Data
Diachronic resources
Diachronic corpora
The New York Times Annotated Corpus

- Contains over 1.8 million articles written and published by the New York Times between January 1, 1987 and June 19, 2007
  - with article metadata

- The corpus provides:
  - over 1.8 million articles
  - over 650,000 article summaries
  - over 1,500,000 articles manually tagged with tags drawn from a normalized indexing vocabulary of people, organizations, locations and topic descriptors
  - over 275,000 algorithmically-tagged articles that have been hand verified
  - Java tools for parsing corpus XML documents

- Text is formatted according to the News Industry Text Format (NITF)

Corpus of Contemporary American English

- Monitor corpus for America English
- Time period: 1990-2009
- 400 million words corpus is evenly divided between spoken, fiction, popular magazines, newspapers, and academic journals.
- The genre balance stays almost exactly the same from year to year

Corpus of Historical American English (COHA)

- Time period: 1810 - 2009
- Genres: newspapers, popular magazines, fiction and nonfiction books
- Size: 406 million words and around 107,000 texts.
- Balanced by genre, sub-genre and domain across decades.

<table>
<thead>
<tr>
<th>Decade</th>
<th>Fiction</th>
<th>Magazines</th>
<th>Newspaper</th>
<th>NF Books</th>
<th>Total</th>
<th>Percent fiction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1810s</td>
<td>641,164</td>
<td>88,316</td>
<td>0</td>
<td>451,542</td>
<td>1,181,022</td>
<td>0.54</td>
</tr>
<tr>
<td>1820s</td>
<td>3,751,204</td>
<td>1,714,789</td>
<td>0</td>
<td>1,461,012</td>
<td>6,927,005</td>
<td>0.54</td>
</tr>
<tr>
<td>1830s</td>
<td>7,590,350</td>
<td>3,145,575</td>
<td>0</td>
<td>3,038,062</td>
<td>13,773,987</td>
<td>0.55</td>
</tr>
<tr>
<td>1840s</td>
<td>8,850,886</td>
<td>3,554,534</td>
<td>0</td>
<td>3,641,434</td>
<td>16,046,854</td>
<td>0.55</td>
</tr>
<tr>
<td>1850s</td>
<td>9,094,346</td>
<td>4,220,558</td>
<td>0</td>
<td>3,178,922</td>
<td>16,493,826</td>
<td>0.55</td>
</tr>
<tr>
<td>1860s</td>
<td>9,450,562</td>
<td>4,437,941</td>
<td>262,198</td>
<td>2,974,401</td>
<td>17,125,102</td>
<td>0.55</td>
</tr>
<tr>
<td>1870s</td>
<td>10,291,968</td>
<td>4,452,192</td>
<td>1,030,560</td>
<td>2,835,440</td>
<td>18,610,160</td>
<td>0.55</td>
</tr>
<tr>
<td>1880s</td>
<td>11,215,065</td>
<td>4,481,568</td>
<td>1,355,456</td>
<td>3,820,766</td>
<td>20,872,855</td>
<td>0.54</td>
</tr>
<tr>
<td>1890s</td>
<td>11,212,219</td>
<td>4,679,486</td>
<td>1,383,948</td>
<td>3,907,730</td>
<td>21,183,383</td>
<td>0.53</td>
</tr>
<tr>
<td>1900s</td>
<td>12,029,439</td>
<td>5,062,650</td>
<td>1,433,576</td>
<td>4,015,567</td>
<td>22,541,232</td>
<td>0.53</td>
</tr>
<tr>
<td>1910s</td>
<td>11,935,701</td>
<td>5,694,710</td>
<td>1,489,942</td>
<td>3,534,899</td>
<td>22,655,252</td>
<td>0.53</td>
</tr>
<tr>
<td>1920s</td>
<td>12,539,681</td>
<td>5,841,678</td>
<td>3,552,699</td>
<td>3,698,353</td>
<td>25,632,411</td>
<td>0.49</td>
</tr>
<tr>
<td>1930s</td>
<td>11,876,996</td>
<td>5,910,095</td>
<td>3,545,527</td>
<td>3,080,629</td>
<td>24,413,247</td>
<td>0.49</td>
</tr>
<tr>
<td>1940s</td>
<td>11,946,743</td>
<td>5,644,216</td>
<td>3,497,509</td>
<td>3,056,010</td>
<td>24,144,478</td>
<td>0.49</td>
</tr>
<tr>
<td>1950s</td>
<td>11,986,437</td>
<td>5,796,823</td>
<td>3,522,545</td>
<td>3,092,375</td>
<td>24,398,180</td>
<td>0.49</td>
</tr>
<tr>
<td>1960s</td>
<td>11,578,880</td>
<td>5,803,276</td>
<td>3,404,244</td>
<td>3,141,582</td>
<td>23,927,982</td>
<td>0.48</td>
</tr>
<tr>
<td>1970s</td>
<td>11,626,911</td>
<td>5,755,537</td>
<td>3,383,924</td>
<td>3,002,933</td>
<td>23,769,305</td>
<td>0.49</td>
</tr>
<tr>
<td>1980s</td>
<td>12,152,603</td>
<td>5,804,320</td>
<td>4,113,254</td>
<td>3,108,775</td>
<td>25,178,952</td>
<td>0.48</td>
</tr>
<tr>
<td>1990s</td>
<td>13,272,162</td>
<td>7,440,305</td>
<td>4,060,570</td>
<td>3,104,303</td>
<td>27,877,340</td>
<td>0.48</td>
</tr>
<tr>
<td>2000s</td>
<td>14,590,078</td>
<td>7,678,830</td>
<td>4,088,704</td>
<td>3,121,839</td>
<td>29,479,451</td>
<td>0.49</td>
</tr>
<tr>
<td>Total</td>
<td>207,633,395</td>
<td>97,207,399</td>
<td>40,124,656</td>
<td>61,266,574</td>
<td>406,232,024</td>
<td>0.51</td>
</tr>
<tr>
<td>-----</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>to_y</td>
<td>86</td>
<td>129</td>
<td>156</td>
<td>178</td>
<td>281</td>
<td>383</td>
</tr>
<tr>
<td>V-&lt;i&gt;ing&lt;/i&gt;</td>
<td>1</td>
<td>8</td>
<td>13</td>
<td>12</td>
<td>22</td>
<td>30</td>
</tr>
<tr>
<td>% V-&lt;i&gt;ing&lt;/i&gt;</td>
<td>0.01</td>
<td>0.06</td>
<td>0.08</td>
<td>0.06</td>
<td>0.07</td>
<td>0.07</td>
</tr>
</tbody>
</table>
Bank of English (Cobuild Corpus)

Bank of English (BoE), also known as the Cobuild Corpus, and (in its most recent incarnation online) as Word Banks Online (http://wordbanks.harpercollins.co.uk).

- Based on the Collins Cobuild dictionaries
- Time period: 1980
- 455 million words by 2005

<table>
<thead>
<tr>
<th>Time period</th>
<th>Fiction</th>
<th>Total</th>
<th>% fiction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1960–79</td>
<td>1,030,000</td>
<td>1,414,000</td>
<td>72.8%</td>
</tr>
<tr>
<td>1980–89</td>
<td>3,087,000</td>
<td>8,792,000</td>
<td>35.1%</td>
</tr>
<tr>
<td>1990–99</td>
<td>6,049,000</td>
<td>20,833,000</td>
<td>29.0%</td>
</tr>
<tr>
<td>2000–4</td>
<td>18,800,000</td>
<td>123,055,000</td>
<td>15.3%</td>
</tr>
</tbody>
</table>
Other English corpora

- ARCHER Corpus (Biber et al., 1994; Yáñez-Bouza 2011): 1700s - 1900s; two million words
- Diachronic Corpus of Present-day Spoken English or DCPSE (Davies, 2009b): 1950s - 1990s british english less than one million words
Google N-gram Viewer

- Search and visualize n-gram statistics from Google Books
- N-gram: sequence of n words
- Google Books digitalizes millions of books
Google N-gram

“Google Books digitalizes millions of books”

1-gram
Google, Books, digitalizes, millions, of, books

2-gram
Google Books, Books digitalizes, digitalizes millions, millions of, of books

3-gram
Google Books digitalizes, Books digitalizes millions, digitalizes millions of, millions of books
Google N-gram: Data Format

ngram TAB year TAB match_count TAB volume_count NEWLINE

Example:
circumvallate   1978   335    91
circumvallate   1979   261    91

Each language for each n-gram has several .gz archives with one n-gram for each line.

Download: https://storage.googleapis.com/books/ngrams/books/datasetsv3.html
CULTUROMICS

A form of computational lexicology that studies human behavior and cultural trends through the quantitative analysis of digitized texts.

Culturomics: Grammar Evolution
Culturomics: Forgot the Old
Culturomics: Forgot the Old
Culturomics: Popularity
Culturomics: Censorship

Marc Chagall (English)
Culturomics: Censorship

Marc Chagall (German)

Nazi censorship
Culturomics: Events

- Russian Flu
- Spanish Flu
- Asian Flu
Google N-gram Viewer: Part-of-Speech
Google N-gram Issues
Overabundance of Scientific Literature
long s in old books looks a lot like a f
Popularity Contests

a book only appears once, whether it’s been read once or millions of times
DUKweb, diachronic word representations from the UK Web Archive corpus
UK Internet Web Archive

- UK Web Archive collects, makes accessible and preserves web resources of scholarly and cultural importance from the UK domain
- JISC UK Web Domain Dataset (1996-2013)
  - resources from the Internet Archive that were hosted on domains ending in ‘.uk’
**DUKweb: source data**

- **ARC format**: used to store "web crawls" as sequences of content blocks
- **WARC format**: is an enhancement of ARC format for supporting metadata, duplicate detection events and more

It is necessary to extract the textual content from HTML pages and discard all other types of content.
From ARC/WARC to WET

**WET format:** contains extracted plaintext from the data stored in ARC/WARC archives
The creation of DUKweb

1. UK Web Archive
2. WARC/ARC to WET
3. Tokenizer
4. Word2Vec
5. Procrustes
6. Change point detection
7. Matrix builder
8. co-occ. matrices
9. Temporal Random Indexing
10. Word Spaces
11. Time series builder
12. Evaluation

Flowchart:
- UK Web Archive feeds into WARC/ARC to WET, which outputs WET.
- WET goes to Tokenizer, which provides Tokens.
- Tokens are fed into Word2Vec, which outputs Embeddings.
- Embeddings are fed into Procrustes.
- Procrustes outputs Tokens.
- Tokens are fed into Matrix builder.
- Matrix builder outputs co-occ. matrices.
- co-occ. matrices are fed into Temporal Random Indexing.
- Temporal Random Indexing outputs Word Spaces.
- Word Spaces are fed into Time series builder.
- Time series builder outputs Evaluation.
- Evaluation can be used for Cumulative or Point-wise analysis.
Co-occurrence matrices

<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>linux</td>
<td>173</td>
</tr>
<tr>
<td>google</td>
<td>255</td>
</tr>
<tr>
<td>xp</td>
<td>237</td>
</tr>
<tr>
<td>manufacturer</td>
<td>64</td>
</tr>
<tr>
<td>job</td>
<td>705</td>
</tr>
<tr>
<td>install</td>
<td>454</td>
</tr>
<tr>
<td>security</td>
<td>137</td>
</tr>
<tr>
<td>cgi</td>
<td>56</td>
</tr>
<tr>
<td>operating</td>
<td>44</td>
</tr>
<tr>
<td>host</td>
<td>255</td>
</tr>
<tr>
<td>performance</td>
<td>69</td>
</tr>
<tr>
<td>sharing</td>
<td>47</td>
</tr>
</tbody>
</table>

One matrix for each year

D-2012_merge_occ.gz
Co-occurrence matrices

Target word

linux swapping 4 google 173 xp 454 manufacturer
237 job 64 install 255 security 137 cgi 47
operating 705 host 69 performance 44 sharing
56...
<table>
<thead>
<tr>
<th>word</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>linux</td>
<td>64</td>
</tr>
<tr>
<td>google</td>
<td>173</td>
</tr>
<tr>
<td>xp</td>
<td>454</td>
</tr>
<tr>
<td>manufacturer</td>
<td>237</td>
</tr>
<tr>
<td>job</td>
<td>454</td>
</tr>
<tr>
<td>install</td>
<td>255</td>
</tr>
<tr>
<td>security</td>
<td>137</td>
</tr>
<tr>
<td>cgi</td>
<td>47</td>
</tr>
<tr>
<td>operating</td>
<td>56</td>
</tr>
<tr>
<td>host</td>
<td>705</td>
</tr>
<tr>
<td>performance</td>
<td>69</td>
</tr>
<tr>
<td>sharing</td>
<td>56</td>
</tr>
</tbody>
</table>
Word Embeddings
Available data

- **Co-occurrence matrices** for each year
- **Word2vec** aligned embeddings
- **Temporal Random Indexing** embeddings
- Download: [https://doi.org/10.23636/1209](https://doi.org/10.23636/1209)

Adam Tsakalidis, Pierpaolo Basile, Marya Bazzi, Mihai Cucuringu & Barbara McGillivray. DUKweb, diachronic word representations from the UK Web Archive corpus. Nature Scientific Data
[https://www.nature.com/articles/s41597-021-01047-x](https://www.nature.com/articles/s41597-021-01047-x)
Data Annotation of Lexical Semantic Changes
Synchronic Word Sense Annotation

Strategies for the annotation of diachronic semantic change are similar to those applied for the annotation of word senses and polysemy at the synchronic level (see e.g. Erk et al 2013)

1) Annotation as classification of word occurrences with sense ID or dictionary senses.
2) Annotation as rating the applicability of dictionary senses on a graded scale to a word occurrence.
3) Annotation as rating the similarity between pairs of usages of the same word, also on a graded scale.
The DURel framework

- A framework that extends synchronic polysemy annotation to diachronic changes in lexical meaning.
- Use of a scale of semantic proximity among sentences (as previously done for synchronic research on polysemy, e.g. Soares da Silva, 1992; Brown, 2008; Erk et al., 2013)

The DURel framework

Identity
Context Variance
Polysemy
Homonymy

4: Identical
3: Closely Related
2: Distantly Related
1: Unrelated

Table 4: Blank (1997)’s continuum of semantic proximity (left) and the DURel relatedness scale derived from it (right).
The DURel framework

- Given sentences of a target word $w$ from two time periods $t_1$ and $t_2$, the semantic relatedness of pair of sentences in each time period is annotated using the relatedness scale above.
- Low mean proximity in a period indicates polysemy or homonymy, while high mean proximity indicates meaning identity.
- The mean proximity values of the two time periods are compared: decrease or increase in the mean relatedness value of $w$ from $t_1$ to $t_2$ indicate respectively innovative or reductive meaning change.
Figure 1: Two-dimensional use spaces (Tuggy, 1993; Zlatev, 2003) in two time periods with a target word $w$ undergoing innovative meaning change. Dots represent uses of $w$. Spatial proximity of two uses means high relatedness.
Annotators judged the relatedness between a use of a word $w$ and a sense definition from a dictionary on the relatedness scale above.

DIACR-Ita 2020 framework

- Annotators were asked to assign each occurrence to one of the meaning of the lemma according to those reported in the Sabatini-Coletti dictionary