DIGILAB WORKSHOP SERIES:
TEXT ANALYSIS 101

TEXT ANALYSIS BASICS

KATIE IRELAND KUIPER
1 APRIL 2021
WHAT IS A TEXT?

- any coherent stretch of language (R. Nordquist)
- a piece of written or spoken material in its primary form
- a text is any object that can be ‘read’; a coherent set of signs that transmits a message (Wikipedia)
- the main body of a piece of language
- the written words in a book, on the internet, etc (Cambridge Dictionary)
WHAT IS TEXT ANALYSIS?

Encompasses the processes involved in analyzing computerized natural language databases (ie. Corpora) in order to:

- Organize/reformat
- Describe
- Understand
- Investigate linguistic and rhetorical devices
- Study specific genres, contexts, and/or discourses
- Other names: text analytics, text mining, stylistics,
TEXT ANALYSIS IS INTERDISCIPLINARY
Image courtesy of Wikimedia commons
Thomas C Mendenhall (1887) wrote one of the first statistical text analyses (Norman; Madigan & Lewis)
• Early studies can be also traced back to automatic translation projects in the 1940s and 50s (Witten).

• Kucera and Nelson Francis’s work on the Brown corpus, a 1 million word database of American English (1967).

• First transcribed corpus of spoken language was created in 1971 by the Montreal French Project, 1 million words (Sankoff & Sankoff 1973).

• Another highly influential study is Kretzschmar et al. 2004’s work on the US Tobacco Industry Documents Corpus.
GETTING YOUR TEXT DATA
DATA SOURCES

- University of Georgia Corpus Server
- Linguistic Data Consortium
- The World Wide Web
- UGA Library Databases
- The Linguistic Atlas Project
- The Hathi Trust Digital Library
- Project Gutenberg
- MONK: Metadata Offer New Knowledge: text analysis suite and public domain TEI texts
- TAPoR: Text Analysis Portal for Research at McMaster Uni
- Martin Weisser’s list of historical corpora
- CLARIN historical corpora
• The corpus server utilizes CQP. (Corpus Query Processor).
• For access to the corpus server, please email linglab@uga.edu.
• On the corpus server, we have access to many awesome databases of language, including:
  • EuroParl
  • SpokenBNC 2014
  • COHA: Corpus Of Historical American English
  • The Brown Corpus
  • The Digital Archive of Southern Speech (DASS)
  • Ancora
ANALYZING YOUR DATA

• Example: proper in British vs American English (tognini Bonelli)
• Example: extracting sentiment from a group of reviews
• Example:
more see imperfection in his face than I now do in his heart.
six or seven and twenty; her face was handsome, her figure tall.
ve-and-thirty; but though his face was not handsome, his count-
ervising rain set full in their face. Chagrined and surprised, the
10
8
12
13
14
15
16
17
18

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18

...
polmineR::corpus()
sAttributes("BNC-BABY")
proper <- cooccurrences("BNC-BABY", query = "proper")
print(proper)
propD <- dispersion("BNC-BABY", query = "proper", s_attribute = "text_author sexe", progress = FALSE)
barplot(height = propD[["count"]], names.arg = propD[["text_author sexe"]], las = 2)

> propD <- dispersion("BNC-BABY", query = "proper", s_attribute = "text_author sexe", progress = FALSE)
> barplot(height = propD[["count"]], names.arg = propD[["text_author sexe"]], las = 2)
METHODS: TEXT ANALYSIS

- distant reading
- natural language processing
- machine learning
- corpus linguistics, corpus-based analysis
METHODS: CORPUS LINGUISTICS

- Frequency analysis
- Analysis of multiword units (ngrams)
- Collocation analysis
- Keyword analysis
METHODS: NLP & MACHINE LEARNING

- document classification
- topic modeling
- sentiment analysis
- named entity recognition
TEXT ANALYSIS IS SCALABLE.
• **spacy**: pos tagging, tokenization, dependency parsing, etc. Check out this [tutorial](#) for more about NLP with spacy

• **CoreNLP**: lemmatization, pos tagging, tokenization, named entity recognition

• **NLTK**: Natural Language ToolKit; contains over 50 corpora, includes options for tokenization, tagging, parsing, document classification

• **Gensim**: useful for various types of topic modeling

• **PyNLPI**: open-source NLP library; great for of tasks ranging from building simplistic models and extraction of n-grams and frequency lists, with support for complex data types and algorithms

• **Pattern**: useful for web-crawling (webscraping) for creating your own corpora; includes options for tokenizing, pos tagging, etc

• **Polyglot**: very useful library for other languages than English

• **TextBlob**: includes options for pos-tagging, noun phrase extraction, classification, translation and sentiment analysis
• **Tidytext**: helpful for data formatting and visualization; works well with other packages in the Tidyverse (Silge & Robinson 2016)

• **Textmining/tm**: includes options for data processing, metadata management, and creation of term-document matrices (Feinerer 2020; Feinerer et al. 2008)

• **Syuzhet**: package created specifically for sentiment analysis by Jockers

• **Text2vec**: dtm, vectorizing data, supports topic modeling and collocational analysis, too

• **StringR**: supports regex, pattern matching, useful for string manipulation

• **spacyR**: NLP package originally created for Python; useful for tokenization and works well with quanteda and tidytext

• **Quanteda**: incredibly useful package; includes preprocessing abilities, dtm function, as well as statistical analyses options like document classification and topic modeling

• **Ggplot2**: great way to visualize your data
RESOURCES AT UGA

- Corpus Server
- Upcoming Courses
- Digilab Resources
- Data Office Hours
COURSES AT UGA

• This Fall 2021:

• Natural Language Processing: LING 4570/6570

• Style: ENGL/LING 4826/6826

• American English: ENGL/LING 4010/6010

• Note: These all count toward the Digital Humanities Undergraduate certificate!
• **AntConc**: A free corpus analysis toolkit for concordancing and corpus-based methods

• **Voyant Tools**: web-based text reading and analysis environment

• **Google Books Ngram Viewer**: online search engine that charts the frequencies of any set of comma-delimited search strings

• **Wordseer**: text analysis environment that combines visualization, information retrieval, and nlp methods

• **Tapor**: web-based set of text analysis tools
• **TextArc:** A visual representation of a text.
• **MALLET:** Maps patterns across texts with various tools.
• **Perl:** was originally created to be a general purpose programming language to help with reports; includes many excellent text-specific functions; supports powerful regular expressions, string processing, and parsing
• **SketchEngine:** text mining app based out of the EU; includes options for your own corpora and includes 500+ other corpora
• **http://corsis.sourceforge.net/**: open source corpus software written in C
• **ICECUP 3.1, Fuzzy Tree Fragments:** based at UCL, set of corpus tools for parsed corpora like **ICE-GB** and **DCPSE**
DATA OFFICE HOURS

CONSULTATIONS FOR DATA CLEANING, STRUCTURING, AND VISUALIZING

Whether just starting your work, or trying to make sense of your research, schedule an appointment for our Data Office Hours and bring your data (text, archival information, numerical data, etc.) for advice and guidance on your project. Expertise in corpus linguistics, Excel, and R, among other tools for data structuring and visualization.

TUESDAYS • 4:00-5:00
WEDNESDAYS • 2:00-3:00

To schedule an appointment visit:
DIGI.UGA.EDU/RESOURCES
RECOMMENDED RESOURCES

• Brezina’s *Statistics in Corpus Linguistics*

• Evert’s work on collocations and corpus methods

• University of Lancaster Corpus for Schools

• *Natural Language Processing with Python* by Bird et al.; Na-Rae Han’s python tutorials

• Silge and Robinson’s Text Mining with R

• University of Birmingham, UK Centre for Corpus Research

• HELSINKI’s VARIENG Center for Research
COMING UP NEXT…

8 April: Text Analysis for Literature and Beyond

15 April: Creating your own Social Media Corpus

22 April: Text Analysis Applications: Social Media
IN PREPARATION FOR NEXT WEEK

Download and install: R and R Studio
THANKS FOR LISTENING!

KATHERINE.KUIPER25@UGA.EDU

PLEASE FILL OUT THIS Survey.


• Evert, Stefan. 2003. The CQP Query Language Tutorial.


• Feinerer et al. 2008.


• Firth, JR. 1957. Papers in Linguistics. London: OUP.


• Han, Na-Rae. Python 3 tutorials. http://www.pitt.edu/~naraehan/python3/.

• HathiTrust. https://www.hathitrust.org/about.


• https://monkeylearn.com/text-analysis/
• Millot, Thomas. Photo. Unsplash
• Project Gutenberg. https://www.gutenberg.org