - hear: 0
  - eat: 3
  - kill: 0

- **cat**
  - get: 52
  - see: 58
  - use: 4
  - hear: 4
  - eat: 6
  - kill: 26

- **???**
  - get: 115
  - see: 83
  - use: 10
  - hear: 42
  - eat: 33
  - kill: 17

- **boat**
  - get: 59
  - see: 39
  - use: 23
  - hear: 4
  - eat: 0
  - kill: 0

- **cup**
  - get: 98
  - see: 14
  - use: 6
  - hear: 2
  - eat: 1
  - kill: 0

- **pig**
  - get: 12
  - see: 17
  - use: 3
  - hear: 2
  - eat: 9
  - kill: 27

- **banana**
  - get: 11
  - see: 2
  - use: 2
  - hear: 0
  - eat: 18
  - kill: 0

\[
sim(???, \text{cat}) = 0.961
\]

**DSM with Web1T5: nearest neighbours**

- Neighbours of **linguistics** (cosine angle):
  - sociology (24.6), sociolinguistics (24.6), criminology (29.5), anthropology (30.8), mathematics (31.2), phonetics (33.1), phonology (33.2), philology (33.2), literatures (33.5), gerontology (35.3), prosemear (35.5), geography (35.8), humanities (35.9), archaeology (35.9), science (36.5), ...

- Neighbours of **spaniel** (cosine angle):
  - terrier (23.0), schnauzer (26.5), pinscher (27.0), weimaraner (28.3), keeshond (29.1), pomeranian (29.4), pekingese (29.6), bichon (30.1), vizsla (30.5), labradoodle (30.6), apso (31.1), spaniels (32.0), frise (32.0), yorkie (32.1), sheepdog (32.3), dachshund (32.4), retriever (32.7), whippet (32.9), havanese (33.1), westie (34.5), mastiff (34.7), dandie (34.7), chihuahua (34.9), dinnmont (35.0), elkhood (35.0), ...
Evaluating distributional similarity

- Correlation with human similarity ratings
  - RG65: Rubenstein and Goodenough (1965)
- Multiple-choice tasks
  - TOEFL synonym questions
  - SAT analogy questions
- Decision tasks for consistent/inconsistent prime
  - SPP: Semantic Priming Project (Hutchison et al. 2013)
  - GEK: Generalized Event Knowledge (Ferretti et al. 2001; McRae et al. 2005; Hare et al. 2009)
- Noun clustering vs. semantic classification
  - AP: Almuhareb/Poesio (Almuhareb 2006)
  - Battig: Battig/Montague norms (Van Overschelde et al. 2004)
  - ESSLLI: basic-level concrete nouns (ESSLLI '08 shared task)
- Psycholinguistic norms & experiments
  - free association norms
  - property norms
  - priming effects ($\Delta$RT)

Correlation with human similarity ratings

- Direct comparison with semantic similarity ratings (WordSim-353, Finkelstein et al. 2002)
  - 353 noun-noun pairs with “relatedness” ratings
  - rated on scale 0–10 by 16 test subjects
  - closely related: money/cash, soccer/football, type/kind, ...
  - unrelated: king/cabbage, noon/string, sugar/approach, ...
  - NB: not all “nouns” are nouns in a traditional sense (five, live, eat, stupid, …)
- Correlation with DSM distances for different corpora
  - DSM parameters: term-term matrix, L4/R4 surface window, 30k feature terms, log $G^2$ weighting, cosine similarity, SVD to 500 dimensions / 50 skipped (Lapesa and Evert 2014)
  - quantitative measure: Spearman’s rank correlation $\rho$ (robust against non-linearities)
Correlation with human relatedness ratings
(Finkelstein et al. 2002)

![WordSim353: British National Corpus](image)

|rho| = 0.644, p = 0.0000, |r| = 0.500 .. 0.640 (1 pairs not found)

distance vs. human rating

Why is Web1T5 so terrible?

Which words are semantically similar to **hot** (in DSM)?

- I hope there are no minors in the room!

  - big (29.5), butt (31.1), ass (31.1), wet (31.2), naughty (31.6), pussy (31.6), sexy (31.6), chicks (32.0), cock (32.2), ebony (32.3), fat (32.4), girls (32.4), asian (32.7), cum (33.1), babes (33.2), dirty (33.2), bikini (33.3), granny (33.4), teen (33.8), pics (33.8), gras (34.1), fucking (34.1), galleries (34.2), fetish (34.3), babe (34.3), blonde (34.5), pussies (34.5), whores (34.6), fuck (34.6), horny (34.7)

Please don’t ask about cats and dogs . . .
Result overview: Frequency comparison

**Frequency Comparison: Basic English (lemmatized)**

Pearson correlation

- DESC
- Gigaword
- WP500
- Wackypedia
- ukWaC
- WebBase
- UKCOW
- Joint Web
- LCC
- Web1T5

<table>
<thead>
<tr>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>

S. Evert (stefan.evert@fau.de) Evaluation of Web Corpora 17 Feb 2015 53 / 64

---

Result overview: MWE identification

**MWE Extraction: English VPC (L0/R3 span)**

Average precision (best AM)

- BNC
- DESC
- Gigaword
- WP500
- Wackypedia
- ukWaC
- WebBase
- UKCOW
- Joint Web
- LCC
tagged LCC
- Web1T5

<table>
<thead>
<tr>
<th>20</th>
<th>25</th>
<th>30</th>
<th>35</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>

S. Evert (stefan.evert@fau.de) Evaluation of Web Corpora 17 Feb 2015 54 / 64

---

Result overview: BBI collocations

**AP 50 | L3/R3 span | gold: BBI**

- BNC
- DESC
- Gigaword
- WP500
- Wiki
- UKWAC
- WEBBASE
- UKCOW
- JOINT
- ENCOW
- LCC
- LCC5
- LCC10
- WEB1T5
- BooksGB
- BooksGBwf

<table>
<thead>
<tr>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>●</td>
<td></td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>

S. Evert (stefan.evert@fau.de) Evaluation of Web Corpora 17 Feb 2015 55 / 64

---

Result overview: BBI collocations

**AP 50 | syntactic dependency | gold: BBI**

- BNC
- DESC
- Gigaword
- WP500
- Wiki
- UKWAC
- ENCOW
- ENCOWgp
- BooksGB
- BooksGBwf
- BooksEN
- BooksENwf

<table>
<thead>
<tr>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>●</td>
<td></td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>

S. Evert (stefan.evert@fau.de) Evaluation of Web Corpora 17 Feb 2015 55 / 64
### Result overview: OCD2 collocations

**AP 50 | L3/R3 span | gold: OCD2**

- **AP50 (%)**
  - BNC
  - DESC
  - Gigaword
  - WikipediA
  - UKWaC
  - joint
  - LCC
  - LCC[f >= 5]
  - LCC[f >= 10]
  - WEB1T5
  - BooksGB
  - BooksGBwf

### Result overview: Distributional semantics

**DSM Evaluation: WordSim−353**

- **rank correlation ρ (%)**
  - BNC
  - DESC
  - Gigaword
  - WP500
  - Wackypedia
  - ukWaC
  - WebBase
  - UKCOW
  - joint Web
  - LOC
  - LOC[f >= 5]
  - LOC[f >= 10]
  - WEB1T5

**DSM Evaluation: multiple choice (SPP)**

- **accuracy**
  - BNC
  - DESC
  - Gigaword
  - WP500
  - Wackypedia
  - ukWaC
  - WebBase
  - UKCOW
  - joint Web
  - LOC
  - LOC[f >= 5]
  - LOC[f >= 10]
  - WEB1T5
Result overview: Distributional semantics

The diagram shows the DSM Evaluation: noun clustering (AP) for various corpora. The x-axis represents different corpora such as BNC, DESC, Gigaword, WP500, Wackypedia, ukWaC, WebBase, UKCOW, Joint Web, LCC, LCC [f >= 5], LCC [f >= 10], and Web1T5. The y-axis represents cluster purity in percentages ranging from 40 to 80.

References I


References II


Biemann, Chris; Bildhauer, Felix; Evert, Stefan; Goldhahn, Dirk; Quasthoff, Uwe; Schäfer, Roland; Simon, Johannes; Swiezinski, Leonard; Zesch, Torsten (2013). Scalable construction of high-quality Web corpora. Journal for Language Technology and Computational Linguistics (JLCL), 28(2), 23–59.


References III

Evert, Stefan (2010). Google Web 1T5 n-grams made easy (but not for the computer). In Proceedings of the 6th Web as Corpus Workshop (WAC-6), pages 32–40, Los Angeles, CA.


References IV

Hare, Mary; Jones, Michael; Thomson, Caroline; Kelly, Sarah; McRae, Ken (2009). Activating event knowledge. Cognition, 111(2), 151–167.


Hutchison, Keith A.; Balota, David A.; Neely, James H.; Cortese, Michael J.; Cohen-Shikora, Emily R.; Tse, Chi-Shing; Yap, Melvin J.; Bengson, Jesse J.; Niemeyer, Dale; Buchanan, Erin (2013). The semantic priming project. Behavior Research Methods, 45(4), 1099–1114.

References V


McRae, Ken; Hare, Mary; Elman, Jeffrey L.; Ferretti, Todd (2005). A basis for generating expectancies for verbs from nouns. Memory & Cognition, 33(7), 1174–1184.


References VI
