What do learners actually do with corpus data?

Friday 7th June
14.15 - 15.00
Tom Cobb
Université du Québec à Montréal

Get this PPT
at https://lextutor.ca/bawe_conf_tom.pptx (6.5 Mb)
Looking over old PPTS

• Presentations about computers in language learning used to be quasi-promotional ‘CALL’

• Promotion is no longer needed
  – All over education, computer as learning tool is here to stay, for better or worse
    • Facebook’s “Summit” Saga

– And in language learning...
  • Rosetta, Babel, Duoling, One grain of rice, Memrise...
  • Probably many current learners’ main experience of instructed learning
However...

• Many computational language learning tools have good record keeping & motivational/game elements but simplistic learning models
  • Mainly single-word learning by imitation and association
    – OK as far as it goes
    – Zero research
      » (other than market research)
• Now what needs promotion is **what to do** with computers in learning
  – Through integration with SLA research
One CALL idea with some research behind it

• “Data driven language learning”
  – As a proven, research-based, computer-assisted component of a language learning program

• Meta-analysis on DDL in LL (2017)
  – Over 20 years, many uses, and varied environments...
    • Within-groups effect size (pre-post) effect size: $d = 1.5$
    • Between-groups e.s. (experimental-control) : $d = .95$
    • Both “strong for the field” (~Plonsky)

• And not exactly a specialty interest
  – *Language Learning*’s most-cited paper in 2017
EMPIRICAL STUDY

Corpus Use in Language Learning: A Meta-Analysis

Alex Boulton and Tom Cobb
Université de Lorraine and Université du Québec à Montréal

This study applied systematic meta-analytic procedures to summarize findings from experimental and quasi-experimental investigations into the effectiveness of using the tools and techniques of corpus linguistics for second language learning or use, here referred to as data-driven learning (DDL). Analysis of 64 separate studies representing 88 unique samples reporting sufficient data indicated that DDL approaches result in large overall effects for both control/experimental group comparisons ($d = 0.95$) and for pre/posttest designs ($d = 1.50$). Further investigation of moderator variables revealed that small effect sizes were generally tied to small sample sizes. Research has barely begun in some key areas, and durability/transfer of learning through delayed posttesting remains an area in need of further investigation. Although DDL research demonstrably improved over the period investigated, further changes in practice and reporting are recommended.
What is Data-Driven Learning?
What questions does it still raise?

The source of language acquisition is...

intuition
input
What is Data-Driven Learning?

The source of language acquisition is...

intuition

input

natural input

input as data
What is Data-Driven Learning?

source of language acquisition

intuition

natural input

expert inspected

input

input as data

learner inspected
What is Data-Driven Learning?

source of language acquisition

intuition

natural input

input as data

input

expert inspected

learner inspected
What is Data-Driven Learning?

• Lextutor’s *Web Concordancer* is an extended test of the ‘learner inspected’ model
  – A free resource, easy access, including mobile phone, used by about 5,000 users per week
    • Presumably teachers & their students
      – (more on this later)
8 Examples of “input as data,” or, learnability-enhanced input transformed by computer software

• Note:
  – this means where the input is significantly processed by software
    • Not just Skype, e-pen-pals, Gutenberg or Google books, online submission
      – Valuable as those may be
Enhancement 1: Data assembly

- Problem: if a pattern is not clear from one piece of input...

the issue by granting an injunction in advance of publication."

If the plaintiff can prove immediately and convincingly that the defendant is intending to publish palpable untruths, an injunction could be granted. Otherwise, the rule against prior restraint must prevail in libel actions.
... it may be clear from several

attraction. Their movement consisted
of magnetized bodies magnetizing one
another." As for the fedayeen, they
"observed a smiling rigour. But the
eroticism was palpable. I could
sense its vibrations, though I wasn't
bothered by it" (p.126). But this
eroticism, and the masculinity of these
men, frequently confounds the

in present tariff charges upon the
necessaries of life. The benefits of
such a reduction would be palpable and
substantial, seen and felt by thousands
who would be better fed and better
clothed and better sheltered. These
gifts should be the willing benefactions
of a Government whose highest function
Enhancement 2: Useful effort-level raising

• Problem: Language work needs cognitive engagement to be effective
  – Involvement load hypothesis
    • Laufer & Hulsteijn (2001)

• Yet just raising effort *per se* is not efficient
  – Writing lines?
  – Committing texts to memory?

• Solution:
  Corpus work raises effort usefully
Why *usefully*?

E.g. concordance vs. dictionary look-up for novel vocabulary

- Dictionary is an easy strategy but with limited involvement or transferability
  - To real life & work

- While assembling/examining/interpreting language data is a highly transferable strategy
  - Can become a habit of mind
    - And produces stronger knowledge with better transfer to novel contexts
      - Cobb (1997)
Enhancement 3:
Data transformation to compensate for gaps in input

- Example: TTS
Enhancement 3: Data transformation to compensate for gaps in input

• Problem: Spoken words map to written words but the mapping is not obvious
  – Words recognized in writing are not recognized in speech, and v-v.

• Solution: TTS technology translates text to speech
  – makes every written word, sentence etc. click-pronounceable
Enhancement 4:
Data transformation to focus on a research-indicated problem skill
Example: Text de/re-construction
Enhancement 4
Data transformation to focus on a skill
Example: Text de/re-construction

• Word knowledge is consolidated through many retrievals

Problem: Opportunities for **word-to-meaning** retrieval are numerous
  – ... while **meaning-to-word** retrieval is rare
    • Productive lexicon = 40% of receptive

• Solution: Transform written/spoken texts into cloze passage
  • Retrieve word/phrase from overall meaning
  • Is effortless for teacher/computer to make
  • Is usefully effortful for learner
Clearly these learnability enhancements can function in consort – Combining data-assembly & meaning-word retrieval
Enhancement 5: Data-Driven Instructional Design

E.g., Assess difficulty of text through frequency profiling

**Problem:** Teachers and materials writers have a poor sense of word frequency

  Or what constitutes an appropriate text for Ss at a certain level

**Solution:** Select or modify texts according to a frequency profile (VP)
A: Well, it was a narrow escape. But we did it. Canadians have preserved their liberties and independence against the always rapacious American beast.

We knew there were powerful elements in the United States that wanted us to kowtow and genuflect to a simplistic worldview, that knuckle-dragging Good-versus-Evil script they have been remorselessly propagandizing all over the world since 9/11.

They have been trying to drag Canada into this simpleton’s game for years, mauling truth and banishing nuance with a continuous stream of invective posing as reason, and caricature passing itself off as accuracy.

It’s a difficult thing to resist the mighty United States at any time, and especially difficult in all the dust and storm of a national election. But we did it.

B: My name is Dr Frederick Treves. I am a doctor at the London Hospital.

One day I saw a picture in the window of a shop near the hospital. I stopped in front of the shop and looked at the picture. At first I felt interested, then I felt angry, then afraid.

It was a horrible, ugly picture. There was a man in the picture, but he did not look like you and me. He did not look like a man. He looked like an elephant. I read the writing under the picture. It said: Come in and see the Elephant Man. 2 pence.

I opened the door and went in. There was a man in the shop. He was a dirty man in an old coat with a cigarette in his mouth. 'What do you want?' he asked. 'I'd like to see the elephant man, please,' I said. The man looked at me angrily. 'Well, you can't,' he said. 'The shop's closing now. You can come back tomorrow.'
Enhancement 6:
Data assembly again, but (a) by learners & (b) with a network

Example

- Problem:
  - Too many new words for learners to look up
    - In reading a challenging text

- Solution:
  Divide the look-ups
  - then re-assemble with a network
<table>
<thead>
<tr>
<th>NEW WORD</th>
<th>EXAMPLE</th>
<th>WORD CLASS</th>
<th>DEFINITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>perennial</td>
<td>The debate on whether voice is teachable in L2 classroom is perennial.</td>
<td>N</td>
<td>The improvement of something.</td>
</tr>
<tr>
<td>insight</td>
<td>Endurance could provide insight to the environment. History of the art but there is risk.</td>
<td>ADJ</td>
<td>(of a problem) able to be solved.</td>
</tr>
<tr>
<td>betterment</td>
<td>I believe politicians are vital to the betterment of society.</td>
<td>N</td>
<td>A principle or set of principles laid down as incontrovertibly true.</td>
</tr>
<tr>
<td>dovetail</td>
<td>Listening and communication must dovetail with other skills.</td>
<td>V</td>
<td>Fit or cause to fit together easily and smoothly.</td>
</tr>
<tr>
<td>cadre</td>
<td>- He trained cadre engineers and health districts in schools in Iraq.</td>
<td>N</td>
<td>The capacity to gain an accurate and understanding of someone or something.</td>
</tr>
<tr>
<td>soluble</td>
<td>Not all problems are soluble, not all can be haggled at acceptable cost.</td>
<td>V</td>
<td>Fit or cause to fit together easily and smoothly.</td>
</tr>
<tr>
<td>hypnotize</td>
<td>I was fascinated them, hypnotize them - until you up and realise they're nuts.</td>
<td>ADJ</td>
<td>Relating to meaning in language or logic.</td>
</tr>
<tr>
<td>dogma</td>
<td>On the contrary, you renew your doctrines and dogmas, make them relevant to the times in which you seek support.</td>
<td>N</td>
<td>A small group of people specially trained for a particular purpose or profession.</td>
</tr>
<tr>
<td>contiguity</td>
<td>Contiguity is necessary in all forms of learning</td>
<td>N</td>
<td>The sequential occurrence or proximal response, causing their association in</td>
</tr>
<tr>
<td>perennial</td>
<td>The debate on whether voice is teachable in L2 classroom is perennial.</td>
<td>N</td>
<td>long-lasting</td>
</tr>
</tbody>
</table>
Enhancement 7: Automated transfer

So we have ~

– Data assembly
– Useful effort raising
– Contextual TTS
– Meaning-to-word retrieval
– Collaborative learning through a network...

† the elusive learning transfer?
Automated transfer

• Problem: Language learning is contextual and learning may not be transferred to a novel context
  – E.g., word knowledge
    • Word learned in one text
      – Is not recognized in another text
    • It is hard to get practice in transfer
      – Word is forgotten when re-encountered
• A DDL approach, again with Group Lex
  – Again ‘bringing it all together’

26
<table>
<thead>
<tr>
<th>#</th>
<th>Qz</th>
<th>WORD</th>
<th>EXAMPLE</th>
<th>POS</th>
<th>DEFINITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✓</td>
<td>perennial</td>
<td>The debate on whether voice is teachable in L2 classroom is perennial.</td>
<td>ADJ</td>
<td>long-lasting</td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
<td>insight</td>
<td>Endurance could provide insight into the environmental history of the area, but there are risks.</td>
<td>N</td>
<td>The capacity to gain an accurate and deep understanding of someone or something.</td>
</tr>
<tr>
<td>3</td>
<td>✓</td>
<td>betterment</td>
<td>I believe politics and politicians are vital to the betterment of society.</td>
<td>N</td>
<td>The improvement of something.</td>
</tr>
<tr>
<td>4</td>
<td>✓</td>
<td>dovetail</td>
<td>Listening and communication skills must dovetail with other skills.</td>
<td>V</td>
<td>Fit or cause to fit together easily and conveniently.</td>
</tr>
<tr>
<td>5</td>
<td>✓</td>
<td>cadre</td>
<td>- He trained cadres of engineers and built health clinics and schools in Iraq.</td>
<td>N</td>
<td>A small group of people specially trained for a particular purpose or profession.</td>
</tr>
<tr>
<td>6</td>
<td>✓</td>
<td>soluble</td>
<td>Not all problems are soluble, not all risks can be hedged at acceptable cost.</td>
<td>ADJ</td>
<td>(of a problem) able to be solved.</td>
</tr>
<tr>
<td>7</td>
<td>✓</td>
<td>hypnotize</td>
<td>I was fascinated by them, hypnotized by them - until you wake up and realise they're nuts.</td>
<td>V</td>
<td>Capture the whole attention of (someone); fascinate.</td>
</tr>
<tr>
<td>8</td>
<td>✓</td>
<td>dogma</td>
<td>On the contrary, you renew your doctrines and dogmas, making them relevant to the times in which you seek support.</td>
<td>N</td>
<td>A principle or set of principles laid down by an authority as incontrovertibly true.</td>
</tr>
<tr>
<td>9</td>
<td>✓</td>
<td>contiguity</td>
<td>Contiguity is necessary in all forms of learning</td>
<td>N</td>
<td>The sequential occurrence or proximity of stimulus and response, causing their association in the mind.</td>
</tr>
<tr>
<td>10</td>
<td>✓</td>
<td>semantic</td>
<td>Its elegant script is unknown from any other source, and not one clue as to its semantic content has emerged.</td>
<td>ADJ</td>
<td>Relating to meaning in language or logic.</td>
</tr>
<tr>
<td>11</td>
<td>✓</td>
<td>traverse</td>
<td>The reader traverses the beam in the opposite direction to that of the object under test to reduce image blur.</td>
<td>V</td>
<td>Move back and forth or sideways.</td>
</tr>
<tr>
<td>12</td>
<td>✓</td>
<td>expound</td>
<td>He used it as an opportunity to expound the idea of flexible framing.</td>
<td>V</td>
<td>Present and explain (a theory or principle) at length and in detail.</td>
</tr>
<tr>
<td>NEW WORD</td>
<td>EXAMPLE</td>
<td>WORD CLASS</td>
<td>DEFINITION</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>----------------------------------------------</td>
<td>------------</td>
<td>-----------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>I believe politics and politicians are vital to the betterment of society</td>
<td>N</td>
<td>The improvement of something</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Not all problems are acceptable cost. (solved at)</td>
<td>ADJ</td>
<td>(of a problem) able to be solved.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>On the contrary, you renew your doctrines and making contiguity relevant to the times you seek other skills.</td>
<td>N</td>
<td>A principle or set of principles laid down as incontrovertibly true.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Endurance could provide into the social history of the area, but there are risks.</td>
<td>N</td>
<td>The capacity to gain an accurate and understanding of someone or something.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Listening and communication skills must dogma dovetail</td>
<td>V</td>
<td>Fit or cause to fit together easily and securely.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>I trained engineers and built schools in Iraq.</td>
<td>N</td>
<td>A small group of people specially trained for a purpose or profession.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Its elegant script is unknown from any other script content has emerged.</td>
<td>ADJ</td>
<td>Relating to meaning in language or literature.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>I was fascinated by them, hypnotized by the insight they're nuts.</td>
<td>V</td>
<td>Capture the whole attention of (someone).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>is necessary in all forms of learning</td>
<td>N</td>
<td>The sequential occurrence or proximately in response, causing their association in memory.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>The debate on whether voice is teachable in</td>
<td>N</td>
<td>Long-lasting.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**NEW WORDS:**
- betterment
- cadre
- contiguity
- dogma
- dovetail
- hypnotize
- insight
- perennial
- semantic
- soluble
<table>
<thead>
<tr>
<th>NEW WORD</th>
<th>EXAMPLE</th>
<th>WORD CLASS</th>
<th>DEFINITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I believe politics and politicians are vital to the <strong>betterment</strong> of society.</td>
<td>N</td>
<td>The improvement of something.</td>
</tr>
<tr>
<td>2</td>
<td>Not all problems are soluble, not all risks can be hedged at acceptable cost.</td>
<td>ADJ</td>
<td>(of a problem) able to be solved.</td>
</tr>
<tr>
<td>3</td>
<td>On the contrary, you renew your doctrines and <strong>dogma</strong>s, making them relevant to the times in which you seek support.</td>
<td>N</td>
<td>A principle or set of principles laid down by an authority as incontrovertibly true.</td>
</tr>
<tr>
<td>4</td>
<td>Endurance could provide <strong>insight</strong> into the environmental history of the area, but there are risks.</td>
<td>N</td>
<td>The capacity to gain an accurate and deep understanding of someone or something.</td>
</tr>
<tr>
<td>5</td>
<td>Listening and communication skills must <strong>dovetail</strong> with other skills.</td>
<td>V</td>
<td>Fit or cause to fit together easily and conveniently.</td>
</tr>
<tr>
<td>6</td>
<td>He trained <strong>cadre</strong>s of engineers and built health clinics and schools in Iraq.</td>
<td>N</td>
<td>A small group of people specially trained for a particular purpose or profession.</td>
</tr>
<tr>
<td>7</td>
<td>Its elegant script is unknown from any other source, and not one clue as to its <strong>semantic</strong> content has emerged.</td>
<td>ADJ</td>
<td>Relating to meaning in language or logic.</td>
</tr>
<tr>
<td>8</td>
<td>I was fascinated by them, <strong>hypnotized</strong> by them - until you wake up and realise they’re nuts.</td>
<td>V</td>
<td>Capture the whole attention of (someone), fascinate.</td>
</tr>
<tr>
<td>9</td>
<td><strong>Contiguity</strong> is necessary in all forms of learning</td>
<td>N</td>
<td>The sequential occurrence or proximity of stimulus and response, causing their association in the mind.</td>
</tr>
<tr>
<td>10</td>
<td>The debate on whether voice is teachable in L10 classroom is <strong>perennial</strong>.</td>
<td>ADJ</td>
<td>long-lasting</td>
</tr>
</tbody>
</table>
QUIZ 2 - Contexts from a CORPUS (bawe_sampler.txt)

Which word/phrase fits all the gaps in each set? (Corpus=bawe_sampler.txt)

Click words for Dictionary Eng_Eng

"DOVETAIL, CADRE, HYPNOTIZE, DOGMA, CONTINUITY," not found in Corpus bawe_sampler.txt

工程质量? Completed: 0 Tries: 0 Per cent: 0 History >> []

1. _______

[001] reactants that contain a hydrophobic portion, which is _______ in oil-like solutions, and a hydrophilic portion, which is _______ in water. This characteristic results from _______. sugars (1) and the precursor of these is _______. by radicals. (The theorem is quoted from _______.)
[002] to correspond with whether the plant stores starch or _______. fn ? 5 then the general polynomial of degree n is not _______. to a better understanding. Section 6 We now define _______. nomials by radical. Definition 6.1 A finite group is _______. isomorphism theorems which will enhance our knowledge about _______. level. We now look at some conditions for groups to be _______.

2. _______

[001] The mental organisation of the _______. mental lexicon is still an open question. The organisation of the _______. lexicon is an important question in cognitive space. According to the first view (henceforth association-based organisation) _______. in the environment that organises the _______. space. For example, 'apple' and 'plum' _______. space than 'apple' and 'basket'. The second view on _______. should be closer neighbours in the _______. not of 'basket'. The second view on _______. provide support for both theories of _______. organisation. Buchanan (1999) showed _______.
Enhancement 8: Negative evidence

Steven Pinker ~

The strongest argument for intuition as a source of language acquisition is the lack of ‘negative evidence’ in natural input... *(paraphrase)*

*How would a learner ever know what is ‘not language’?*
* (= not possible in his/her particular target language)*

A concordance assembly from a large corpus can show what does not exist in a language

*For example, the Arab learner’s “I go to home”*
Zero cases in the intended meaning in BAWE Written (8 million wds in 30 disciplines)
What is Data-Driven Learning?

What questions does it still raise?

- Source of language acquisition
  - Intuition
    - Natural input
    - Expert inspected
  - Input
    - Input as data
    - Learner inspected
Is it better...

• For learners to work with raw data themselves
  – Pushing the software in their own way to raise data comprehensibility & learnability
  • From a few hints and demos

• Or for teachers, course designers, materials writers
  – To do this on learners’ behalf
  – Offering learners themselves just the final step in the process?

Ex: Ss determine the collocations of *ride* and *drive* in a concordance that was basically set up to make this obvious

  With messy or ambiguous data eliminated
A case study: Concordancing on Lextutor with BAWE

- BAWE on Lextutor since 2018
  >>>Well promoted
- One of about 25 open choice corpora & text collections
Research Questions

• Which resources do users exploit most in Lextutor’s concordance program?
• Which resources do they exploit least?
• Which resources do they fail to notice?
• And is this same or different for BAWE/BASE corpora?
What can the BAWE/BASE do? (as realized on Lextutor)

Main affordances

• Size

• Speed
  < 10 secs for basic word + complete output (‘sex’=2119 lines)
  • Cf Mark Davies’ Coca, which gets speed by giving only 100 lines per screen
    – So everything cannot be seen together

• Number and nature of academic sub-divisions

• By far Lextutor’s best corpus

Specific examples ➔
1. Count + examples by discipline

2119 hits
Standardized to 311 per million (hits/corpus size x 1,000,000)

FREQUENCY BY SUB-CORPUS:
sociology=540 law=290 english=189 anthropology=176 classics=172 history=114 linguistics=113 psychology=97 us_sudies=78 politics=69 archeology=40 medicine=37 publishing=37 philosophy=34 mathematics=29 health=28 business=22 biology=15 hospitality=13 economics=9 food_science=8 agriculture=4 cybernetics=2 planning=2 engineering=1 architecture=0 chemistry=0 computer_sci=0 meteorology=0 physics=0

RANGE: family sex appears in 25 out of 30 sub-corpora (range=83.33%)

Click any KEYWORD for more context.

001. ck cross sows in rotary field system 16,000 mixed SEX free range broiler chickens 180 South Devon an [agriculture]
002. led by the breed, the type of nourishment and the SEX. Above is a drawing of a cow and the parts of [agriculture]
003. xers when fully matured. Manyday could use both SEXES for meat production, and also sell the large [agriculture]
005. ions associated with contact. Thus we can see how SEXUAL division, class division, moral ideals and [anthropology]
006. sis for social definitions. What cultures make of SEX differences is almost infinitely variable, so [anthropology]
007. s to be humorous; homely, while a Geisha is to be SEXY, etc. While this image does not appeal to the [anthropology]
008. at this and establish a clear distinction between SEX and gender. During this time it became clear t [anthropology]
009. reviously, the biological differences between the SEXES could not be used as a universal determinant [anthropology]
010. alternative theory that they propose is that both SEX and gender are socially constructed (Moore, 19 [anthropology]
011. e of socially constructed understandings of them. SEXUAL intercourse and human reproduction are not [anthropology]
012. pose that there is a further distinction between SEX and gender: 'Sex', represents a particular con [anthropology]
013. is a further distinction between sex and gender: 'SEX', represents a particular construct of human b [anthropology]
014. trust of human bodies, particularly Euro-America. SEX, refers to the physical nature of human bodies [anthropology]
015. gender, relates to what different cultures see as SEX (Moore, 1994). One fact that must be indicated [anthropology]
016. indicated is that every culture cannot understand SEX without 'sex' (their own definition); this can [anthropology]
017. that every culture cannot understand sex without 'SEX' (their own definition); this can therefore be [anthropology]
018. If even the biological definitions are variable, SEX cannot exist as a universal concept. Gender or [anthropology]
019. concept. Gender or the socially applied meaning of SEX has the prevailing role and consequently gene [anthropology]
020. ale geladas have large shaggy manes of hair. Both SEXES have a naked bright pink hourglass shape on [anthropology]
021. ked bright pink hourglass shape on their chest, a SEXUAL displays such as estrus in females is signa [anthropology]
022.iduals rear feeding have moved the necessity of SEXUAL display to the chest of the individual. The [anthropology]
2. Second-Sort direct from output

- E.g., Frequency of family members

rly childhood are not sexually attracted to each other when SEXUALLY mature, this explains partly why humans rarely mate [anthropology]. Being too physical, too hard. This black masculinity is a 'SEXUALLY' defined masculine ideal rooted in physical domination [sociology]. A male could also imply that the female is either inferior or SEXUALLY promiscuous for example master/mistress, bachelor's [linguistics]. In a way that females may have more experience with being SEXUALLY harassed than males, and as a result, have schemas [psychology]. Is a higher demand for prostitution and thus the spread of SEXUALLY transmitted diseases such as AIDS. Housing provision [planning].

ale publications played upon the popular images of women as SEXUALLY insatiable, nagging scolds, as well as playing upon [history]. The contradictions within it; for example, that women were SEXUALLY uncontrollable, yet the moral guardians of society [history].

respectability, money, and class to declare sisterhood with SEXUALLY disgraced women' in an attempt to relieve tensions [english]. Thus, women develop an instinct for nurturing and men are SEXUALLY promiscuous to provide the best chances for their g [sociology].

to 885,000 by the next period. SUSI competes with SAFE and SEXY (segment market leader) in the Singles market, and SAMA [business].

 whereas Encolpio is a feminine male that opposes the ideal, SEXY gladiator. The spectacle of the male, therefore, is com [classics].

The ideal of the independent woman however these women were SEXY. Connell in Segal 1997: 60 Richardson D in Holland and [anthropology]. To be humorous; homely, while a Geisha is to be SEXY, etc. While this image does not appeal to the average W [anthropology].

% in the first half of 1999. The use of the "funny, useful, SEXY" formula is apparently one which works for the 18-30 age [publishing].

le that have survival value. This is also described as the 'SEXY sons' hypothesis, Eysenck (2001). The female looks for a [psychology].

v, 1996, p79). The females in this anime are very strong, SEXY and don't seem to wear much clothing, this may be to hel [anthropology].

cidation for family sex: sex=890 sexual=733 sexuality=305 sexes=83 sexually=55 sexist=17 sexism=10 sexy=8 sexing=5 sexualised=5 sexed=3 sexless=1 sexualisation=1 sexualise=1 sexualising=1 sexualized=1
3. Click key for larger context

Barbara Ehrenreich and Deirdre English argued that the witchcraft phenomenon ‘was a ruling class campaign of terror directed against the female peasant population...’ However, recent historiography has strayed from this simplistic line of argument and developed many other reasons for the high numbers of women in witchcraft prosecutions. As Brian Levack argues women were more likely to be accused of witchcraft but there was nothing that excluded men. This point is interesting, as although there were many more female than male witches, men were still prosecuted and found guilty, which does not fit into Ehrenreich's and English's thesis. One of the most common arguments in recent years has been that ‘...witchcraft was sex-related but not a sex-specific crime’. Understanding the witchcraft phenomenon as a 'persecution of women'
4. Separable collocation

16 hits
Standardized to 2 per million (hits/corpus size x 1,000,000)

FREQUENCY : sort=15 sorts=1

Click any **KEYWORD** for more context

001. he EAP support class as being there to 'SORT out' all the language problems of t [linguistics]
002. her brother was also angry and told to "SORT yourself out you skinny bitch". Aft [medicine]
003. ments to self (made by student), e.g. "SORT out quote and footnote". 2 2004-10 [english]
004. out, grow up, want, upset, stop, bury, SORT out, fester, perceive, make, feel, [linguistics]
005. chemical shift. The PENDANT carbon NMR SORT the signals out into how many hydro [chemistry]
006. s and thus does suggest an influence of SORTS. As Kuhl points out, eugenicists [history]
007. what constituted reality requires some SORT of outside existence to the body, w [psychology]
008. ater galago karyotypes in an attempt to SORT out ambiguities regarding interspec [anthropology]
009. can generally depend on the business to SORT out the problems quickly and to the [business]
010. e of its whole legal system if it is to SORT out this crises, and even that may [law]
011. ersity of Natal, South Africa, aimed to SORT out the Galagidas current phylogene [anthropology]
012. ites in Central Africa in an attempt to SORT out the taxonomic ambiguities of th [anthropology]
013. ntries across the world, which aimed to SORT out their own problems without Amer [politics]
014. s around the cement mixer. And I try to SORT out what has happened to me. The bu [english]
015. literature review of each species will SORT out what type of research that has [anthropology]
016. isible hands', in which the market will SORT an unethical behaviour out, thus pr [agriculture]
5. Sort by VP level

Most comprehensible contexts first ...

... ask to Schopenhauer, who, "has made it [pity] the VIRTUE" (A7) and for whom no action could better differ. 

... ly from the outside; rather it sublates itself by VIRTUE of its own nature, and passes over, of its own accord, so that nobody knows how to do it. The lower form of VIRTUE, on the other hand, has nothing to do with them.

... whether it is fair to connect these two. In the end it would appear to me that her VIRTUE is certainly possible, and that man would not believe that for a man to be entirely without VIRTUE is another very Augustan VIRTUE, a love of peace, and he cares for his men...
6. With easy extraction for worksheet etc.

Concordance extract for *family* VIRTUE With no associate on EITHER side sorted level

002. With easy extraction for worksheet etc.

003. of all three in so far as they are the VIRTUE of a functionary unity." (Annas, [1.00]

005. writes, "this necessary perspectivism by VIRTUE of which every center of force - [1.00]

007. Philosophy 0215 Does the Meno show that VIRTUE is teachable? The question about [1.00]

008. opinion (doxa), the more often we will be VIRTUOUS. After all, as Guthrie says, 'd [1.00]

010. since Socrates' death in 399BC. Either VIRTUE is teachable or it is not, and it [1.00]

011. solve the problem of the teachability of VIRTUE is, but rather to demonstrate in [1.00]

013. the Brazilian blacks became a national VIRTUE. In Mexico, the corrido and new p [1.00]

014. to have a true knowledge of the form of VIRTUE, he could teach it, but as no one [1.07]

016. it is not taught at present because such VIRTUE as exists at present is not knowl [1.08]

017. whether it is fair to connect these two VIRTUES. In the end it would appear to m [1.08]
7. Disambiguation with collocation

Keyword(s): family ▼ bank  
(Max chars. 30)  In corpus: BAWE Written All (8m)*

OPTION: With associated word(s)  
river far left opposite west

and NONE of these words  
federal committee westminster funds loans capital

CONTROLS:  
NEW SORT ▼ Sort Lines by Mean VP (Demo)

Sorted By ▼ word(s) to ▼ of keyword  | or Keyword ▼  | Line Width 130 ▼  Number of Lines 10,000 ▼
Gapped

On STARTS or FAMILY searches, "sort by keyword" gives distribution of forms before concordance lines (+ again at bottom)

+ Scan for any recurring word (potential collocate) within ▼ words presenting <= ▼ times
(Where would such a list come from?)
Click any KEYWORD for more context.

001. low direction of the stream on the left BANK from which the flow feeds the grazi [agriculture]
002. azing land. While the slope of the left BANK is relatively constantly flat, that [agriculture]
003. ntrolled with sandbags piled up at both BANKS as the river reaches the mill. BAN [agriculture]
004. th banks as the river reaches the mill. BANK TYPE: Vertical or sloping earth cli [agriculture]
005. was very steep. ADJACENT LAND USE: Left BANK: The land was under rough grazing, [agriculture]
006. nd S. cinerea. RB: Similar to the left BANK, with more diversity of flora, incl [agriculture]
007. ant during rearing of juveniles. Beetle BANKS that join the River Thames to the [agriculture]
008. tlement. Woodland areas, river and lake BANKS, and wetland areas such as the Som [archeology]
009. of the tombs. They are also on the same BANK of the river, spaced a fair distanc [archeology]
010. ated on the east side of Oxford, on the BANK of the River Cherwell. The college [architecture]
011. n belt and within 50m meters of a river BANK, 500m of Ancient Woodland, 25m of a [architecture]
012. the shimmering crane bird on the river BANK White like fresh milk! His roar is [english]
013. the shimmering crane bird on the river BANK', here the line breaks without inst [english]
014. Vosges#. In the Rhine valley, the west BANK faces the east and so is warmed by [food_science]
015. is warmed by the morning sun. The west BANKS receive the suns rays for the maxi [food_science]
016. wars such as in the Gaza Strip and West BANK where people were forcibly removed [planning]
017. e, Robert (2001) 'Showdown at the World BANK', New Left Review 7, Accessed on 08 [politics]
018. ion over Israeli occupation of the West BANK, there are more significant problem [politics]
019. at Kind of Women: Stories from the Left BANK and Beyond . London: Virago Russell [sociology]
363 hits

Standardized to 53 per million (hits/corpus size x 1,000,000)

FREQUENCY BY SUB-CORPUS: politics=139 economics=68 agriculture=49 law=40 hospitality=34 business=7 publishing=4 history=3 us_studies=3 sociology=2 computer_sci=1 cybernetics=1 planning=1 anthropology=0 archeology=0 architecture=0 biology=0 chemistry=0 classics=0 food science=0 health=0 mathematics=0 medicine=0 meteorology=0 philosophy=0 physics=0

RANGE: family bank appears in 15 out of 30 sub-corpora (range=50.00%)

Click any KEYWORD for more context

001. many internal donors, such as the World BANK, which has just released a credit o [agriculture]
002. cept to Agricultural Innovation , World BANK Project Presentation in India Harlo [agriculture]
003. h Working Paper Series, No. 3332, World BANK, [WWW], Accessed in 25th November [agriculture]
005. g-run (Stern, 1999). Accordingly, World BANK advocated the discouragement of the [agriculture]
008. untry more or less protectionist? World BANK Policy Research Working Paper no. W [agriculture]
011. 002). A report carried out by the World BANK (2001) showed these afore mentioned [agriculture]
012. everity of the problem comes from World BANK (2001) where 23.4 billion tonnes of [agriculture]
013. vers and from cultivable land, (World BANK, 2004). This is not only a problem [agriculture]
014. Loess Plateau with help from the World BANK aimed to improve the potential for [agriculture]
015. tallied sediment control dams. The World BANK played an important role in the pro [agriculture]
016. nd the initial investment lasts, (World BANK, 2004). As was mentioned earlier, p [agriculture]
017. UN is involved together with the World BANK to help China in this period of tra [agriculture]
018. cal scale, with the help from the World BANK and FAO, enabling farmers to take c [agriculture]
020. ment Programme was financed from World BANK funds. PHARE programmes were the fi [agriculture]
021. rs for educational projects, the World BANK, has been currently funding 157 pro [agriculture]
022. nd Koynan (2000) produced for the World BANK shows that paid extension is "fossi [agriculture]
023. Technology Transfer in Nicaragua, World BANK, [WWW], Accessed in October 10th 2 [agriculture]
7. Challenging gap worksheets

34 hits  Hits/million cannot be calculated if Number of Lines user has chosen is set below the number a

FREQUENCY BY SUB-CORPUS: medicine=20  english=6  economics=1  engineering=1  health=1  his politics=1  sociology=1  us_studies=1  agriculture=0  anthropology=0  archeology=0  archite business=0  chemistry=0  classics=0  computer_sci=0  cybernetics=0  food_science=0  hospital linguistics=0  mathematics=0  meteorology=0  physics=0  planning=0  psychology=0  publishin

Click any KEYWORD for more context

001. e and the threat of hostile takeover is _______. (b) large groups of firms are i [economics]
002. of 3 mm to a final diameter of 1 mm are _______. greater than, for instance, dra [engineering]
003. otif in the final stanza ensures that a _______. sense of loss and bereavement p [english]
004. tation." As with 'The Tell Tale Heart', _______. instability and distorted reali [english]
005. nt of view; so shallow, so delusive, so _______. in every other." Images of d [english]
006. ustration thus making dissatisfaction a _______. force; Marlow's tangible disill [english]
007. e narrator's presence, intensifying the _______. melancholic despondency that p [english]
008. eeing air was stunned' (l.9) engenders a _______. aura of paranoia. Depicting a b [english]
009. f the questions to be asked there was a _______. imbalance of power, with the in [health]
010. ty, the sublime landscape, unity with a _______. nature, and Pantheism, all of w [history]
011. ed pain. Synovium. The synovium was not _______. Muscle bulk. The leg was swoll [medicine]
012. r. Distension/hepatomegaly/spleomegaly/ _______. faeces in colon. BS +. Respirat [medicine]
013. ves on palpation. The apex beat was not _______. No pain was elicited on palpat [medicine]
014. No left or right parasternal heaves or _______. thrills. The apex beat was loca [medicine]
015. patient was supine Palpation: Liver was _______. five fingers breadth below the [medicine]
016. th below the subcostal marginSpleen was _______. to the umbilicusNo other masses [medicine]
017. ilicNo other masses felt. Kidneys not _______. Percussion: Liver was percusse [medicine]
018. scites, there was no loin tenderness or _______. kidneys. Again, this diagnosis [medicine]
019. and lymphoedema. There is no history of _______. cervical lymph nodes or yellowi [medicine]
020. qually on both sides. The apex beat was _______. in the 5 th ICS MCL. No pain wa [medicine]
021. ess are also unlikely, as there were no _______. masses or pain elicited on abdo [medicine]
022. nderness in the right iliac fossa and a _______. lump may indicate the presence [medicine]
RESULTS
Summary of feature use – all corpora

From 49,461 independent consultations since April 2018, here are the proportion seeking...

– A particular corpus
– Search options (equals, family, starts, ends)
– Sort options (VP-Level, discipline, right, left)
  • Esp. Partitioned corpora
– Collocation
  • Fixed and separable
  • Esp. Disambiguation by multiple
– Larger context
– Output shaping
– Gapped output
Data restriction

The following data does **not** include these:

- Concordance DEMO routines from input page
  - \( \leq 20 \) in number and probably \( >50\% \) of searches

- Automated concordance calls from other routines
  - With pre-set defaults
<table>
<thead>
<tr>
<th>CORPUS</th>
<th>SIZE (wds)</th>
<th>hits</th>
<th>% of 542,777</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>1 m</td>
<td>189,272</td>
<td>34.87</td>
</tr>
<tr>
<td>BNC_legal</td>
<td>2.3 m</td>
<td>58,723</td>
<td>10.82</td>
</tr>
<tr>
<td>Brown_BNCW*</td>
<td>2 m</td>
<td>41,339</td>
<td>7.62</td>
</tr>
<tr>
<td>BNC_sp_v_wrt*</td>
<td>2 m</td>
<td>41,339</td>
<td>7.62</td>
</tr>
<tr>
<td>Brown_BNCW</td>
<td>2 m</td>
<td>22,904</td>
<td>4.22</td>
</tr>
<tr>
<td>Acada Corp</td>
<td>6 m</td>
<td>21,161</td>
<td>3.90</td>
</tr>
<tr>
<td>BAWE*</td>
<td>8 m</td>
<td>19,749</td>
<td>3.64</td>
</tr>
<tr>
<td>bncoca_1-2</td>
<td>14 m</td>
<td>19,611</td>
<td>3.61</td>
</tr>
<tr>
<td>BNC Speech Sampler</td>
<td>1 m</td>
<td>19,162</td>
<td>3.53</td>
</tr>
<tr>
<td>BNC Medicine</td>
<td>1.4 m</td>
<td>17,191</td>
<td>3.17</td>
</tr>
<tr>
<td>BNC Spoken</td>
<td>10 m</td>
<td>15,500</td>
<td>2.86</td>
</tr>
<tr>
<td>BNC Written Sampler</td>
<td>1 m</td>
<td>15,171</td>
<td>2.80</td>
</tr>
<tr>
<td>BNC_humanities</td>
<td>3.4 m</td>
<td>12,790</td>
<td>2.36</td>
</tr>
<tr>
<td>TV Talk</td>
<td>.5 m</td>
<td>11,920</td>
<td>2.20</td>
</tr>
<tr>
<td>Wiki*</td>
<td>1.1 m</td>
<td>10,639</td>
<td>1.96</td>
</tr>
<tr>
<td>BNC Commerce</td>
<td>2.2 m</td>
<td>8,740</td>
<td>1.61</td>
</tr>
<tr>
<td>Graded 2k</td>
<td>1 m</td>
<td>8,406</td>
<td>1.55</td>
</tr>
<tr>
<td>Graded 1k</td>
<td>1 m</td>
<td>7,470</td>
<td>1.38</td>
</tr>
<tr>
<td>Dr House*</td>
<td>.8 m</td>
<td>3,801</td>
<td>0.70</td>
</tr>
<tr>
<td>Graded*</td>
<td>1 m</td>
<td>3,511</td>
<td>0.65</td>
</tr>
</tbody>
</table>

* = partitioned corpus, otherwise 1-file corpus
<table>
<thead>
<tr>
<th>SEARCH TYPE</th>
<th># HITS</th>
<th>% of 542,777</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQUALS</td>
<td>478,611</td>
<td>88.18</td>
</tr>
<tr>
<td>FAMILY</td>
<td>48,891</td>
<td>9.01</td>
</tr>
<tr>
<td>STARTS</td>
<td>7,374</td>
<td>1.36</td>
</tr>
<tr>
<td>CONTAINS</td>
<td>5,927</td>
<td>1.09</td>
</tr>
<tr>
<td>ENDS</td>
<td>1,841</td>
<td>0.34</td>
</tr>
</tbody>
</table>
### MULTIWORD SEARCH

<table>
<thead>
<tr>
<th># WDS</th>
<th># HITS</th>
<th>% of 542,777</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Word</td>
<td>373,062</td>
<td>68.7</td>
</tr>
<tr>
<td>2 WDS</td>
<td>57,879</td>
<td>10.66</td>
</tr>
<tr>
<td>3 WDS</td>
<td>67,751</td>
<td>12.48</td>
</tr>
<tr>
<td>4 WDS</td>
<td>6,082</td>
<td>1.12</td>
</tr>
<tr>
<td>5 WDS</td>
<td>2,318</td>
<td>0.43</td>
</tr>
<tr>
<td>SORT TYPE</td>
<td># HITS</td>
<td>% of 542,777</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
<td>--------------</td>
</tr>
<tr>
<td>KEY (default)</td>
<td>257,153</td>
<td>47.4</td>
</tr>
<tr>
<td>LEFT</td>
<td>207,708</td>
<td>38.3</td>
</tr>
<tr>
<td>RIGHT</td>
<td>54,240</td>
<td>9.9</td>
</tr>
<tr>
<td>VP-LEVEL</td>
<td>3,864</td>
<td>0.7</td>
</tr>
<tr>
<td>SUB-CORPUS</td>
<td>1,248</td>
<td>0.2</td>
</tr>
<tr>
<td>REFIRE</td>
<td>% of 542,777</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>--------------</td>
<td></td>
</tr>
<tr>
<td>266,796</td>
<td>49.15</td>
<td></td>
</tr>
<tr>
<td># OF ASSOC_WORD</td>
<td># HITS</td>
<td>% of 542,777</td>
</tr>
<tr>
<td>-----------------</td>
<td>---------</td>
<td>--------------</td>
</tr>
<tr>
<td>0 WORDS</td>
<td>465813</td>
<td>85.8</td>
</tr>
<tr>
<td>1 WORD</td>
<td>75314</td>
<td>13.9</td>
</tr>
<tr>
<td>2 WDS</td>
<td>4304</td>
<td>0.79</td>
</tr>
<tr>
<td>3 WDS</td>
<td>1883</td>
<td>0.35</td>
</tr>
<tr>
<td>4 WDS</td>
<td>749</td>
<td>0.14</td>
</tr>
<tr>
<td>5 WDS</td>
<td>460</td>
<td>0.08</td>
</tr>
</tbody>
</table>

ASSOC_WORD \w*.\w \w*.\w \w*.\w \w*.\w \w*.\w
Requests for larger contexts on two random days

<table>
<thead>
<tr>
<th>Popular Pages Complete Lexical Tutor - Thu 23rd May, 2019</th>
<th>Popular Pages Complete Lexical Tutor - Tue 28th May, 2019</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hits</strong></td>
<td><strong>Title</strong></td>
</tr>
<tr>
<td>974</td>
<td>VP 1-WORD (HOME)</td>
</tr>
<tr>
<td>907</td>
<td>Complete Lexical Tutor</td>
</tr>
<tr>
<td>594</td>
<td>CORPUS CONCORDANCE OUTPUT</td>
</tr>
<tr>
<td>588</td>
<td>VP 1-WORD (EDIT-TO-PROFILE)</td>
</tr>
<tr>
<td>563</td>
<td>VOCABPROFILE COMPLETE - INPUT</td>
</tr>
<tr>
<td>552</td>
<td>VOCABPROFILE COMPLETE - OUTPUT</td>
</tr>
<tr>
<td>494</td>
<td>VOCABPROFILE ENGLISH</td>
</tr>
<tr>
<td>446</td>
<td>VOCABPROFILE ENGLISH OUTPUT</td>
</tr>
<tr>
<td>253</td>
<td>CORPUS CONC ENG</td>
</tr>
<tr>
<td>234</td>
<td>VP HOME</td>
</tr>
<tr>
<td>219</td>
<td>CORPUS CONCORDANCE OUTPUT</td>
</tr>
<tr>
<td>187</td>
<td>VST (computer/test)</td>
</tr>
<tr>
<td>178</td>
<td>CORPUS CONC ENG</td>
</tr>
<tr>
<td>170</td>
<td>TESTS HOME</td>
</tr>
<tr>
<td>153</td>
<td>ONLINE CONCORDER</td>
</tr>
<tr>
<td>138</td>
<td>VST (computer/practice)</td>
</tr>
<tr>
<td>121</td>
<td>Complete Lexical Tutor</td>
</tr>
<tr>
<td>116</td>
<td>VOCABPROFILE COMPLETE - INPUT</td>
</tr>
<tr>
<td>115</td>
<td>VOCABPROFILE COMPLETE - OUTPUT</td>
</tr>
<tr>
<td>98</td>
<td>FREQUENCY HOME</td>
</tr>
<tr>
<td>81</td>
<td>LEVELS TEST (PRODUCTIVE)</td>
</tr>
<tr>
<td>78</td>
<td>LARGER CONTEXT</td>
</tr>
<tr>
<td>73</td>
<td>HYPERTEXT 2</td>
</tr>
<tr>
<td>69</td>
<td>TEXT-LEX COMPARE</td>
</tr>
<tr>
<td>64</td>
<td>FREQUENCY INDEXER ENGLISH</td>
</tr>
<tr>
<td>41</td>
<td>LARGER CONTEXT</td>
</tr>
<tr>
<td>40</td>
<td>404 Not Found</td>
</tr>
</tbody>
</table>

119/813 = 15%

82/1049 = 8%
Interesting gap-tasks?

- Negligible across corpora - despite their popularity with learners in my experience
  and proven learning power particularly in transfer dimension
Frequency of look-up words by VP

By k1=0-999 etc. by combined family count

Roughly

- 50% 1k-3k
- 25% 4k-25k
- 25% “off-list”

(mainly misspellings not found in VP’s BNC-Coca family lists of 25k x 5 = 125,000 word-forms)
And specifically BAWE Complete

19,749 searches

<table>
<thead>
<tr>
<th># SEARCH WORDS</th>
<th>0</th>
<th>%</th>
<th>1</th>
<th>%</th>
<th>2</th>
<th>%</th>
<th>3</th>
<th>%</th>
<th>4</th>
<th>%</th>
<th>5</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>12,763</td>
<td>64.6</td>
<td>3,307</td>
<td>16.7</td>
<td>1340</td>
<td>6.8</td>
<td>427</td>
<td>2.2</td>
<td>65</td>
<td>0.3</td>
</tr>
<tr>
<td># ASSOC WORDS</td>
<td>16,453</td>
<td>83.3</td>
<td>2,834</td>
<td>14.4</td>
<td>9</td>
<td>0.0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SEARCH TYPE</th>
<th>equals</th>
<th>%</th>
<th>family</th>
<th>%</th>
<th>starts</th>
<th>%</th>
<th>contains</th>
<th>%</th>
<th>ends</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>18,065</td>
<td>91.5</td>
<td>1,037</td>
<td>5.3</td>
<td>245</td>
<td>1.2</td>
<td>320</td>
<td>1.6</td>
<td>78</td>
<td>0.39</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SORT TYPE</th>
<th>key</th>
<th>%</th>
<th>left</th>
<th>%</th>
<th>right</th>
<th>%</th>
<th>VP-Level</th>
<th>%</th>
<th>subcorp</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>13,527</td>
<td>68.5</td>
<td>2,998</td>
<td>15.2</td>
<td>2,596</td>
<td>13.1</td>
<td>159</td>
<td>0.8</td>
<td>465</td>
<td>2.35</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SEARCH WORD</th>
<th>k1</th>
<th>%</th>
<th>k2</th>
<th>%</th>
<th>k3</th>
<th>%</th>
<th>k4</th>
<th>%</th>
<th>k5-25</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15,404</td>
<td>53.0</td>
<td>4,900</td>
<td>16.9</td>
<td>4,417</td>
<td>15.2</td>
<td>1,049</td>
<td>3.6</td>
<td>2,181</td>
<td>7.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>K-LEVEL</th>
<th>OFF</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1,085</td>
<td>3.74</td>
</tr>
</tbody>
</table>
Look-ups by frequency for BAWE Complete

For all corpora
What does frequency level of look-up items tell us?

• Hypothesis:
  – If look-ups are mainly 1k-3k, then the interest is usage characteristics of common words
    • Collocation, part of speech, register, how-to-use
  – If mainly 4k+, then meaning of uncommon words is the interest
    • Polysemy, homophony, inferencing from context, meaning within a discipline

– BAWE pattern suggests targeting of usage characteristics
What does fewer off-list items tell us?

• Hypothesis:
  – Off-list items are mainly misspellings
  – High rate of misspellings suggests learners are working independently
  – Low rate suggests teacher has supplied a BAWE task

• Which in turn supports the usage hypothesis
What distinguishes BAWE usage?

• Fewer off-list look-ups
• More common word look-ups
• Hence focus on usage in teacher-led tasks

Everything else is the same as other corpora
  – Low interest in corpus partitioning
    • search, sort, collocation options
    • broader context
    • collocation
Putting it all together

• Most users get one-word example from small, unpartitioned corpus
  – Minus interesting sort or collocation

• Interesting corpora + searches get low use

• Just enough to indicate…
  – They *are* discoverable
  – But that few discover them

• High quality corpus like BAWE makes minor difference
But who is this user?

- If **weekend** is heavy, it is mainly students doing assignments
- If **weekday** is heavy, it is mainly teachers preparing and delivering in-class demos
  – Confirmed by emails
Hypothesis: Users are > 50% teachers

• The remainder are probably learners told to look at concordance by their teachers
  — But learners seem not adequately prepared to exploit resources most usefully
How many TESL training programs include *anything* about corpus?

- How to search, interpret, generalize patterns, apply patterns...
  - Let alone get learner into DDL spin-offs as offered by Lextutor

- **Mini-Experiment**: Google the keywords “tesl training programs corpus DDL”
  - With the small problem that ‘DDL’ picks up ‘didactique des langues’
tesl training programs corpus

About 56'700 results (0.50 seconds)

**TEFL/TESL Certificate Programs - Distance Learning Programs**
https://www.tesol.org/...your.../certificate-programs-and-distance-learning-programs

Those who decide not to begin their teaching career with a TESL/TEFL/TESOL degree or endorsement may choose to pursue a shorter term certificate program, ...

*Missing: corpus | Must include: corpus*

**People also ask**

- How can I get my TEFL certification online for free?
- How do I get ESL certified?
- What qualifications do you need to teach English as a second language?
- What is a TESL certificate?

**Accredited TESL Training Programs | TESL Ontario**
www.teslontario.org/accreditation/tesl-training-programs

The TESL Ontario accredited English teachers work for a variety of employers ... For information about the TESL program admission requirements, registration ...

*Missing: corpus | Must include: corpus*
Conclusion

• Better corpus and search tools are not enough
  – Users will only scratch the surface
    • Whether learners or teachers

• Training in search & interpretation is essential
  – TESL MA programs are the logical place to start
    • But these have shrinking course loads, not expanding!

• The “corpus revolution” is reaching the classroom
  – But it is a minimal version
Conclusion

• So to answer the starting question

What questions does it still raise?

- Source of language acquisition
  - Intuition
  - Natural input
  - Expert inspected

- Input
  - Input as data
  - Learner inspected
Conclusion

• DDL only makes sense if learners’ feel they are making their own discoveries in the language data
  – In some sort of problem-solving framework

• But this takes some knowledge, preparation & planning to set up
  – I suspect we need the hidden hand of the expert a bit closer to the surface of DDL than it is at present
    • Make sure interesting data gets ‘discovered’
Conclusion

• This hidden hand is needed to make sure that high quality corpora and search tools get used and are stretched
  —And people like us are pushed to make them even better

• But the Lextutor experiment shows this is unlikely to happen by itself

~END~