Describing our texts: complexity, similarity, readability

Mate Akos

2019 September

MTA TK PTI

For today

1. Describing texts
2. Lexical diversity
3. Collocations
4. Readability and complexity
5. Similarity
6. Exploring keywords

Describing texts

Descriptive statistics of wordcounts, unique types, sentences, etc.

- range (min-max)
- mean or median (depending on the shape of the distribution)
- sum
- lenght (in tokens, sentences, paragraphs, etc.)

Lexical diversity

- Most basic measure: **Type to token ratio**
  - $\frac{\text{total types}}{\text{total tokens}}$
- Sensitive to differences in text lengths
- The larger the text, the smaller the TTR

Quanteda provides a huge amount of lexical diversity indices via the **texstat_lexdiv()** function.

The complete list: `https://quanteda.io/reference/textstat_lexdiv.html`

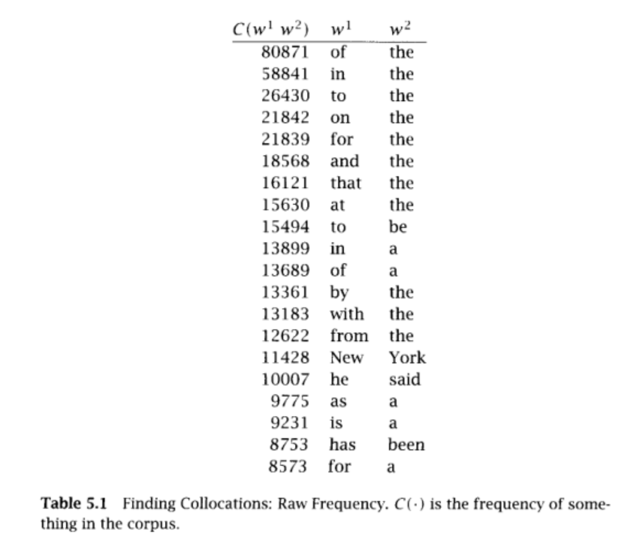
Beyond unigrams: collocations

- Unigrams do not provide the context for our single word tokens. We also have no information if a not/very is preceding our keywords or not.
- A way around it is to identify meaningful collocations in our corpus.
- Collocation ”is an expression consisting of two or more words that correspond to some conventional way of saying things” (Manning and Schütze, FSNLP, 1999: 152)
- E.g.: ethnic cleansing, income inequality, international monetary fund
- Problem: most collocations are noise (I am, on the, etc.)

Identifying collocations

- Frequency
- part-of-speech filtering
- Hypothesis testing: χ^2 test or t-test

Frequency based

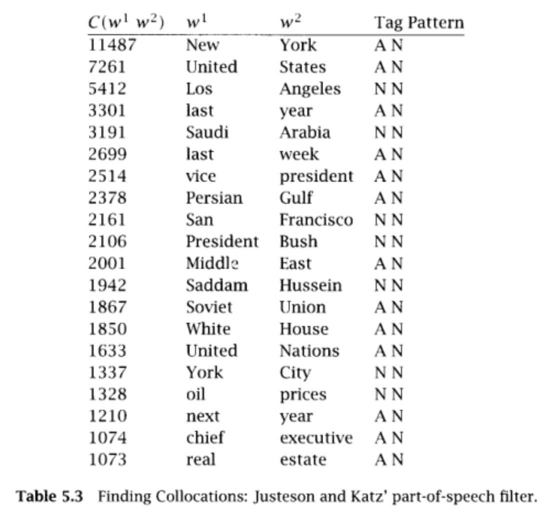


$C(w^1\ w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

Table 5.1 Finding Collocations: Raw Frequency. $C(\cdot)$ is the frequency of something in the corpus.

Figure 1: Manning and Schütze, 1999: 154

Part of speech filtering

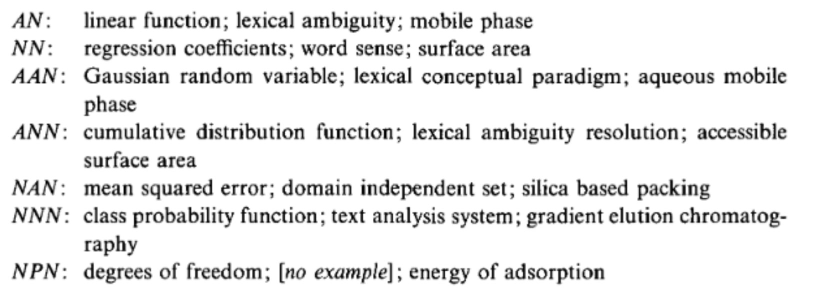


$C(w^1\ w^2)$	w^1	w^2	Tag Pattern
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N
1337	York	City	N N
1328	oil	prices	N N
1210	next	year	A N
1074	chief	executive	A N
1073	real	estate	A N

Table 5.3 Finding Collocations: Justeson and Katz' part-of-speech filter.

Figure 2: Manning and Schütze, 1999: 155

Part of speech filtering



AN:	linear function; lexical ambiguity; mobile phase
NN:	regression coefficients; word sense; surface area
AAN:	Gaussian random variable; lexical conceptual paradigm; aqueous mobile phase
ANN:	cumulative distribution function; lexical ambiguity resolution; accessible surface area
NAN:	mean squared error; domain independent set; silica based packing
NNN:	class probability function; text analysis system; gradient elution chromatography
NPN:	degrees of freedom; [*no example*]; energy of adsorption

Figure 3: Justeson and Katz, 1995: 17

Hypothesis testing

χ^2 is preferred to t-test as we cannot assume normally distributed propabilities (Dunning 1993).

$$X^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

The expected frequencies of $E_{i,j}$ are computed from the marginal probabilities (totals of rows and columns converted into proportions)

Quanteda implementation of collocation detection textstat_collocation()

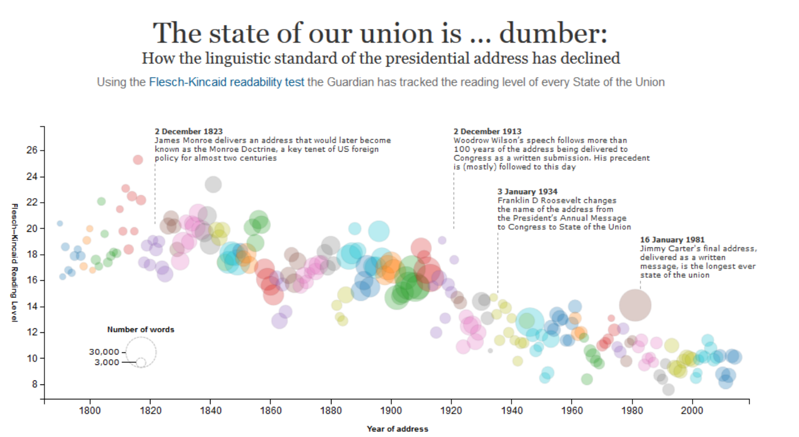
Readability and text complexity

How ”complex” is a given text?

- taking sentence lenght and combination of syllables into account
- Possible application: how complex are various political communications?

Flesch-Kincaid readability index

- $0.39(\frac{\text{total words}}{\text{total senteces}}) + 11.8(\frac{\text{total syllables}}{\text{total words}})$
- Rescaled to US grade levels (1-12)



The state of our union is ... dumber:

How the linguistic standard of the presidential address has declined

Using the Flesch-Kincaid readability test the Guardian has tracked the reading level of every State of the Union

2 December 1823
James Monroe delivers an address that would later become known as the Monroe Doctrine, a key tenet of US foreign policy for almost two centuries

2 December 1913
Woodrow Wilson's speech follows more than 100 years of the address being delivered to Congress as a written submission. His precedent is (mostly) followed to this day

3 January 1934
Franklin D Roosevelt changes the name of the address from the President's Annual Message to Congress to State of the Union

16 January 1981
Jimmy Carter's final address, delivered as a written message, is the longest ever state of the union

Figure 4: https://www.theguardian.com/world/interactive/2013/feb/12/state-of-the-union-reading-level

Similarity (i)

How similar our documents/unit of analysis are? How to measure (dis)similarity? Frequent approaches:

- Euclidean distance
    - $d_2(x_i, x_j) = (\sum_{k=1}^{d}(x_{i,k} - x_{j,k})^2)^{1/2}$
- Minkowski metric
    - $d_p(x_i, x_j) = (\sum_{k=1}^{d}(x_{i,k} - x_{j,k})^p)^{1/p}$
    - when $p = 1$, it is the Manhattan distance, when $p = 2$, it is the Euclidean distance

Similarity (ii)

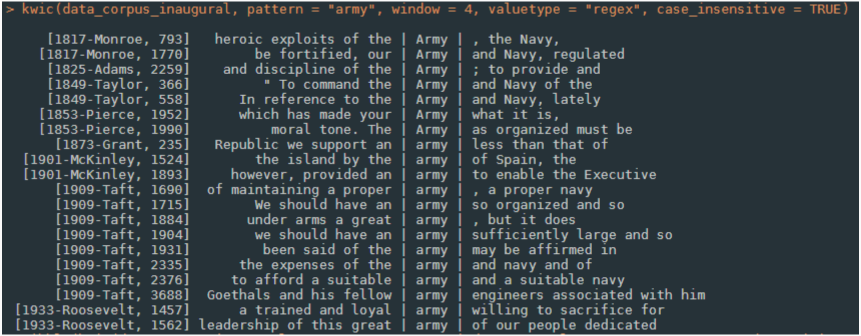
- cosine similarity
  - $cos(A, B) = \frac{A \cdot B}{\|A\| \|B\|}$
  - where $A \cdot B = \sum_{k=1}^{n} a_k b_k$ (dot product) and $\|A\| = \sqrt{\sum_{k=1}^{n} a_k^2}$ (norm or length of the vector)
  - Important feature: **document length does not matter** as it ultimately measures the cosine of the angle between the two vectors
- Jaccard similarity coefficient
  - $J = \frac{|A \cap B|}{|A \cup B|}$
- correlation

quanteda implementation

- textstat_simil()
- textstat_dist()

Exploring our texts: context around keywords

Motivation: what is the **context** that our keyword appears in throughout our corpus (or one document)? The quanteda shorthand is KWIC (key words in context)



```
> kwic(data_corpus_inaugural, pattern = "army", window = 4, valuetype = "regex", case_insensitive = TRUE)

     [1817-Monroe, 793]   heroic exploits of the | Army | , the Navy,
    [1817-Monroe, 1770]       be fortified, our | Army | and Navy, regulated
     [1825-Adams, 2259]    and discipline of the | Army | ; to provide and
     [1849-Taylor, 366]       " To command the | Army | and Navy of the
     [1849-Taylor, 558]      In reference to the | Army | and Navy, lately
    [1853-Pierce, 1952]      which has made your | Army | what it is,
    [1853-Pierce, 1990]          moral tone. The | Army | as organized must be
      [1873-Grant, 235]   Republic we support an | army | less than that of
  [1901-McKinley, 1524]       the island by the | army | of Spain, the
  [1901-McKinley, 1893]     however, provided an | army | to enable the Executive
      [1909-Taft, 1690]  of maintaining a proper | army | , a proper navy
      [1909-Taft, 1715]       We should have an | army | so organized and so
      [1909-Taft, 1884]     under arms a great | army | , but it does
      [1909-Taft, 1904]       we should have an | army | sufficiently large and so
      [1909-Taft, 1931]        been said of the | army | may be affirmed in
      [1909-Taft, 2335]      the expenses of the | army | and navy and of
      [1909-Taft, 2376]     to afford a suitable | army | and a suitable navy
      [1909-Taft, 3688]  Goethals and his fellow | army | engineers associated with him
 [1933-Roosevelt, 1457]       a trained and loyal | army | willing to sacrifice for
 [1933-Roosevelt, 1562] leadership of this great | army | of our people dedicated
```

Figure 5: