ACCESSING ALL AREAS: Corpus Analysis methods in inter-disciplinary applications

Ramesh Krishnamurthy
Aston University
March 2011
Abstract

• Every human activity and every academic discipline generates texts.

• Applying Corpus Analysis methods to these texts can often generate new insights into the underlying themes and threads in various fields, and discover aspects that are often missed by qualitative approaches.

• This talk will describe and discuss a wide range of applications in, for example, pedagogy, translation, literature, forensics, sociology, politics, computer science, and the natural sciences.
“Every human activity and every academic discipline generates texts”

**e.g. BUSINESS**

<table>
<thead>
<tr>
<th>“REAL WORLD”</th>
<th>ACADEMIC PURPOSES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social</td>
<td>Travel</td>
</tr>
<tr>
<td>Committees</td>
<td>Presentations</td>
</tr>
<tr>
<td>Dealing with customers</td>
<td>Buying and selling</td>
</tr>
<tr>
<td></td>
<td>Seminars and lectures</td>
</tr>
<tr>
<td></td>
<td>Meetings</td>
</tr>
<tr>
<td>Telephone</td>
<td></td>
</tr>
<tr>
<td>Telex</td>
<td></td>
</tr>
<tr>
<td>Letters</td>
<td>Order, Supply, Remind,</td>
</tr>
<tr>
<td></td>
<td>Complain, Apologize ...</td>
</tr>
<tr>
<td>Memos</td>
<td>Report writing</td>
</tr>
<tr>
<td>Email</td>
<td>Case studies</td>
</tr>
<tr>
<td></td>
<td>Summary writing</td>
</tr>
<tr>
<td>Documentation</td>
<td>Import/Export</td>
</tr>
<tr>
<td>Promotional literature</td>
<td>Text books</td>
</tr>
<tr>
<td>Job specifications and advertisements</td>
<td></td>
</tr>
<tr>
<td>WEBSITES</td>
<td>Financial press</td>
</tr>
<tr>
<td></td>
<td>WEBSITES</td>
</tr>
</tbody>
</table>
Applying Corpus Analysis methods to these texts...

... but WHAT IS A CORPUS?

- “a collection of naturally occurring language texts, chosen to characterize a state or variety of a language” (Sinclair, 1991)

- “a collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a language” (Crystal, 1991)

- “a collection of actually occurring texts (either spoken or written), stored and accessed by means of computers, and useful for investigating language use” (Thornbury, 2006)

- Like a library, but in which you know not only where each text is, but where every word in each text is! (Krishnamurthy, 20..?)
Why bother with corpora?

• For hundreds of years, **astronomers and sailors** observed the stars in the sky with their eyes. Astronomers were able to construct satisfactory theories, and sailors were able to navigate over large distances successfully. However, once **telescopes** were invented, astronomers and sailors realized that **many of their theories and assumptions were wrong**, e.g. that the Earth moved round the Sun and not vice versa, and the Earth was round and not flat.

• Similarly, **corpus** research has shown that although we may speak, read and communicate very well, **our intuitions about language are in fact unreliable**. What we **think** we have heard or read is often very different from what we have actually heard or read. (Indeed what we think we ourselves have said or written is often different from what we have actually said or written!)
Why bother with corpora?

- “Language users cannot accurately report language usage, even their own” (Sinclair, 1987)
- “Using a language is a skill that most people are not conscious of; they cannot examine it in detail, but simply use it to communicate” (Sinclair 1995)
- “There are many facts about language that cannot be discovered by just thinking about it, or even reading and listening very intently” (Sinclair, 1995)
- As language teachers and professionals, we often have strong intuitions about language use... **Corpus-based research, however, shows us that our intuitions are often completely wrong.** (Biber 2005)
<table>
<thead>
<tr>
<th>Why not use the Web and Search Engines?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Web and Search Engines</strong></td>
</tr>
<tr>
<td><strong>SIZE</strong></td>
</tr>
<tr>
<td><strong>PROCESSING SPEED</strong></td>
</tr>
<tr>
<td><strong>ANALYSES</strong></td>
</tr>
<tr>
<td><strong>CONTENT RANGE</strong></td>
</tr>
<tr>
<td><strong>CONTENT STABILITY</strong></td>
</tr>
<tr>
<td><strong>CONTENT QUALITY</strong></td>
</tr>
<tr>
<td><strong>SOFTWARE</strong></td>
</tr>
</tbody>
</table>
Corpus Analysis methods: bottom-up!

1. Identification *(tokenization)*
   - to identify the objects of study (words, phrases, etc)

2. Frequency
   - to know which objects are common and which are rare

3. Behaviour
   - eg *concordance*
   - to observe the behaviour of the objects in context

4. Analysis
   - eg *collocation*
   - to use the power of computer software to identify patterns

5. Interpretation and Generalisation
   - on the basis of the observed behaviour and analyses
Corpus Analysis 1. IDENTIFICATION: *(tokenisation)* of objects of study – *problems of definition*...
Corpus Analysis 2. FREQUENCY

In Descending Frequency order:

<table>
<thead>
<tr>
<th>Most frequent words</th>
<th>BCET (18 m words)</th>
<th>Bank of English (418 m words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>1,081,654</td>
<td>22,849,031</td>
</tr>
<tr>
<td>of</td>
<td>535,391</td>
<td>10,551,630</td>
</tr>
<tr>
<td>and</td>
<td>511,333</td>
<td>9,787,093</td>
</tr>
<tr>
<td>to</td>
<td>479,191</td>
<td>10,429,009</td>
</tr>
<tr>
<td>a</td>
<td>419,798</td>
<td>9,279,905</td>
</tr>
<tr>
<td>in</td>
<td>334,183</td>
<td>7,518,069</td>
</tr>
<tr>
<td>that</td>
<td>215,332</td>
<td>4,175,495</td>
</tr>
<tr>
<td>s</td>
<td>(8,570)</td>
<td>4,072,762</td>
</tr>
<tr>
<td>is</td>
<td>(166,691)</td>
<td>3,900,784</td>
</tr>
<tr>
<td>it</td>
<td>198,578</td>
<td>3,771,509</td>
</tr>
</tbody>
</table>

In Alphabetical order:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>receivability</td>
<td>5</td>
</tr>
<tr>
<td>receivable</td>
<td>147</td>
</tr>
<tr>
<td>receivables</td>
<td>182</td>
</tr>
<tr>
<td>receival</td>
<td>2</td>
</tr>
<tr>
<td>receivals</td>
<td>3</td>
</tr>
<tr>
<td>receive</td>
<td>29942</td>
</tr>
<tr>
<td>received</td>
<td>48309</td>
</tr>
<tr>
<td>receiver</td>
<td>3129</td>
</tr>
<tr>
<td>receivers</td>
<td>1311</td>
</tr>
<tr>
<td>receivership</td>
<td>737</td>
</tr>
<tr>
<td>receiverships</td>
<td>28</td>
</tr>
<tr>
<td>receives</td>
<td>4705</td>
</tr>
<tr>
<td>receiveth</td>
<td>2</td>
</tr>
<tr>
<td>receiving</td>
<td>11340</td>
</tr>
</tbody>
</table>

APPLICATIONS: linguistics: language description; language teaching: lexical syllabus; lexicography: inclusion/ exclusion, spelling forms, frequency markers
however, have trimmed its interest receivable figure by £22 million. I couldn't tell whether accounts receivable had been paid off set by an increase in accounts receivable. Later, when would routinely tell her they had receivables of thousands of credit-card receivables it has so far.

The value of debtors (or receivables) has to be paid foreign-currency payables against receivables. This type

of a glance at Andrei. Replacing the receiver, she spoke hurr...
Corpus Analysis 4.1 software – patterns: e.g. collocates of *hard* (BoE 448m, 2001)

1. By frequency: = co-occurrence

|  |  
|---|---|
| to | 58444 |
| the | 39225 |
| it | 37451 |
| and | 27875 |
| a | 24531 |
| is | 18314 |
| s | 16947 |
| of | 14980 |
| for | 11462 |
| work | 11378 |

2. By various statistics: [e.g. t-score] = ‘significant’ collocation

|  |  
|---|---|
| it | 37451 | 146.778704 |
| to | 58444 | 136.672099 |
| work | 11378 | 99.794844 |
| very | 7297 | 73.848184 |
| is | 18314 | 64.938617 |
| find | 4915 | 64.513063 |
| working | 4347 | 61.992590 |
| worked | 4039 | 61.399611 |
| s | 16947 | 55.243460 |
| so | 5902 | 49.255196 |

Many ‘corpus-frequent’ words

Calculates ‘actual vs expected’ frequency; thus promotes ‘more than random’ co-occurrences; and eliminates many ‘corpus-frequent’ words
Corpus Analysis 4.2 software – patterns: e.g. collocates of **hard** (BoE 448m, 2001)

[**NODE** = word being analysed = **hard**]

**LEXICO-GRAMMAR**: (impersonal) *it, hard+to-infinitive, hard+for/on (sb?)*

**PHRASEOLOGY**: *hard + time(s), line, core, pressed, currency, hitting*

**IDIOMATICITY**: *hard (on the?) heels (of?)*

**PRAGMATICS**: advice, excuses, warnings

**MODALITY**: *will/would, degree (very, so, too, how)*

**SEMANTICS**: *work/time/currency; lexical sets (perception: believe /imagine /see /understand; practical: get, find, say, beat, tell, keep, come, make)*
LEXIS: (E)FL Dictionaries (Cobuild, Longman, Cambridge, Macmillan); General Dictionaries (Collins, NODE, ODE); Bilingual Dictionaries (Collins, Oxford-Hachette); many new National Dictionaries (Czech, Hungarian, Slovene, South African multilingual); Thesaurus (A. Kilgarriff); (E)LT Coursebooks


FORMULAIC: N. Schmitt (2004); M. McCarthy (1997)


SPOKEN v WRITTEN: D. Biber (2006); M. McCarthy (2001); S. Adolphs (2010)

GRAMMAR: S. Hunston/G. Francis (Pattern Grammar, 2000); D. Biber (Spoken and Written, 1999)

VARIETIES/DIALECTS: ICE; Leeds (Atwell: Br/Am, 2007); Scots; Irish

HISTORICAL/DIACHRONIC: Helsinki

CONTRASTIVE: Larrivee & Krishnamurthy (rare syntactic sequences: def det+indef pron; Eng/Fr, 2009); GeWiss
APPLICATION 1: LANGUAGE STUDIES

- Most frequent lexical words: *time, people* (Cobuild, 1987)
- common verbs... are used ‘delexically’: *take a bath/an exam, make a decision/mistake, have a rest/an idea* (Cobuild, 1987)
- *see...* means ‘understand’ (esp. in spoken): *I see / You see*
- *thing / this / that* refer most often to abstract entities (e.g. a proposition or argument), not physical objects: *A strange thing happened... Is that why you had a few days off?... This is why I'm opposed to the plan* (Willis, 1990)
- -ly adverbs: *lamely* used for excuses (not walking), *crisply* used with verbs of speech (Cobuild; cf ~ in other dictionaries)
- *of ...*is not a preposition: rarely used in adjuncts, usually in noun phrases, *X of Y* (Sinclair, 1991)
- New words (1985-1995): (0 in 1985 corpus) > 1995 corpus = camcorder (1214), virtual reality (458), imaging (463), mobile phone (455), satellite dish (236), laptop (184), videophone (144), palmcorder, smart card, microsurgery, teleworker, email, helipad, hypertext (Cobuild 1995)
APPLICATION 1: LANGUAGE STUDIES

  – http://ijl.oxfordjournals.org/content/2/1/44.short

• = corpus & pre-corpus dictionaries

• Headwords/senses: “...marks COBUILD out as closer to the ideal of a learner's dictionary than LDOCE, whose wealth of extra information seems rather peripheral and thus superfluous”

• Examples: “Any learner's dictionary stands or falls by the number and quality of its examples, for it is the examples above all which help the foreign learner to understand a word (decoding function), learn it (learning function) and use it (encoding function). LDOCE... is clearly inferior to COBUILD both as far as the number and length of its examples are concerned”

• http://www.antimoon.com/how/cobuild-review.htm

• “friendliest, easiest-to-understand definitions on the market... show you how to use a word naturally, not just what it means... the definition and the example sentences give you an almost complete picture of how a word is used in the English language. After you read them, the word is usually "yours" — you can use it easily in your own sentences”
APPLICATION 1: LANGUAGE STUDIES

“in speech, progressive verbs are more common than simple verbs” (Biber 2005)

Figure 5: Frequency of simple, perfect, and progressive aspect in four registers (based on Biber et al, 1999, Figure 6.2)
would is itself hypothetical (= 50% of corpus examples): I think The Tempest would make a wonderful film. Conditional use is a sub-category: It would surprise me if sterling strengthened. Cf coursebooks: would is only 2nd conditional (Willis, 1990)

Sequences of tense, person, and phrases of time and place always depend on the standpoint of the writer/speaker. They are not rules of reported speech (Willis, 1990)

Lexical Syllabus – Willis (1990), Lewis (1993)

Collocation - Lewis (2000), Barfield & Gyllstad (2009)

Authentic texts and examples - Ranalli (2003), Gavioli & Aston (2001), Koprowski (2005)

Learner corpora - ICLE (Granger, 1993), L1-specific vs generic L2 problems, overuse and underuse (Berber-Sardinha, 1996, 2004)
APPLICATION 2: LANGUAGE PEDAGOGY

- **CALL** (Computer-Assisted Language Learning) and **CorpusCALL**
- **Data-Driven Learning** (Tim Johns)
- **Lexical Syllabus** (Willis 1990)
- **TALC** (Teaching and Language Corpora) conferences (1994 – 2006)
  - [http://talc7.eila.jussieu.fr/previous_sites.en.shtml](http://talc7.eila.jussieu.fr/previous_sites.en.shtml)
- **CLLT** (Corpus Linguistics and Language Teaching) newsgroup
Teachers analyse students’ errors

A corpus of 50 essays written by Chinese EAP (English for Academic Purposes) foundation students

5232 errors identified. Most frequent errors were:

1. Missing definite article 10.1%
2. Bare singular count noun for plural 8.8%
3. Redundant definite article 8.5%
4. Mis-selection of preposition 6.1%
5. Lexical misconception 5.8%
6. Wrong tense and aspect 3.8%
7. S-V non-agreement 2.4%
8. Wrong collocation 2.1%
9. Missing ‘a’/’an’ 2.0%
10. Comma splice 2.0%

APPLICATION 2: LANGUAGE PEDAGOGY
(Fei-Yu Chuang 2005)
APPLICATION 2: LANGUAGE PEDAGOGY
(Kosem & Krishnamurthy 2007)

General English: *say* = “speak” http://acorn.aston.ac.uk/

Academic English: *say* = “write” http://acorn.aston.ac.uk/

*argue* = “quarrel” (General English); = “propose/claim” (Academic English) http://corpus.byu.edu/bnc/
APPLICATION 2: LANGUAGE PEDAGOGY
(Kosem & Krishnamurthy 2007)

(chemistry, education, etc)

(lecture, discussion, etc)

(chemistry, education, etc)

(article, essay, etc)

GENRE

GENRE

LEVEL

expert

learner

SPOKEN

WRITTEN

MODE
Corpora vs Coursebooks:

• Coursebooks often use made-up or heavily edited, unnatural text
• Glossaries and grammar explanations are severely restricted to specific context
• Often out-of-date
• Follow a grammatical syllabus
• Limited varieties of text
Parallel Corpora (contents: source/translated) shows Non-Word Equivalence

Comparable corpora (same topic/domain)

Theory vs data (e.g. text length)
change to the system. A system, such as the current one, which is capable of making so few decisions, be they approvals or prohibitions, is something less than a valid system.

Frankly, I believe that this modernisation has been satisfactory. Through his management, Commissioner Monti achieves good results, as did his predecessor and, clearly, he has established a guarantee which, in parallel with the creation and development of the European internal market, has been capable of establishing the corresponding corrective measures so that the market economy within the Union may work adequately, without the distortions which we economists know may arise when the market is enlarged, as has happened in the European Union since 1993.

The credibility of the efforts we make here in Parliament to employ our finances properly stands and falls with the efforts we make to deal with such scandals and prevent them from arising in the future. This does not just mean administrative changes but also structural changes; in other words we must create instruments which are actually capable of affording support to our position in the European Union.
Table 3: A Bilingual Concordance Based on Aligned Sentences

drug/médicaments ("good" drug)

208848: it states in the form of changes to these patents

18251978: on aux États-Unis en modifiant ces lois sur les médicaments brevetés, pendant ou avant les ne

680714: s of our lumber to the United States. SENT %s.c. drug patent law amendments SENT %brev SENT >|

18731544: . les modifications à la loi sur les brevets de médicaments SENT %brev SENT >| m. steven w . l

894609: nely the human body, so that it resists whatever drug was used to fight it. SENT there is ample pr

18974531: de sorte qu'il résiste maintenant à tous les médicaments. SENT nous avons maintenant la preu

952490: trade talks and a possible Canadian move to hike drug prices in Canada when documents released yest

19035379: étude notamment sur le rapport entre les prix des médicaments et le libre-échange ? SENT >| des

952520: e the relationship between, among other things, drug prices and free trade ? SENT >| some bon . m

19035379: étude notamment sur le rapport entre les prix des médicaments et le libre-échange ? SENT >| des

1106045: the Americans a commitment to bring in a generic drug bill which will increase the drug costs of ca

19194437: ricains de présenter un projet de loi sur les médicaments non brevetés qui fera perdre 200 o

drug/drogue ("bad" drug)

941162: ted that the chief promoters of that program were drug smugglers and gun-runners and in no way were

19023479: eurs de ce programme étaient des trafiquants de drogue et d'armes, parfaitement incapables de fa

1288675: ned out that these people were in fact a group of drug smugglers and gun runners who knew nothing at

19384390: taient en réalité un groupe de trafiquants de drogue et de marchands d'armes qui ne connaissaie

1525398: r. speaker, is the minister suggesting that the drug traffickers, offenders and those with armed

19630303: ordre le statut de réfugié aux trafiquants de drogue, aux criminels et aux personnes armées ?

1808214: al with problems they might have with alcohol and drug abuse, as well as generally adjusting to com

19924350: nt régler leurs problèmes d'alcoolisme et de drogue et c'est à ce milieu enfin qu'ils devro

2224048: my riding was one which helped young people with drug problems. SENT that was cut back altogether

20355996: roupies aidait les jeunes qui s'adonnaient à la drogue. SENT il ne touche plus un sou. SENT >si
APPLICATION 3: TRANSLATION

COMENEGO (Corpus Multilingüe de Economía y Negocios): 4-grams

| 1 5779 de l article L | 1 2587 de millones de euros |
| 2 4152 à l article L | 2 2491 del Mercado de Valores |
| 3 3247 l article L du | 3 2470 de la Ley de |
| 4 1916 dans les conditions prévues | 4 2033 la Ley de de |
| 5 1749 article L du code | 5 1829 Nacional del Mercado de |
| 6 1621 à l article R | 6 1826 Comisión Nacional del Mercado |
| 7 1598 alinéa de l article | 7 1756 del Consejo de Administración |
| 8 1576 dispositions de l article | 8 1615 la Comisión Nacional del |
| 9 1512 de l article R | 9 1382 el Consejo de Administración |
| 10 1456 dans le cadre de | 10 1281 artículo de la Ley |
| 11 1329 du Code de commerce | 11 1206 KONINKLIJKE AHOLO N V |
| 12 1324 application de l article | 12 1197 en el caso de |
| 13 1316 www economie cours fr | 13 1182 el artículo de la |
| 14 1259 à compter de la | 14 1083 lo dispuesto en el |
| 15 1216 c est à dire | 15 1073 en el artículo de |
| 16 1210 Autorité des marchés financiers | 16 1068 los millones de euros |
| 17 1199 prévues à l article | 17 1016 Consejo de Administración de |
| 18 1198 l de l article | 18 1008 Real Decreto de de |
| 19 1109 L du code de | 19 945 de pérdidas y ganancias |
• Theory vs data (e.g. text length)

• Translated texts tend to be longer than originals

• “There is almost general tendency, irrespective of the translator’s identity, language, genre, period and the like, to explicitate in the translation information that is only implicit in the original text.”

  (Toury, 1980)
## APPLICATION 3: TRANSLATION

George Orwell *1984* (info from TELRI)

<table>
<thead>
<tr>
<th>Language</th>
<th>Sentences</th>
<th>Paragraphs</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>6701</td>
<td>1286</td>
<td>104,302</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>6649</td>
<td>1321</td>
<td>87,235</td>
</tr>
<tr>
<td>Czech</td>
<td>6714</td>
<td>1285</td>
<td>80,366</td>
</tr>
<tr>
<td>Estonian</td>
<td>6658</td>
<td>1289</td>
<td>79,334</td>
</tr>
<tr>
<td>Hungarian</td>
<td>6732</td>
<td>1292</td>
<td>81,147</td>
</tr>
<tr>
<td>Romanian</td>
<td>6487</td>
<td>1335</td>
<td>101,460</td>
</tr>
<tr>
<td>Slovene</td>
<td>6689</td>
<td>1288</td>
<td>91,619</td>
</tr>
</tbody>
</table>
APPLICATION 4: LITERATURE, STYLISTICS

• Michaela Mahlberg (Clusters in Dickens, 2007)

• Krishnamurthy, 1995. The Macrocosm and the Microcosm: The Corpus and The Text
  http://acorn.aston.ac.uk/acorn_publication.html

• Bill Louw (1993, 2006)
• Shakespeare (Culpeper, 2007)
<table>
<thead>
<tr>
<th>Three-word clusters</th>
<th>F</th>
<th>T</th>
<th>Four-word clusters</th>
<th>F</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>out of the</td>
<td>1,209</td>
<td>23</td>
<td>as if he were</td>
<td>452</td>
<td>21</td>
</tr>
<tr>
<td>as if he</td>
<td>1,157</td>
<td>23</td>
<td>as if he had</td>
<td>302</td>
<td>22</td>
</tr>
<tr>
<td>there was a</td>
<td>1,091</td>
<td>23</td>
<td>at the same time</td>
<td>289</td>
<td>22</td>
</tr>
<tr>
<td>it was a</td>
<td>1,050</td>
<td>23</td>
<td>in the course of</td>
<td>263</td>
<td>21</td>
</tr>
<tr>
<td>one of the</td>
<td>1,005</td>
<td>23</td>
<td>as if it were</td>
<td>249</td>
<td>23</td>
</tr>
<tr>
<td>i don't know</td>
<td>1,001</td>
<td>23</td>
<td>in the midst of</td>
<td>234</td>
<td>23</td>
</tr>
<tr>
<td>that he was</td>
<td>959</td>
<td>23</td>
<td>in a state of</td>
<td>232</td>
<td>22</td>
</tr>
<tr>
<td>the old man</td>
<td>941</td>
<td>20</td>
<td>for the first time</td>
<td>231</td>
<td>23</td>
</tr>
<tr>
<td>that it was</td>
<td>852</td>
<td>23</td>
<td>with an air of</td>
<td>225</td>
<td>18</td>
</tr>
<tr>
<td>that he had</td>
<td>808</td>
<td>23</td>
<td>i don't know what</td>
<td>224</td>
<td>23</td>
</tr>
<tr>
<td>he had been</td>
<td>783</td>
<td>22</td>
<td>i beg your pardon</td>
<td>222</td>
<td>20</td>
</tr>
<tr>
<td>i am not</td>
<td>720</td>
<td>23</td>
<td>on the part of</td>
<td>222</td>
<td>20</td>
</tr>
<tr>
<td>i apologize</td>
<td>633</td>
<td>22</td>
<td>it was his book</td>
<td>310</td>
<td>20</td>
</tr>
</tbody>
</table>
Spring. [Larkin, 1954]

Green-shadowed people sit, or walk in rings,
Their children finger the awakened grass,
Calmly a cloud stands, calmly a bird sings,
And, flashing like a dangled looking-glass,
Sun lights the balls that bounce, the dogs that bark,
The branch-arrested mist of leaf, and me,
Threading my pursed-up way across the park,
An indigestible sterility....

REVIEW
After he has established this idyllic but commonplace vision of Nature and humanity in harmony, he shocks us with the image of himself, 'an indigestible sterility'.

(Hartley 2000)
Krishnamurthy (1995):

“idyllic and commonplace”? YES, 6 corpus collocates of spring are also in the poem... flower, grass, green, sun, water, leaf

“shocks us with... indigestible sterility”? NO, negative connotations are signalled from the very first word:

green-shadowed (dangled, arrested, threading, pursed-up)

green and shadow do not collocate AT ALL in the corpus

green collocates with shade

shade has positive connotations; shadow has negative ones

Other features discovered

/s/ sounds = 23 out of 99 words: arrested, best, glass, grass, gratuitous, immodest, indigestible, least, lights, mist, most, most, pursued, seasons, see, sings, sit, spring, spring, stands, sterility, sun, use

shorter words (compared to corpus average)

very common words: untaught is the rarest; not difficult: un+taught

the hyphenated forms are very unusual – 0 occurrences in corpus
APPLICATION 5: FORENSIC LINGUISTICS

- Malcolm Coulthard & Aston Centre for Forensic Linguistics:
- Bentley & Craig/ “Let Him Have It”
- West Midlands Serious Crime Squad
- Witness statements
- Suicide notes
- Author identification
- Text messages

- Criminal/victim mediation – Lynne Cameron (Leeds)
Derek William Bentley (30 June 1933 – 28 January 1953) was a British teenager hanged for the murder of a police officer, committed in the course of a burglary attempt. The murder of the police officer was committed by a friend and accomplice of Bentley's, Christopher Craig, then aged 16. Bentley was convicted as a party to the murder, by the English law principle of "joint enterprise". This created a cause célèbre and led to a 45-year-long campaign to win Derek Bentley a posthumous pardon, which was granted partially in 1993, then completely in 1998.

... a "confession" recorded by Bentley, which was claimed by the prosecution to be a "verbatim record of dictated monologue", was shown by forensic linguistics methods to have been largely edited by policemen. Linguist Malcolm Coulthard showed that certain patterns, such as the frequency of the word "then" and the grammatical use of "then" after the grammatical subject ("I then" rather than "then I"), was not consistent with Bentley's use of language (his idiolect), as evidenced in court testimony. These patterns fitted better the recorded testimony of the policemen involved. This is one of the earliest uses of forensic linguistics on record.

http://en.wikipedia.org/wiki/Derek_Bentley_case
Among my high profile cases are the Birmingham Six Appeal, the Bridgwater Four Appeal and the Ronald Bolden trial, at the conclusion of which, in 1989, the West Midlands Serious Crime Squad was disbanded. 

[http://www1.aston.ac.uk/lss/staff/coulthardm/](http://www1.aston.ac.uk/lss/staff/coulthardm/)

**Professor Malcolm Coulthard**, Director of the Centre, was instrumental in securing the conviction of David Hodgson for the murder of Jenny Nicholl, a teenager from North Yorkshire who disappeared in 2005. By analysing text messages sent from Jenny’s phone after she disappeared, and comparing those with messages sent before her disappearance and those from the defendant’s phone, Professor Coulthard was able to deduce that Hodgson had sent the messages on Jenny’s behalf. His expert testimony was instrumental in securing the conviction.

[http://www1.aston.ac.uk/research/case-studies/forensic-linguistics/](http://www1.aston.ac.uk/research/case-studies/forensic-linguistics/)
APPLICATION 6: SOCIOLOGY

- **Sociology** > **Social theory** > Karl Marx, Antonio Gramsci, Louis Althusser, Jürgen Habermas, Michel Foucault, Pierre Bourdieu, etc

- **Linguistics** > **Sociolinguistics** > **Critical Discourse Analysis** (CDA) = interdisciplinary study (humanities and social sciences) of discourse > language as social practice > social and political issues as evidenced in discourse > Norman Fairclough *Language and Power* 1989; Ruth Wodak, Teun A. van Dijk, Theo Van Leeuwen, Christina Schäffner, Gunther Kress, etc


- **Critical Discourse Analysis** > **Corpus Linguistics Methods** > Lancaster (McEnery, Baker, Gabrielatos, etc)

- **Sociology** > **Critical Discourse Analysis**
APPLICATION 6: SOCIOLOGY

• Krishnamurthy (1996): [http://corpus.aston.ac.uk/acorn_publication.html](http://corpus.aston.ac.uk/acorn_publication.html)

Ethnic, racial and tribal: The language of racism? 4 news articles:

• YUGOSLAVIA: multi-ethnic state, six main ethnic groups and three major religions
• KENYA: tribal violence, tribal fighting, tribal fighting, tribesmen, ethnic violence
• UK: ethnic recruiting rate, ethnic minority recruitment, Commission for Racial Equality, black recruits, ethnic minorities, ethnic minority police officers, black and Asian people, ethnic minorities
• SOUTH AFRICA: mixed-race politician, inter-racial marriage

NOTES

• tribe is never used for Yugoslavia or UK (NB humour)
• If tribe is the technical term for Kenya, why one use of ethnic?
• UK: ethnic means ‘black /black and Asian’ (not Irish/Welsh, or Italian/Polish)
• UK: ‘impersonal’ use: help applications not help applicants
• UK/Kenya: ‘impersonal’ use of numbers and statistics
## APPLICATION 6: SOCIOLOGY

Krishnamurthy (1996):  
[http://corpus.aston.ac.uk/acorn_publication.html](http://corpus.aston.ac.uk/acorn_publication.html)  
*Ethnic, racial and tribal: The language of racism?*

<table>
<thead>
<tr>
<th>ethnic:</th>
<th>racial:</th>
<th>tribal:</th>
</tr>
</thead>
<tbody>
<tr>
<td>groups</td>
<td>discrimination</td>
<td>assembly</td>
</tr>
<tr>
<td>minorities</td>
<td>non</td>
<td>grand</td>
</tr>
<tr>
<td>and</td>
<td>multi</td>
<td>leaders</td>
</tr>
<tr>
<td>minority of</td>
<td>equality</td>
<td>chiefs</td>
</tr>
<tr>
<td>violence in</td>
<td>and</td>
<td>and</td>
</tr>
<tr>
<td>cleansing</td>
<td>of</td>
<td>groups</td>
</tr>
<tr>
<td>group</td>
<td>south</td>
<td>killings</td>
</tr>
<tr>
<td>albanians</td>
<td>a</td>
<td>navajo</td>
</tr>
<tr>
<td></td>
<td>commission</td>
<td>in</td>
</tr>
<tr>
<td></td>
<td>ethnic</td>
<td>a</td>
</tr>
</tbody>
</table>
FEMINISM

Corpus Analysis: Bank of English

Collocation list by t-score

<table>
<thead>
<tr>
<th>Collocate</th>
<th>T-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>of</td>
<td>1211</td>
</tr>
<tr>
<td>women</td>
<td>204</td>
</tr>
<tr>
<td>and</td>
<td>965</td>
</tr>
<tr>
<td>radical</td>
<td>131</td>
</tr>
<tr>
<td>feminism</td>
<td>110</td>
</tr>
<tr>
<td>has</td>
<td>248</td>
</tr>
<tr>
<td>is</td>
<td>415</td>
</tr>
<tr>
<td>wave</td>
<td>70</td>
</tr>
<tr>
<td>lesbian</td>
<td>64</td>
</tr>
<tr>
<td>political</td>
<td>76</td>
</tr>
<tr>
<td>post</td>
<td>65</td>
</tr>
<tr>
<td>feminist</td>
<td>58</td>
</tr>
<tr>
<td>anti</td>
<td>55</td>
</tr>
<tr>
<td>socialist</td>
<td>42</td>
</tr>
<tr>
<td>her</td>
<td>123</td>
</tr>
<tr>
<td>modern</td>
<td>45</td>
</tr>
<tr>
<td>men</td>
<td>55</td>
</tr>
<tr>
<td>about</td>
<td>115</td>
</tr>
<tr>
<td>politics</td>
<td>40</td>
</tr>
<tr>
<td>new</td>
<td>96</td>
</tr>
<tr>
<td>second</td>
<td>54</td>
</tr>
</tbody>
</table>

[ NB check semantic or pragmatic value of collocates in concordances ]

a) political: radical, political, post, anti, socialist, politics, movement, Marxism, therapy, theory, liberal, minority, gender, black
b) sexual: lesbian, gay, lesbianism
c) qualitative/evaluative: wave, modern, new, social, equality, backlash, liberation, contemporary (?)
d) western (?)
e) second (?), scale, rise
f) 1970s
## Corpus Analysis: IDS corpus

Collocates of ‘Feminismus’ categorised into 8 thematic clusters

<table>
<thead>
<tr>
<th>Category</th>
<th>Collocates</th>
</tr>
</thead>
<tbody>
<tr>
<td>political movements/ideas</td>
<td>radikal (8) Ökologie (16) kämpferisch (8) Sozialismus (11) Vertreter (11) militant (7) Kommunismus (10) Marxismus (3) existieren real (6) extrem (3) Pazifismus (3) Ideologie (6) politisch (15) Nationalismus (3) Antisemit (3) Revolution (3)</td>
</tr>
<tr>
<td>sex, gender roles/body</td>
<td>Frau (119) Machismo (10) Sexismus (10) Geschlecht (11) männlich (14) feminin (5) weiblich (12) sexual (6) schwul (4) sexuell (3) sexy (3) Weib (3) Mutter (5)</td>
</tr>
<tr>
<td>feminist ideas,</td>
<td>Emanzipation (13) Ikone (12) Gleichberechtigung (20) neu (61) feministisch (10) Feministin (6) Wegbereiter (3) Schwarz Alice (6)</td>
</tr>
<tr>
<td>places</td>
<td>amerikanisch (11) westlich (7) irisch (3)</td>
</tr>
<tr>
<td>time</td>
<td>70er (7) siebziger Jahr (4) Zeit (5) Jahr (14) Jahrzehnt (13) heute (6)</td>
</tr>
<tr>
<td>religion</td>
<td>Schabbat (3) Bibel (3)</td>
</tr>
<tr>
<td>social trends</td>
<td>Konsumgesellschaft (7) Esoterik (3) sozial (4) Pop (5)</td>
</tr>
</tbody>
</table>
• Krishnamurthy (2003)
• Orpin (2005)
• UK Party Manifestos 1900-2010 (ACORN – Aston Corpus Network)
• BBC: Cameron speeches
• USA: Presidential Candidate debates; Obama speeches

The word that won the 1997 UK General Election:

- from a German adjective, meaning ‘from Silesia’, referring to a thin, cheap cloth
- entered English language as adjective in C17
- gradual change of meaning: ‘cheap’ > ‘shoddy’ > ‘poor quality’ > (places) ‘dirty, disreputable’ > (people) clothes, appearance, behaviour, character
- abstract noun 1: ‘state of being ADJ’ formed with –ness
APPLICATION 7: POLITICS

• not in OED (1971)
• not in 20m-word BCET corpus (1986)

• was in Collins English Dictionary (1986)...
• ... and even the tiny Collins Gem (1985-87)
• was in the Bank of English corpus (1990-95)
  1990 (c. 50m-words) : 0.6 per 1m
  1992 (c. 100m-words) : 1.9 per 1m
  1995 (211m-words) : 8.7 per 1m
### APPLICATION 7: POLITICS

**sleaze: 1997**

**COLLOCATES** in Bank of English (323m-words)

<table>
<thead>
<tr>
<th>Word</th>
<th>Count</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>of</td>
<td>455</td>
<td>10.209692</td>
</tr>
<tr>
<td>allegations</td>
<td>69</td>
<td>8.264774</td>
</tr>
<tr>
<td>tory</td>
<td>67</td>
<td>8.111882</td>
</tr>
<tr>
<td>and</td>
<td>329</td>
<td>6.222366</td>
</tr>
<tr>
<td>nolan</td>
<td>36</td>
<td>5.987291</td>
</tr>
<tr>
<td>scandal</td>
<td>29</td>
<td>5.341044</td>
</tr>
<tr>
<td>government</td>
<td>40</td>
<td>5.090982</td>
</tr>
<tr>
<td>political</td>
<td>32</td>
<td>5.069198</td>
</tr>
<tr>
<td>party</td>
<td>31</td>
<td>4.798490</td>
</tr>
<tr>
<td>labour</td>
<td>25</td>
<td>4.576461</td>
</tr>
<tr>
<td>inquiry</td>
<td>21</td>
<td>4.470166</td>
</tr>
<tr>
<td>tories</td>
<td>20</td>
<td>4.406132</td>
</tr>
</tbody>
</table>

- **NOTE:** *Tory/Tories, government* (i.e. the Tory one) ...
- ... **BUT** *WHY Labour*?
sleaze: 1997

from CONCORDANCE of Labour as collocate:

...Labour allegations of Tory ‘sleaze’...

...Labour had no truck with sleaze?...

...Labour's campaigning over sleaze...

...Labour Seeks Anti-Sleaze Act...

...Labour critics of sleaze...

...veteran Labour campaigner against sleaze...
BLAIR OR BROWN?
Number of mentions of Tony Blair and Gordon Brown in David Cameron's speeches since becoming Conservative leader

Brown becomes leader of Labour Party (27 June 2007)

DATE OF SPEECH

SOURCE: BBC Analysis & Research
SHADOW CABINET
Number of mentions of members of the shadow cabinet and other Tory "big guns" in David Cameron's speeches since becoming Conservative leader

Scaled indicator of number of number of mentions

- Iain
- Duncan Smith
- George
- Osborne
- David
- Davis
- Andrew
- Lansley
- William
- Hague

Number of mentions

SOURCE: BBC Analysis & Research
THATCHER REMEMBERED
Number of mentions of Margaret Thatcher in David Cameron's speeches since becoming Conservative leader

SOURCE: BBC Analysis & Research/ picture: Getty images
VARIATIONS ON A THEME
Themes and issues referred to in David Cameron's speeches since becoming Conservative leader

SOURCE: BBC Analysis & Research
Most-used words in David Cameron's speeches since becoming Conservative leader

people (1933)
government (1456)
social (964) just (856) new (847) change (766)
society (752) work (746) country (743) local (711)
britain (696) conservative (673) responsibility (658)
world (646) public (632) party (589) state (512) tax (497)
children (492) help (472)

() indicate number of times word spoken

SOURCE: BBC Analysis & Research
• Computational linguistics
• Information retrieval, search algorithms
• Semantic Web
• Anaphora resolution, named entity recognition, question-answering
• Artificial Intelligence: talking fridges, etc
• See corpora list
Item #13896 (30 Jul 2009 08:03) - [Corpora-List] Languages in Biology and Medicine (LBM 2009): Second Call for Papers

LBM is an interdisciplinary forum that brings together researchers in biology, chemistry, medicine, public health and informatics to discuss and exploit cutting edge language technology. Language, in its many forms, is the universal means to represent, question, and convey knowledge. Likewise applied technologies such as information extraction, summarization, and translation are key technologies for advancing biomedical research and healthcare provision. The automation of all these solutions empowers our ability to discover new knowledge.

Papers focusing on application aspects of languages in biology, chemistry, medicine, public health as well as theoretical aspects are welcome.

Item #13716 (29 May 2009 19:03) - [Corpora-List] Research associate -- National Centre for Text Mining

We seek to appoint one research associate in text mining to work within a project on “Chemistry using Text Annotations” in cooperation with the University of Cambridge, the Royal Society of Chemistry and Thompson Reuters.

Item #13495 (9 Apr 2009 17:24) - [Corpora-List] Job opening: text mining, National Centre for Text Mining, University of Manchester

The National Centre for Text Mining, School of Computer Science, seeks to appoint a research associate in text mining for 15 months. You will work within a project funded by JISC on Chemistry using Text Annotation in cooperation with the University of Cambridge, the Royal Society of Chemistry and Thompson Reuters.
Science, Artificial Intelligence, or Genetic/Molecular Biology.

---

Item #2622 (4 Mar 2002 12:40) - Corpora: Call for Papers: Computerm 2 at COLING
- domain-dependent IE/NE tasks with reference to terms and terminology, such as the treatment of Biology or Medical concept/nomenclature.
- IR/indexing with domain-dependent lexical perspectives

---

Item #2615 (27 Feb 2002 14:17) - Corpora: NLPBA 2002: CALL FOR PARTICIPATION
different text processing areas (information retrieval, natural language processing...), and experts from medicine and molecular biology (genomics, proteomics...). It is a unique opportunity to have a synthetic view on the...

---

Item #2537 (4 Feb 2002 14:10) - Corpora: ACL Workshop CFP: Effective Tools and Methodologies for Teaching NLP/CL
Linguistics, Information Retrieval, Speech Science, Cognitive
Science, Artificial Intelligence, or Genetic/Molecular Biology.

---

- using previously neglected methods, including those from non-NLP fields, such as biology, nuclear physics, or finance, with promising results and/or reasonable potential for the future;

---

Item #2485 (22 Jan 2002 10:07) - Corpora: CFP: The 6th International Colloquium on Grammatical Inference
grammar induction in natural language acquisition, computational...
• WBE
• COMENEGO
• Sentiment analysis
• Pattern matching – correlating with astrology, etc
USING AN EXISTING CORPUS

- Free, registration, subscription, purchase
- General reference corpora: various languages
- BNC, ANC, BYU
- Specialist corpora
- Sketchengine €40/yr
CREATING YOUR OWN CORPUS:

1. collecting data

- Project Gutenberg for out-of-copyright books
- Lexis-Nexis for newspapers
- Copyright for sharing corpus (not needed for personal research)
- Metadata and Documentation! E.g. web – changes constantly; lose contact with contributors (eg students)
### CREATING YOUR OWN CORPUS:

2. Need to obtain copyright permission

(Wilkinson 2006)

<table>
<thead>
<tr>
<th>Low</th>
<th>Grey Area</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus used for private study &amp; research within an educational institution</td>
<td>Multiple copies accessible to students &amp; colleagues for study or research within an educational institution</td>
<td>Multiple copies accessible to staff and students for study or research outside the educational institution</td>
</tr>
<tr>
<td>Users are able to see only very short concordance lines</td>
<td>Users are able to see extended context (e.g. a few paragraphs*)</td>
<td>Users are able to view the entire text of the corpus</td>
</tr>
<tr>
<td>Research papers and articles read by a relatively small audience containing very limited citations of concordance lines</td>
<td>Articles read by a wide audience containing extensive citations of concordance lines</td>
<td></td>
</tr>
</tbody>
</table>

CREATING YOUR OWN CORPUS:  
3. Corpus Software

- Alice in Wonderland – Wordsmith
- Alice in Wonderland – Antconc
- Sketchengine
- Web as Corpus – Webcorp, Wacky
- WORDLE
- POS tagging services – Leeds/Lancaster
- Frequency lists, UCREL
- Compleat Lexical Tutor
- David Lee
References

• Biber 2005
• Fei-Yu Chuang 2005
• Cobuild, 1987
• Cobuild 1995
• Crystal, 1991
• Hartley 2000
  – http://ijl.oxfordjournals.org/content/2/1/44.short
• Tim Johns
• Kosem & Krishnamurthy 2007
• Krishnamurthy 1994? spring
• Larkin, 1954
• Lewis (1993)
• Sinclair, 1987
• Sinclair, 1991
• Sinclair 1995
• Thornbury, 2006
• Toury, 1980
• Willis, 1990
• **COLLOCATION:** Cobuild CD-Rom (1995); Oxford Students Dictionary (2004)
• **FORMULAIC:** N. Schmitt (2004); M. McCarthy (1997)
• **METAPHOR:** A. Deignan (2005); G. Low/L. Cameron; K. Wikberg (2008)
• **SPOKEN v WRITTEN:** D. Biber (2006); M. McCarthy (2001); S. Adolphs (2010)
• **GRAMMAR:** S. Hunston/G. Francis (Pattern Grammar, 2000); D. Biber
  (Spoken and Written, 1999)
• **VARIETIES/DIALECTS:** ICE; Leeds (Atwell: Br/Am, 2007); Scots; Irish
• **HISTORICAL/DIACHRONIC:** Helsinki
• **CONTRASTIVE:** Larrivee & Krishnamurthy (rare syntactic sequences: def det+indef pron; Eng/Fr, 2009); GeWiss
• **Collocation** - Lewis (2000), Barfield & Gyllstad (2009)
• **Authentic texts and examples** - Ranalli (2003), Gavioli & Aston (2001), Koprowski (2005)
• **Learner corpora** - ICLE (Granger, 1993), L1-specific vs generic L2 problems, overuse and underuse (Berber-Sardinha, 1996, 2004)
• Bilingual concordance (Ken Church?)
• **COMENEGO (Corpus Multilingüe de Economía y Negocios)**
• **TELRI (Multext East?? Orwell)**
• Michaela Mahlberg (Clusters in Dickens, 2007)
• Krishnamurthy, 1995. *The Macrocosm and the Microcosm: The Corpus and The Text*
  [http://acorn.aston.ac.uk/acorn_publication.html](http://acorn.aston.ac.uk/acorn_publication.html)
• Bill Louw (1993, 2006)
• Shakespeare (Culpeper, 2007)
• Malcolm Coulthard & Aston Centre for Forensic Linguistics:
• Bentley & Craig/ “Let Him Have It”
• West Midlands Serious Crime Squad
• Witness statements
• Suicide notes
• Author identification
• Text messages

• Criminal/victim mediation – Lynne Cameron (Leeds)
• Norman Fairclough *Language and Power* 1989
• Ruth Wodak, Teun A. van Dijk, Theo Van Leeuwen, Christina Schäffner, Gunther Kress, etc

• **Corpus Linguistics** > analyses discourse > offers insights into social issues > Stubbs (1992), Stubbs and Gerbig (1993), Caldas-Coulthard (1993), Fox (1993), Louw (1993), Krishnamurthy (1996), etc

• **Critical Discourse Analysis** > **Corpus Linguistics Methods** > Lancaster (McEnery, Baker, Gabrielatos, etc)

• Krishnamurthy (1996):

• Jaworska & Krishnamurthy
• Krishnamurthy (2003)
• Orpin (2005)
• UK Party Manifestos 1900-2010 (ACORN – Aston Corpus Network)
• BBC: Cameron speeches
• USA: Presidential Candidate debates; Obama speeches
• WBE
• COMENEGO
• Wilkinson 2006