Using Online Corpora for Synchronic and Diachronic Studies

Jong-Bok Kim
jongbok@khu.ac.kr
Kyung Hee University, Seoul

2014 Winter Conference of the KASELL
(The Korean Association for the Study of English Language and Linguistics)
Dec 6, 2014
Sungkyunkwan University
Outline of the Talk

1. Introduction
2. Resurgence of Corpus linguistics
3. Using Online Corpora: Synchronous Aspects
4. Diachronic Studies
5. Global English: Dialectal Studies
6. Doing More Serious Studies
7. Conclusion
What is a corpus?

A corpus is a large, principled collection of naturally occurring texts (either written or spoken) stored electronically.
What corpus approach can do?

- It can give insights into how language is **really used**, rather than how people **think** it is used (cf. rationalism vs. empiricalism)
Empiricism vs. Rationalism

- From the early 1900s until the 1950s, linguistics focused on ‘empirical’ data (research on the unknown languages required to gather and organize large amounts of data from native speakers).
- In the 1950s, Chomsky challenged the empirically-oriented research, advocating ‘intuition-based’ approaches.
Arguments against the Empiricism

- Available corpora were simply too small to provide meaningful data
- Corpora do not explain why some constructions are unacceptable
- Data are about ‘real’ world, not about our linguistic faculty.
methodological problems with too much reliance on linguistic ‘intuition’: huge gaps between the intuition-based data by linguists (to support a particular theory or analysis) and the authentic data from 1970s, rapidly growing interest in ‘realistic’ grammars based on the ‘functional’, ‘corpus-based’, ‘performance’ based perspectives around 1990s, rapid development of computer science: big size of corpus (e.g., 1 million to 2 billion words), software, online, etc. with a huge size of balanced corpus data, corpora can produce fairly trivial data as well as linguistically insightful data
More on corpus linguistics

- Corpus linguistics is not a ‘science’, but ‘methodology’
- It argues for the importance of data, a frequency-based approach to linguistic phenomena.
- It focuses on acquiring, organizing, and correctly interpreting the primary data, rather than overtly-abstract theory that may or may not be based on the accurate data.
What corpus approach is NOT?

- NOT able to provide negative evidence
- NOT able to provide all possible language at one time
- NOT able to explain why
Why interpretation is so important: Quantitative vs. qualitative

- Evidence always requires interpretation, and the interpretative, critical skills of the humanities researcher are still highly prized in the discussion about what corpus data actually mean.

- A good corpus linguist should be able to handle both qualitative and quantitative analysis; many linguists supplement quantitative analysis with interpretative, qualitative analyses of corpus data.
Some online corpora for English


- **COHA: Corpus of Historical American English**: 400 million, 1810-2010, Modeling linguistic change (lexical, morphological, syntactic, semantic change).

- **TIME corpus**: 100 million words from *Time* magazine.

- **GloWbe: Corpus of Global Web-based English**: composed of 1.9 billion words from 1.8 million web pages in 20 different English-speaking countries.

- **BNC-Web: British National Corpus**: 100 million, a web-based client program for searching and retrieving lexical, grammatical and textual data (cf. BYU-BNC).
Basic Search Methods

- Using **word lists and frequency**: vocabulary is a central foundation to language learning; corpus tools can generate word lists.

- Using **concordance lines**: knowing which words go together and which words do not go together is a puzzle (e.g., *big/large* corpus tools can generate KWIC (Key Word in Context Indexes)).

- Using **tagged/parsed texts**: words can have different grammatical roles. Tagged texts can be useful in dealing with words that have multiple functions (e.g., *well, can, ...*).

- Checking **Role of register**: spoken vs. written; face-to-face vs. phone; business-phone vs. personal phone conversation (cf. ICE-GB: International Corpus of English, Great Britain).
A Quick Guide: Simple query syntax

- One exact word (e.g., *seedy* originated from *seed*: seedy looking, seedy bar?)
POSt (e.g., V-ing)
Use the dropdown box
POS tagging in the BNC-WEB

- **AJ0** Adjective (general or positive) (e.g. good, old, beautiful)
- **AJC** Comparative adjective (e.g. better, older)
- **AT0** Article (e.g. the, a, an, no)
- **NN0** Common noun, neutral for number (e.g. aircraft, data, committee)
- **NN1** Singular common noun (e.g. pencil, goose, time, revelation)
- **NP0** Proper noun (e.g. London, Michael, Mars, IBM)

**Lemma:** \( \{ \text{light} / V \} \)
Lemma: use the bracket \[be\]

- Lemma (e.g., be, am, was, are, were)
- Option: Group by lemma

**CORPUS OF CONTEMPORARY AMERICAN ENGLISH**

450 MILLION WORDS, 1990-2012 [DOWNLOAD ALL 190,000 TEXTS]

<table>
<thead>
<tr>
<th>CONTEXT</th>
<th>FREQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>'RE GOING</td>
<td>81767</td>
</tr>
<tr>
<td>'S GOING</td>
<td>66965</td>
</tr>
<tr>
<td>IS GOING</td>
<td>49718</td>
</tr>
<tr>
<td>WAS GOING</td>
<td>42082</td>
</tr>
<tr>
<td>ARE GOING</td>
<td>35717</td>
</tr>
<tr>
<td>'M GOING</td>
<td>31439</td>
</tr>
<tr>
<td>IS BEING</td>
<td>19481</td>
</tr>
<tr>
<td>ARE BEING</td>
<td>17517</td>
</tr>
<tr>
<td>'RE TALKING</td>
<td>16370</td>
</tr>
<tr>
<td>WERE GOING</td>
<td>16340</td>
</tr>
<tr>
<td>'RE DOING</td>
<td>15157</td>
</tr>
<tr>
<td>WAS BEING</td>
<td>11094</td>
</tr>
</tbody>
</table>
Regular Expressions (wild cards)

- Wild cards (e.g., single, song, singing ...)
- Wildcard: * = any # letters (e.g., un*ly)
- Wildcard: ? = one letter (e.g., s?ng)
- Wild card for searching syntactic patterns: [make] * Adjectives
Collocation

- The tendency of specific words to be used together.
- Words found in the company of other words
- Meaningful with a large corpus
- Can give us a picture of the typical environment of words and insights into unusual patterning
Collocation

- Nouns after the phrase *look into*: *look into* [\(n\ast\)] 05
- Verbs with the expression *V NP into V-ing*: *into* [\(v\ast g\ast\)] [\(vv\ast\)] 40

---

**CORPUS OF CONTEMPORARY AMERICAN ENGLISH**

<table>
<thead>
<tr>
<th>450 MILLION WORDS, 1990-2012</th>
<th>[DOWNLOAD ALL 190,000 TEXTS]</th>
</tr>
</thead>
</table>

**DISPLAY**
- LIST
- CHART
- KWIC
- COMPARE

**SEARCH STRING**
- WORD(S): thick
- COLLOCATES: [\(n\ast\)]
- POS LIST
- RANDOM:

**SECTIONS**
- SHOW

**SORTING AND LIMITS**
- SORTING: FREQUENCY
- MINIMUM: FREQUENCY

**CONTEXT**

<table>
<thead>
<tr>
<th></th>
<th>CONTEXT</th>
<th>FREQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HAIR</td>
<td>1003</td>
</tr>
<tr>
<td>2</td>
<td>GLASSES</td>
<td>5153</td>
</tr>
<tr>
<td>3</td>
<td>SMOKE</td>
<td>287</td>
</tr>
<tr>
<td>4</td>
<td>LAYER</td>
<td>365</td>
</tr>
<tr>
<td>5</td>
<td>ACCENT</td>
<td>219</td>
</tr>
<tr>
<td>6</td>
<td>SLICES</td>
<td>195</td>
</tr>
<tr>
<td>7</td>
<td>SKIN</td>
<td>193</td>
</tr>
<tr>
<td>8</td>
<td>WALLS</td>
<td>194</td>
</tr>
<tr>
<td>9</td>
<td>FOG</td>
<td>181</td>
</tr>
<tr>
<td>10</td>
<td>CLOUDS</td>
<td>166</td>
</tr>
<tr>
<td>11</td>
<td>GLASS</td>
<td>160</td>
</tr>
<tr>
<td>12</td>
<td>AIR</td>
<td>143</td>
</tr>
</tbody>
</table>
high vs. tall

(1) a. An estimated 50% of U.S. homes have unhealthily high levels of moisture. (COCA 2011 MAG)

b. ... and I do worry that my daughter is at high risk. (COCA 2012 MAG)

c. Articles with a score of greater than 75 were deemed to be high quality. (COCA 2012 ACAD)

(2) a. That’s what I’d call the tall guy in my head. (COCA 2009 FIC)

b. She finds two tall glasses and takes them back to the table. (COCA 2008 FIC)

c. It softened the air and turned the tall pines beyond into gray shadows of themselves. (COCA 2008 FIC)
Use ‘compare’ and ’collocate’
Collocation of *go* vs. *come*

(3)  
  a. I’m going to **go crazy** in this city. (COCA 2010 NEWS)  
  b. They **go bankrupt** at a high rate. (COCA 2010 NEWS)  
  c. I would **go mad** with it. (COCA 2010 FIC)  
  d. Things can **go bad** very quickly out there. (COCA 2012 NEWS)

(4)  
  a. I’ll make one of those wishes **come true** now. (COCA 2012 MAG)  
  b. It’s the details that make the story **come alive**. (COCA 2009 ACAD)  
  c. The gear doesn’t **come cheap**. (COCA 2002 NEWS)  
  d. New information has **come available**. (COCA 1999 FIC)
Using Online Corpora: Synchronic Aspects

Collocation of *go* vs. *come* in COCA

- Adjectives occurring after the verb ‘go’ are positive, but those with the verb ‘come’ are negative
- Bad things go away from us, and good things come to us.

<table>
<thead>
<tr>
<th>WORD 1 (W1): GO (1.39)</th>
<th>WORD 2 (W2): COME (0.72)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WORD</strong></td>
<td><strong>W1</strong></td>
</tr>
<tr>
<td>1 CRAFTY</td>
<td>632</td>
</tr>
<tr>
<td>2 BANKRUPT</td>
<td>776</td>
</tr>
<tr>
<td>3 WILD</td>
<td>2,04</td>
</tr>
<tr>
<td>4 UNNOTICED</td>
<td>378</td>
</tr>
<tr>
<td>5 BROKE</td>
<td>170</td>
</tr>
<tr>
<td>6 MAD</td>
<td>161</td>
</tr>
<tr>
<td>7 BLIND</td>
<td>109</td>
</tr>
<tr>
<td>8 SLOW</td>
<td>211</td>
</tr>
<tr>
<td>9 HAND-IN-HAND</td>
<td>101</td>
</tr>
<tr>
<td>10 HEAD-TO-HEAD</td>
<td>97</td>
</tr>
<tr>
<td>11 UNDETECTED</td>
<td>92</td>
</tr>
<tr>
<td>12 EXTINCT</td>
<td>90</td>
</tr>
<tr>
<td>13 UNANSWERED</td>
<td>90</td>
</tr>
<tr>
<td>14 UNPUNISHED</td>
<td>83</td>
</tr>
<tr>
<td>15 UNCHALLENGED</td>
<td>80</td>
</tr>
<tr>
<td>16 INSANE</td>
<td>77</td>
</tr>
<tr>
<td>17 NUCLEAR</td>
<td>77</td>
</tr>
<tr>
<td>18 NEGATIVE</td>
<td>76</td>
</tr>
<tr>
<td>19 SOUR</td>
<td>60</td>
</tr>
<tr>
<td>20 UNREPORTED</td>
<td>68</td>
</tr>
<tr>
<td>21 WIDE</td>
<td>130</td>
</tr>
<tr>
<td>22 BUST</td>
<td>64</td>
</tr>
<tr>
<td>23 BAD</td>
<td>249</td>
</tr>
</tbody>
</table>

Jongbok Kim (KHU)
More on statistical terms

- simple notions: frequency, mean, standard deviation

- normalization of frequencies: used when comparing two data sets of unequal size. Compare the two sizes in terms of one million

- mutual information (MI): offer substantial evidence of how commonly individual words collocate with others. In general, if an MI score higher than 3 suggests a strong bond between the search item and its collocate

- finding out if the difference in frequencies is statistically significant or not: t-tests, ANOVA, chi-square, log-likelihood test, z-score, etc.
**Frequency and Normalization (Per Million)**

Compare the size of corpus, number of hits, and Normalization: *get* in the BNC-Web

Your query "[word="got" %c]" returned 90106 hits in 3013 different texts, thinned with method random selection to 500 in 1416 texts. [0.163 seconds]

Your initial query result was thinned. Extrapolated word counts and the corresponding frequency per million words figures are given in smaller the thin-factor, the less reliable these figures will be.

<table>
<thead>
<tr>
<th>Categories</th>
<th>General information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categories (for crosstabs only)</td>
<td>no crosstabs</td>
</tr>
</tbody>
</table>

The following distribution was found:

<table>
<thead>
<tr>
<th>Spoken or Written:</th>
<th>No. of words</th>
<th>No. of hits</th>
<th>Dispersion (over files)</th>
<th>Frequency per million words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spoken</td>
<td>10,409,858</td>
<td>2,885</td>
<td>564/908</td>
<td>277.14</td>
</tr>
<tr>
<td>Written</td>
<td>87,903,571</td>
<td>2,115</td>
<td>852/3,140</td>
<td>24.06</td>
</tr>
<tr>
<td>total</td>
<td>98,313,429</td>
<td>5,000</td>
<td>1,416/4,048</td>
<td>50.86</td>
</tr>
</tbody>
</table>
Mutual Information Value

See the collocation of *cause* with the following noun:
More on the key notions

- But normalised scores aren’t proof that what you have is significant.
- Two common tests of significance are chi-square and log likelihood.
- Log likelihood (LL): If the LL for your result is greater than 6.63, the probability of the result happening by chance is less than 1%. So we can be 99% certain that the result actually means something ($p < 0.01$). If the LL is 3.84 or more, the probability of it happening by chance is less than 5%. So we are 95% certain of the result ($p < 0.05$).
Mutual Information Value

See the collocation of *start* with the following -ing verb in the BNC-Web.

<table>
<thead>
<tr>
<th>No.</th>
<th>Word</th>
<th>Total No. in written texts</th>
<th>As collocate</th>
<th>In No. of texts</th>
<th>Log-likelihood value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>crying</td>
<td>1623</td>
<td>8</td>
<td>4</td>
<td>93.833954</td>
</tr>
<tr>
<td>2</td>
<td>talking</td>
<td>8560</td>
<td>6</td>
<td>5</td>
<td>46.996872</td>
</tr>
<tr>
<td>3</td>
<td>thinking</td>
<td>10193</td>
<td>6</td>
<td>6</td>
<td>44.917790</td>
</tr>
<tr>
<td>4</td>
<td>moving</td>
<td>7456</td>
<td>5</td>
<td>5</td>
<td>38.715523</td>
</tr>
<tr>
<td>5</td>
<td>turn</td>
<td>11210</td>
<td>5</td>
<td>2</td>
<td>34.675233</td>
</tr>
</tbody>
</table>
KWIC (Key Word In Context)

The most common format for concordance lines and KWIC index is formed by sorting and aligning the words alphabetically.
Registers – genres

- COCA is a balanced corpus: spoken, fiction, magazine, newspaper
- Use the function of ‘chart’
Genres of *have + PP*

(5) a. Prosecutors there *have pursued* multiple piracy investigations. (COCA 2012 NEWS)

b. Israel *has attacked* nuclear sites in foreign countries before. (COCA 2012 NEWS)

c. Still, Romney *has struggled* to strike the right note with the masses. (COCA 2012 NEWS)

d. The presence of U.S. national team players in the EPL *has helped* to boost ratings. (COCA 2012 NEWS)

e. “It is evident that GM *has made* significant improvement in its manufacturing operations,” president Ron Harbour says. (COCA 2001 NEWS)
Genres of *have* + PP in COCA

- the most common in newspapers
- why? the newspapers are used to report the events that happened in the past but still important or relevant in the present.
Genres of *be*-passive vs. *get*-passive

(6) a. Two of these, strategy and security, are shown as new layers in its protocol stack. (COCA 2012 ACAD)
b. Her thoughts were confirmed by formal anxiety testing. (COCA 2012 ACAD)
c. All 180 audience seats were filled. (COCA 2012 NEWS)

(7) a. I don’t suspect that I will ever choose to get married. (COCA 2012 SPOK)
b. You didn’t get invited to birthday parties at school. (COCA 2012 SPOK)
c. Now he got knocked for singing the song like that. (COCA 2012 SPOK)
Genres of *be*-passive vs. *get*-passive in COCA

**Corpus of Contemporary American English**

**450 Million Words, 1990-2012** [Download All 190,000 Texts]

<table>
<thead>
<tr>
<th>SECTION</th>
<th>ALL</th>
<th>SPOKEN</th>
<th>FICTION</th>
<th>MAGAZINE</th>
<th>NEWSPAPER</th>
<th>ACADEMIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>FREQ</td>
<td>615453</td>
<td>838561</td>
<td>55093</td>
<td>112010</td>
<td>116686</td>
<td>245305</td>
</tr>
<tr>
<td>PER MIL</td>
<td>1,325.47</td>
<td>874.39</td>
<td>618.08</td>
<td>1,180.61</td>
<td>1,274.41</td>
<td>2,704.68</td>
</tr>
</tbody>
</table>

*See all sub-sections at once*
Usage of the verb *undergo*

- **core**: undergo
- **collocation**: surgery, tests, treatment, change, training, test, and so on
- **colligation**: Preceded by passive or modal (e.g., forced to, must), and followed by adjective and abstract noun (e.g., further testing, major change)
- **semantic preference**: Followed by nouns belonging to these sets—medical procedures; changes; nonmedical testing; other unpleasant things
- **discourse prosody**: Indicates that a procedure is unpleasant and involuntary
‘end up V-ing’ construction

(8) a. She’d end up calling you anyway. (COHA 2002 FIC)

b. I ended up taking it home for Ed and myself at the end of the day. (COHA 2003 FIC)

c. You ended up talking around like a mad person. (COHA 2000 MAG)

d. The very companies that do things to block competition end up hurting themselves. (COHA 2000 MAG)

e. You’re so desperate to play that you end up doing something crazy like this. (COHA 2009 NEWS)
‘end up V-ing’ construction in COHA

- the most common in spoken, but the least in academic
- has been increasingly used since 1930s
Must vs. need to in COHA

(9) a. And now I **must rush** home to give your mother the excellent news! (COHA 2000 FIC)

b. You **must give** me an answer. (COHA 2001 FIC)

c. I **must get** back to work. (COHA 1991 FIC)

d. You **must hit** the ball standing on one leg. (COHA 1991 FIC)

- keeps decreasing decade by decade
- becoming kind of old-fashion
(10) a. I need to learn my lines. (COHA 2000 FIC)
b. You need to do your best to say it correct. (COHA 2000 FIC)
c. I need to know that you trust me. (COHA 2001 FIC)
d. I need to see some progress soon. (COHA 2002 FIC)

- keeps increasing (as opposed to ‘must + VP’)
need not in COHA

(11) a. This need not take as long as I feared. (COHA 2001 FIC)

b. You need not accompany us. (COHA 2004 FIC)

c. He need not know that there is little more to follow. (COHA 2004 FIC)

d. However, the moon need not be your enemy. (COHA 2001 MAG)

e. Flora need not have worried. (COHA 2004 NF)
a lot of in COHA

(12)  a. A lot of things were better before. (COHA 2000 FIC)
b. We dumped a lot of sugar on those. (COHA 2001 FIC)
c. George Bush has a lot of power. (COHA 2007 NF)
d. “He wouldn’t be driving a lot of horses,” he interposed quickly. (COHA 1907 FIC)
e. He had a lot of current history to catch up on. (COHA 1955 FIC)
(13) a. How can you stop yourself from getting angry or showing you’re upset? (GloWbE US G)
b. This is stopping them from continuing to sully her with this ridiculous controversy. (GloWbE US G)
c. The cook caught his arm to stop him from being blown past. (GloWbE GB G)
d. It didn’t stop her from buying more dishes. (GloWbE GB G)

more or less the same across different languages
How can I stop it turning off? (GloWbE GB G)

Someone seems no less keen to stop him finding out. (GloWbE GB G)

I stopped him going through the veil. (GloWbE US G)

I stop myself catching it by washing my hands an even number of times. (GloWbE US G)

the most common in British English, but the least in American English, if without the preposition from
Binominal NP structures

English Binominal NPs (BNP) with the skeletal structure of ‘Det1 N1 of Det2 N2’ display many intriguing syntactic and semantic properties. Examples in (15) are naturally occurring BNP data extracted from the BNC:

(15)  
   a. It’s been [a hell of a day] at the office.  
   b. And it introduced her to Budapest, [a jewel of a city].  
   c. Rune nodded [his shaven dome of a head].  
   d. She had [a skullcracker of a headache].  
   e. A door opened; and into the assessment room stepped [a giant of a man].
Issues

- Template: Det1 + N1 + of Det2 + N2
- Syntax: Which one is the syntactic head?
- Semantics: What is the semantic relationship between N1 and N2?
- Pragmatics: Are there any discourse constraints?
So Big a Mess Construction

- [so big] [a mess]
- [too good] [a chance] [to miss]
- . . . French President Nicolas Sarkozy, who is [so happy of a guy] he got drunk at the G8 summit . . .
Adjectival Predeterminer (APD) Constructions

English also allows a limited range of words in the predeterminer position.

- **such type**
  
  (16)  
  a. He’s making [such] a big sandwich.  
  b. [What] a wonderful conference it is!  
  c. They have been for [many] a long day.

- **so type**
  
  (17)  
  a. Hunger was now [so] powerful a force in its life.  
  b. The accounts are all about [how] big a struggle it is.  
  c. We have far [too] great a gap between these two states.  
  d. [This] new a phoneme would have two allophones.  
  e. It’s about [that] big a diameter.  
  f. He proved far [more] successful a dealer than he had a client.
so vs. such

(18) a. Without promotion, there is no such a thing anymore.  
    <BNC HCX 352>
    b. Anything doesn’t give you any such a look ahead information.  <BNC KRM 262>

(19) a. It was so typical (of) a craze
    b. the fortunes of Syhlock have been all too typical (of) a career..
    c. How typical (of) a tech to speel this all out so literally...
    d. He was that talented (of) a football player
Along with the infinitival raising construction as in (20a) and its assumed source sentence in (20b), English also employs the so-called copy-raising construction given in (21a):

(20) a. The lifeguards seem to be dancing across the water.  
     b. It seems that the lifeguards are dancing across the water.

(21) a. The lifeguards seem like they are dancing across the water.  
     b. It seem like the lifeguards are dancing across the water.
Empirical Issues

The authentic corpus data include cases with the pronominal copying is a complex process.

(22)  
   a. The informant sounds as if he or she worked with or for Stone.  
   b. The fact that she went alone seems like she wasn’t afraid.

The pronominal copying is in the specifier of the subject, the prepositional object, and even no pronominal copying at all:

(23)  
   a. He appeared as if his heart were broken by her speech.  
   b. The bed appeared as if someone had recently been dragged from it.  
   c. The President sounded as if the world was helpless to stop the killing in Bosnia.
Nonfinite XP

The type of predicative expression is quite flexible.

(24) **Unaugmented absolutes**

a. All our savings [gone], we started looking for jobs. (Quirk et al. 1972)
b. Job offers [from three major companies], Stacey is happier than ever.

(25) **With-augmented absolutes**

a. With those two [gone], the Devil Rays got younger quicker than they expected. (COCA 2001 NEWS)
b. With eyes [full of laughter], she pushed past my leg and tossed in the boat. (COCA 2002 MAG)
What with absolute constructions (WWAC)

Typical examples:

(26)  
a. My life was pretty hectic what with the job and the writing. (COHA 2003 FIC)

b. I am nearly dead, what with hunger, and thy cruel bonds, and the gag. (COHA 1910 NF)

tenseless free adjuncts functioning as adverbial sentence modifiers

describe reasons for failure, something unfortunate happening, or not happening.
Typical examples:

(27) a. Love at first sight had coerced him into marrying a complete stranger. (COCA 2006 FIC)

b. I probably pressured him into driving around the barricades. (COCA 1997 FIC)

The construction involves causation: the subject referent causes the object referent into the state of affairs expressed by *into* -*ing* clauses.
Main search methods:

- \([vv\ast] \ 0.4 \ into \ [v?g\ast] \ \text{rather than} \ [vv\ast] \ 0.4 \ [n\ast] \ into \ [v?g\ast]\)

  The context 0.4 represents 4 or less (including zero) collocate distances between the main verb and the into gerundive.

\(28\)  
  a. She said she was coaxed into joining a tour of the frat house. (COCA 2006 SPOK)  
  b. He was forced into performing many similar surgical operations. (COCA 2009 FIC)
Filtered out cases:

1. Embedded cases: *try, let*, etc.

   (29)  a. He was also **trying** to manipulate you into changing your testimony (COCA 2012 SPK)

   b. I **let** him goad me into taking a drink (COCA 2005 FIC)

2. Different non-object control usages: *put, pour*, etc.

   (30)  a. Mrs. McDonnell is **putting** a great effort into promoting Virginia wine (COCA 2005 SPOK)

   b. Armstrong decided to **pour** his savings into opening a grocery store (COCA 2009 NEWS)

3. Mistakes in tagging V-*ing* forms but no distinction among *ger, prog, and pres part*:

   (31)  a. Thousands of others **turned** the highways into **parking** lots. (COCA 2012 NEWS)

   b. To **turn** them into **voting** booths just doesn’t make sense at this point in time. (COCA 2002 NEWS)
### Frequency

<table>
<thead>
<tr>
<th># tokens</th>
<th>Corpus Size</th>
<th>Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,848</td>
<td>450 million words</td>
<td>COCA</td>
</tr>
<tr>
<td>3,874</td>
<td>400 million words</td>
<td>COHA</td>
</tr>
<tr>
<td>1,748</td>
<td>100 million words</td>
<td>Tlme</td>
</tr>
<tr>
<td>1,130</td>
<td>100 million words</td>
<td>BNC</td>
</tr>
<tr>
<td>6,735</td>
<td>385 million words</td>
<td>Glowbe-US</td>
</tr>
<tr>
<td>6,416</td>
<td>385 million words</td>
<td>GloWbe-UK</td>
</tr>
<tr>
<td>25,357</td>
<td>1.32 billion words</td>
<td>Total</td>
</tr>
</tbody>
</table>

**Rudanko (2006):** based on 144 million words of BRE and 117 million words of AME which yielded **1,050 tokens** of the construction.
Normalized Frequency in COHA from 1800

The overall increase with all verbs has been quite consistent during the past 200 years.

Figure: Overall increase in frequency (per million words)
Alternative view: web interface

The COHA web interface also shows the overall increase in frequency, as evidenced from its frequency from 1800 to 2009.

Figure: Overall increase in frequency: COHA web interface
### Top frequency verbs in COCA

<table>
<thead>
<tr>
<th>Rank</th>
<th>COCA</th>
<th>Token No.</th>
<th>BNC</th>
<th>Token No.</th>
<th>COHA</th>
<th>Token No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TALK</td>
<td>887</td>
<td>FORCE</td>
<td>79</td>
<td>TALK</td>
<td>473</td>
</tr>
<tr>
<td>2</td>
<td>TRICK</td>
<td>536</td>
<td>TRICK</td>
<td>75</td>
<td>TRICK</td>
<td>221</td>
</tr>
<tr>
<td>3</td>
<td>FOOL</td>
<td>327</td>
<td>FOOL</td>
<td>57</td>
<td>FORCE</td>
<td>152</td>
</tr>
<tr>
<td>4</td>
<td>COERCE</td>
<td>226</td>
<td>TALK</td>
<td>51</td>
<td>FOOL</td>
<td>134</td>
</tr>
<tr>
<td>5</td>
<td>FORCE</td>
<td>226</td>
<td>MISLEAD</td>
<td>46</td>
<td>DECEIVE</td>
<td>125</td>
</tr>
<tr>
<td>6</td>
<td>PRESSURE</td>
<td>171</td>
<td>COERCE</td>
<td>36</td>
<td>FRIGHTEN</td>
<td>121</td>
</tr>
<tr>
<td>7</td>
<td>COAX</td>
<td>166</td>
<td>DECEIVE</td>
<td>36</td>
<td>SPRING</td>
<td>103</td>
</tr>
<tr>
<td>8</td>
<td>SCARE</td>
<td>125</td>
<td>BULLY</td>
<td>35</td>
<td>COAX</td>
<td>100</td>
</tr>
<tr>
<td>9</td>
<td>LURE</td>
<td>114</td>
<td>PROVOKE</td>
<td>30</td>
<td>COERCE</td>
<td>100</td>
</tr>
<tr>
<td>10</td>
<td>MANIPULATE</td>
<td>95</td>
<td>LEAD</td>
<td>29</td>
<td>MISLEAD</td>
<td>82</td>
</tr>
<tr>
<td>11</td>
<td>TRANSLATE</td>
<td>93</td>
<td>CON</td>
<td>28</td>
<td>DELUDE</td>
<td>81</td>
</tr>
<tr>
<td>12</td>
<td>BULLY</td>
<td>87</td>
<td>PRESSURE</td>
<td>21</td>
<td>BULLY</td>
<td>79</td>
</tr>
<tr>
<td>13</td>
<td>MISLEAD</td>
<td>86</td>
<td>BLACKMAIL</td>
<td>18</td>
<td>PRESSURE</td>
<td>77</td>
</tr>
<tr>
<td>14</td>
<td>DELUDE</td>
<td>83</td>
<td>PRESSURISE</td>
<td>18</td>
<td>LURE</td>
<td>71</td>
</tr>
<tr>
<td>15</td>
<td>SEDUCE</td>
<td>78</td>
<td>DRAW</td>
<td>17</td>
<td>SCARE</td>
<td>71</td>
</tr>
<tr>
<td>16</td>
<td>GOAD</td>
<td>73</td>
<td>COAX</td>
<td>15</td>
<td>BEGUILE</td>
<td>70</td>
</tr>
<tr>
<td>17</td>
<td>SHAME</td>
<td>71</td>
<td>SHOCK</td>
<td>15</td>
<td>LEAD</td>
<td>64</td>
</tr>
<tr>
<td>18</td>
<td>DECEIVE</td>
<td>69</td>
<td>DELUDE</td>
<td>14</td>
<td>GOAD</td>
<td>60</td>
</tr>
<tr>
<td>19</td>
<td>FRIGHTEN</td>
<td>68</td>
<td>TRAP</td>
<td>13</td>
<td>PROVOKE</td>
<td>53</td>
</tr>
<tr>
<td>20</td>
<td>CON</td>
<td>67</td>
<td>DUPE</td>
<td>12</td>
<td>PERSUADE</td>
<td>51</td>
</tr>
<tr>
<td>21</td>
<td>DUPE</td>
<td>59</td>
<td>GOAD</td>
<td>12</td>
<td>SEDUCE</td>
<td>51</td>
</tr>
<tr>
<td>22</td>
<td>INTIMIDATE</td>
<td>57</td>
<td>LURE</td>
<td>12</td>
<td>CAJOLED</td>
<td>50</td>
</tr>
<tr>
<td>23</td>
<td>LEAD</td>
<td>57</td>
<td>SEDUCE</td>
<td>12</td>
<td>TRAP</td>
<td>49</td>
</tr>
<tr>
<td>24</td>
<td>PROD</td>
<td>51</td>
<td>MANIPULATE</td>
<td>11</td>
<td>SHAME</td>
<td>48</td>
</tr>
<tr>
<td>25</td>
<td>LULLED</td>
<td>44</td>
<td>MOVE</td>
<td>11</td>
<td>BETRAYED</td>
<td>45</td>
</tr>
<tr>
<td>26</td>
<td>PROVOKE</td>
<td>42</td>
<td>SLIP</td>
<td>11</td>
<td>BREAK</td>
<td>45</td>
</tr>
<tr>
<td>27</td>
<td>THREW</td>
<td>42</td>
<td>FRIGHTEN</td>
<td>10</td>
<td>WHEEDLE</td>
<td>42</td>
</tr>
<tr>
<td>28</td>
<td>CAJOLE</td>
<td>41</td>
<td>TEMPT</td>
<td>10</td>
<td>CON</td>
<td>39</td>
</tr>
<tr>
<td>29</td>
<td>BRAINWASH</td>
<td>36</td>
<td>CAJOLE</td>
<td>9</td>
<td>PUSH</td>
<td>39</td>
</tr>
<tr>
<td>30</td>
<td>ENTICE</td>
<td>27</td>
<td>PANICK</td>
<td>9</td>
<td>TEMPT</td>
<td>39</td>
</tr>
<tr>
<td>31</td>
<td>TRAP</td>
<td>26</td>
<td>BRAINWASH</td>
<td>8</td>
<td>BLACKMAIL</td>
<td>26</td>
</tr>
</tbody>
</table>
Innovative uses – new verbs

an increase in the number of matrix verbs: the overall normalized frequency in each of the seven periods followed by the number of new verbs in that period that occur at least once ($f \geq 1$) and twice ($f \geq 2$), as well as a list of the "new" verbs that occur at least twice.

<table>
<thead>
<tr>
<th>Period</th>
<th>Freq</th>
<th>$f \geq 1$</th>
<th>$f \geq 2$</th>
<th>New verbs (frequency $\geq 2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1810s-1830s</td>
<td>1.87</td>
<td>18</td>
<td>5</td>
<td>call 11, bring 5, coax 2, resolve 2, take 2</td>
</tr>
<tr>
<td>1840s-1860s</td>
<td>4.13</td>
<td>67</td>
<td>25</td>
<td>frighten 10, deceive 5, persuade 5, force 5, cheat 4, delude 3, inveigle 3, irritate 3, seduce 3, shame 3, startle 3, surprise 3, talk 3, trick 3, provoke 2, quicken 2, tease 2, terrify 2, lure 2, draw 2, entice 2, flog 2, construe 2, bully 2, cajole 2</td>
</tr>
<tr>
<td>1870s-1890s</td>
<td>6.54</td>
<td>56</td>
<td>16</td>
<td>coerce 9, flatter 4, trap 4, fool 4, rouse 3, spur 2, stimulate 2, train 2, transform 2, gull 2, hoodwink 2, whip 2, worry 2, cow 2, crush 2, argue 2</td>
</tr>
<tr>
<td>1900s-1920s</td>
<td>9.01</td>
<td>92</td>
<td>20</td>
<td>bluff 8, hypnotize 7, kid 5, nag 4, throw 4, shock 3, bamboozle 3, starve 3, stampede 2, trip 2, push 2, anger 2, astonish 2, jar 2, jeer 2, jolly 2, browbeat 2, conjure 2, dupe 2, enchant 2</td>
</tr>
<tr>
<td>1930s-1950s</td>
<td>11.49</td>
<td>99</td>
<td>22</td>
<td>pressure 9, prod 7, maneuver 6, rush 5, threaten 3, needle 3, josh 3, spur 3, squeeze 2, steer 2, surprise 2, rationalize 2, panic 2, plunge 2, blackjack 2, con 2, corner 2, cow 2, interest 2, jockey 2, shake 2</td>
</tr>
<tr>
<td>1960s-1980s</td>
<td>12.28</td>
<td>119</td>
<td>13</td>
<td>embarrass 4, guide 3, kick 3, lock 2, nudge 2, sucker 2, taunt 2, terrorize 2, trigger 2, harry 2, change 2, co-opt 2, drag 2</td>
</tr>
<tr>
<td>1990s-2000s</td>
<td>14.99</td>
<td>93</td>
<td>15</td>
<td>throw 7, draft 3, move 3, whip 3, will 2, transform 2, chase 2, hook 2, invest 2, jolt 2, nurture 2, sink 2, steer 2, stun 2, sway 2</td>
</tr>
</tbody>
</table>
### New verbs from each of the seven periods in COHA

<table>
<thead>
<tr>
<th>Period</th>
<th>Verb</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1810s-1830s</td>
<td>call</td>
<td>this strife of matter becomes a talisman to call into being the sublimest energies of the soul (MAG, 1829)</td>
</tr>
<tr>
<td></td>
<td>bring</td>
<td>exhausting thy purse, to bring into being a pile of timber, and brick and mortar (MAG, 1832)</td>
</tr>
<tr>
<td>1840s-1860s</td>
<td>frighten</td>
<td>You frighten him into deceiving, if you treat him so (FIC, 1852)</td>
</tr>
<tr>
<td></td>
<td>deceive</td>
<td>and then he deceived me into imagining that she wanted time (FIC, 1866)</td>
</tr>
<tr>
<td>1870s-1890s</td>
<td>coerce</td>
<td>and to coerce the Senate into yielding many changes in laws (NEWS, 1876)</td>
</tr>
<tr>
<td></td>
<td>flatter</td>
<td>provided she is not flattered into going beyond her powers (NEWS, 1892)</td>
</tr>
<tr>
<td>1900s-1920s</td>
<td>bluff</td>
<td>in trying to bluff us into letting him keep his flag up (MAG, 1918)</td>
</tr>
<tr>
<td></td>
<td>hypnotize</td>
<td>her calm assumption that she was a very old lady hypnotized them into thinking she was (FIC, 1919)</td>
</tr>
<tr>
<td>1930s-1950s</td>
<td>pressure</td>
<td>they can be persuaded or pressured into buying almost anything (NF, 1944)</td>
</tr>
<tr>
<td></td>
<td>prod</td>
<td>whose wives campaigned vigorously to prod the government into pressing for their release (NEWS, 1950)</td>
</tr>
<tr>
<td>1960s-1980s</td>
<td>embarrass</td>
<td>until worldwide protests embarrassed the government into releasing him (MAG, 1974)</td>
</tr>
<tr>
<td></td>
<td>guide</td>
<td>the … way she had guided him into taking her and had taken pleasure from him (FIC, 1989)</td>
</tr>
<tr>
<td>1990s-2000s</td>
<td>throw</td>
<td>you're almost thrown into defining us by targeting them (MAG, 1990)</td>
</tr>
</tbody>
</table>
Some verbs used only one corpora:

may indicate the innovative uses of the construction:

(32)  
   a. COCA alone: entice, translate, etc  
   b. BNC alone: slip, panick, channel, hoodwink, charm, dragoon, embarrass, pressurize, prompt, pump, etc  
   c. COHA alone: betray, wheedle, inveigle, badger, stamped, etc  

(33)  
   a. The company has **pressurised** the Health Department into allowing its distribution here. (BNC  
   b. She had been **dragooned** into helping with the housework (BNC EVC)  
   c. No doubt that she had **inveigled** Howard into marring her. (COHA 1909 FIC)
The range of verbs is seemingly endless.

Are there any limits to the creativity of speakers? Can we state categorically that something cannot be said?

With more data, we may find more new verbs used in the construction.

Any patterns in terms of syntax or semantics?
Conclusion

Is language production really a poor reflection of language competence as Chomsky really argued?

Corpus linguistics can surely provide the gaps that the competence-based linguistic research has brought to us.

Corpus is a more powerful methodology from the point of view of the scientific method, as it is open to objective verification of results.


Selected References

- Kim, Jong-Bok and Mark Davies. 2014. The INTO-CAUSATIVE Construction in English: A Construction-based Perspective. (Under Review)