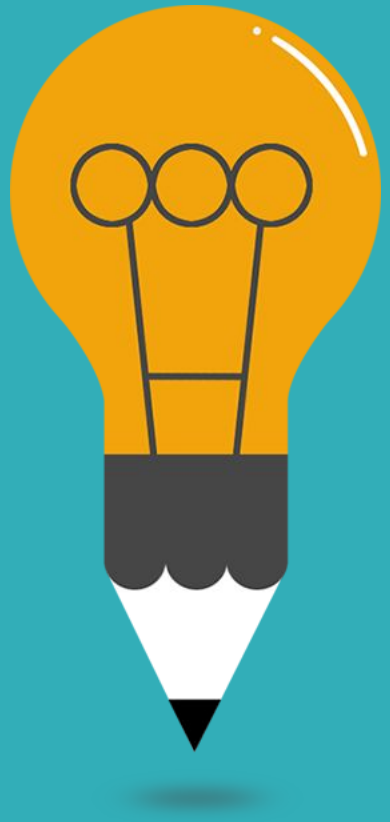
Link to slides: https://tinyurl.com/SIGDOC20

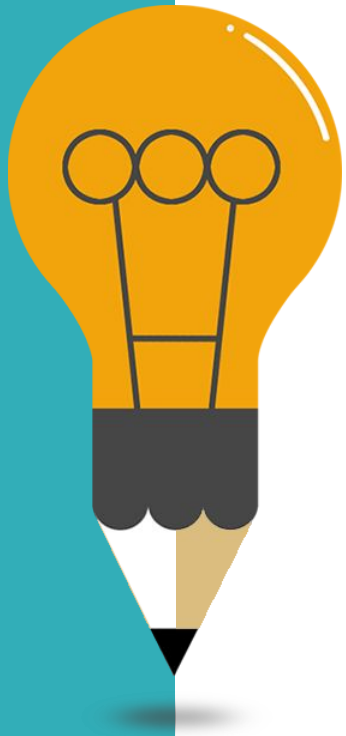


# A Critical Look at Computational Methods in Corpus Analysis

SIGDOC | October 5-9, 2020

Yeqing Kong, Nupoor Ranade, Jianfen Chen, Missy Hannah

# Agenda



**01** **Introduction**
Corpus analysis

**02** **Computational Methods**
Four commonly used approaches

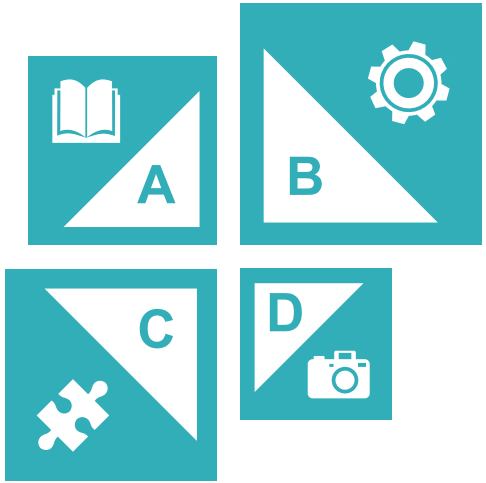
**03** **Factors to Consider**
When choosing a method

**04** **Conclusion**
Q & A

# What is a corpus?

**Machine-readable**

**Sampled**




**Authentic**

**Representative**

A corpus is a collection of naturally-occurring language text, chosen to characterize a state of variety of language (Sinclair, 1991).

# Research objectives

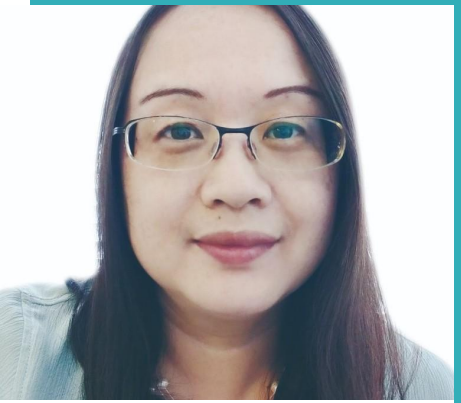
01

Showcase four commonly used approaches in corpus analysis

02

Provide guidance to researchers on the factors to be considered while implementing each of these methods

# Panelists



**Jianfen Chen**

Purdue University

Verbal Data Analysis

**Yeqing Kong**

NC State University

Corpus Linguistics

**Nupoor Ranade**

NC State University

Programming

**Missy Hannah**

NC State University

Sentiment Analysis

# Verbal Data Analysis

Jianfen Chen | Purdue University

# Verbal Data Analysis as Mixed-Method

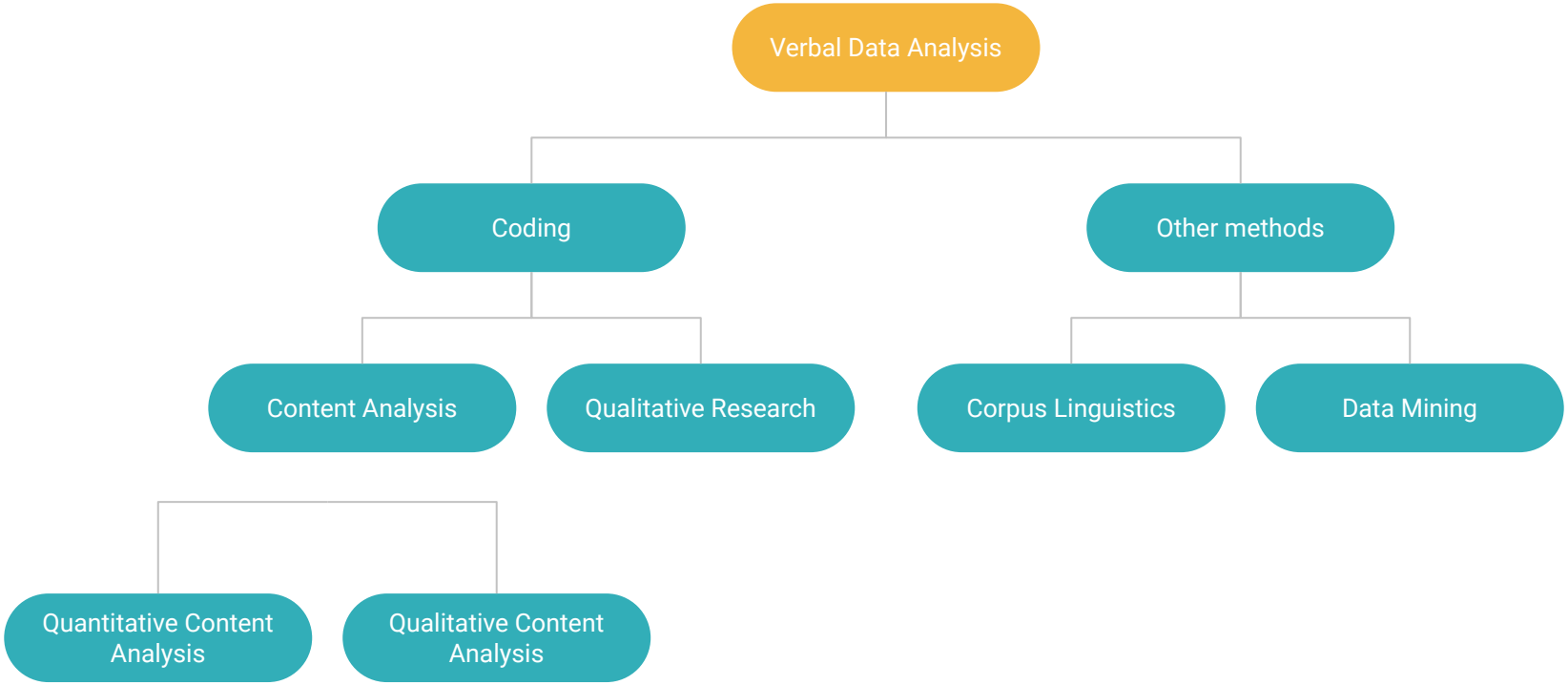



Figure 1.1: Taxonomy of approaches to verbal data analysis from Coding Stream of Language (Geisler & Swarts, 2019)

# Key Skills and Tools for VDA

| Skills | Tools |
| --- | --- |
| Segmenting data | Excel |
| Coding | MAXQDA |
| Examining patterns of distribution | Dedoose |

# Case study

**A Website-based Comparison of Corporate Introductions between Chinese and American Public Companies**

- **Data source**: Forbes website
- **Sampling method**: Criterion sampling
- **Data size**:
  - 10 Chinese companies vs. 10 American companies
  - 8,600 words in total

# Key Steps in VDA

**Segmenting Data** → **Developing Coding Scheme** → **Coding and Analyzing Data**

Choose the language unit you want to use to segment the data based on your research questions

Develop the codes, dimension, and coding scheme you want to use to code the data

Write about the recurring patterns shown in your coded data and analyze them to answer your research questions

# Corpus Linguistics

Yeqing kong | NC State University

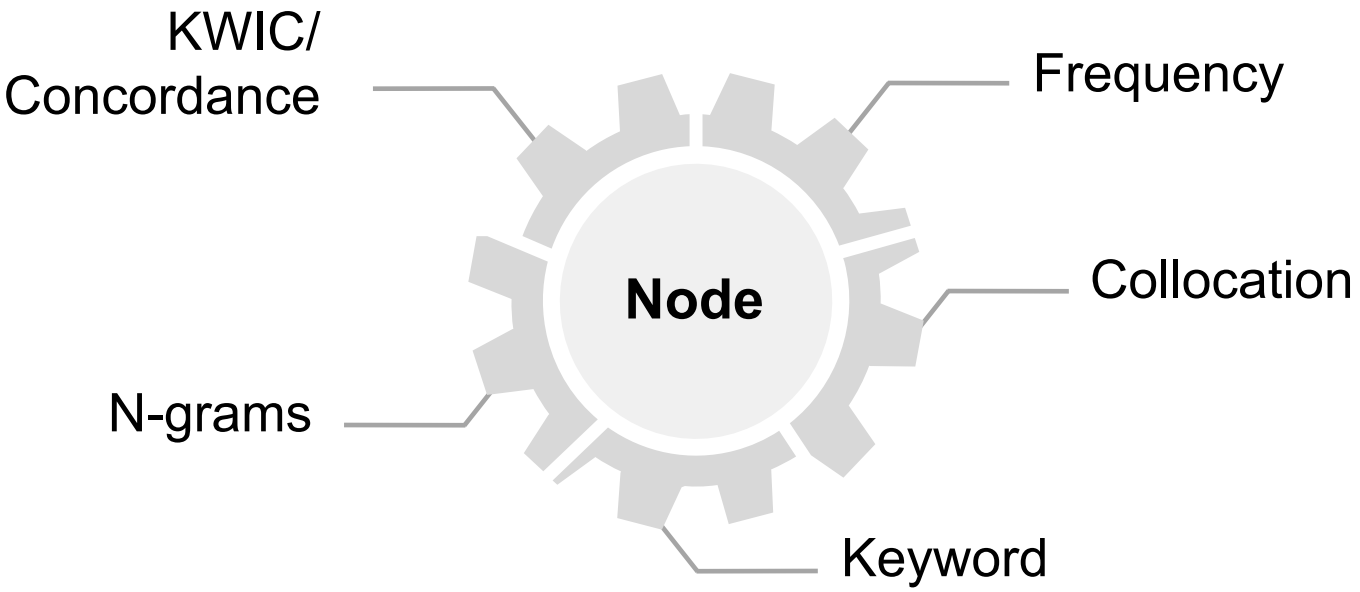
# What is corpus linguistics?

> Corpus linguistics is the study of large collections of texts, often carefully sampled and encoded electronically to be representative of a particular kind of naturally occurring language.

**Paul Baker**

Litosseliti, L. (Ed.). (2018). Research methods in linguistics. Bloomsbury Publishing.

# Key corpus linguistics concepts

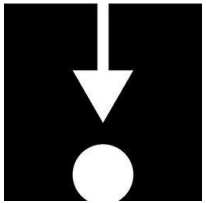


# Corpus software packages

AntConc


WordSmith

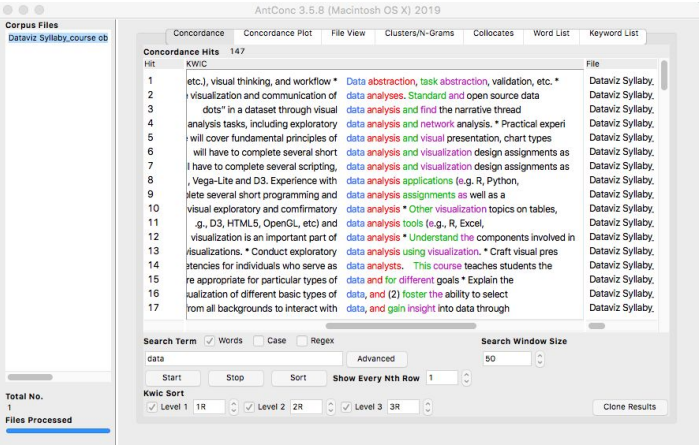

Wmatrix

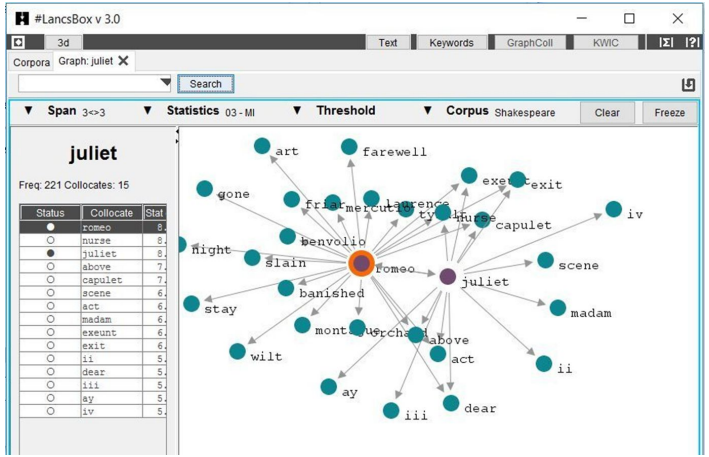


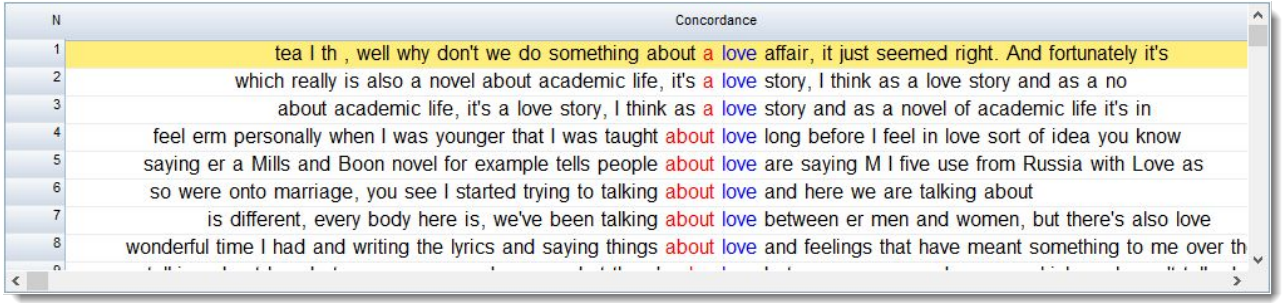
LancsBox


Sketch Engine

# Corpus software interface



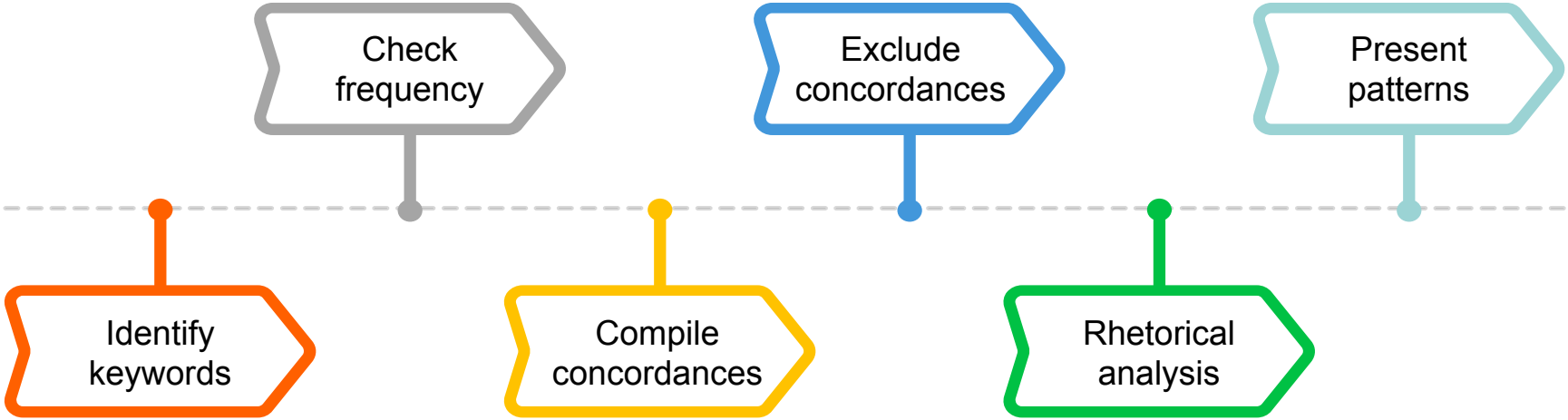
| N | Concordance |
|---|---|
| 1 | tea I th , well why don't we do something about a love affair, it just seemed right. And fortunately it's |
| 2 | which really is also a novel about academic life, it's a love story, I think as a love story and as a no |
| 3 | about academic life, it's a love story, I think as a love story and as a novel of academic life it's in |
| 4 | feel erm personally when I was younger that I was taught about love long before I feel in love sort of idea you know |
| 5 | saying er a Mills and Boon novel for example tells people about love are saying M I five use from Russia with Love as |
| 6 | so were onto marriage, you see I started trying to talking about love and here we are talking about |
| 7 | is different, every body here is, we've been talking about love between er men and women, but there's also love |
| 8 | wonderful time I had and writing the lyrics and saying things about love and feelings that have meant something to me over th |

# Case study

**A Corpus-assisted analysis of global cultural flows of AI in news discourse**

- **Data**: news articles in the U.S. and China
- **Tool**: AntConc
- **Method**:
  - built concordances with keywords
  - examine the concordances and extract patterns

# Corpus-assisted Discourse Analysis Workflow



# Examine “Technology” as a keyword

Corpus Files
CD_cleaned data_39.txt
NYT_cleaned data_59.txt

Concordance | Concordance Plot | File View | Clusters/N-Grams | Collocates | Word List | Keyword List

Concordance Hits 340

| Hit | KWIC | | File |
|---|---|---|---|
| 1 | lag the | technologies." A memorandum | NYT_cleaned da |
| 2 | and information | technology. A visitor | CD_cleaned dat |
| 3 | voice-recognition | technology. According to | CD_cleaned dat |
| 4 | at bay. | Technological advancements are | CD_cleaned dat |
| 5 | of other | technological advances decades | NYT_cleaned da |
| 6 | of recent | technological advances in | NYT_cleaned da |
| 7 | to link | technology advances more | NYT_cleaned da |
| 8 | and Information | Technology. AI has | CD_cleaned dat |
| 9 | and Information | Technology. AI has | CD_cleaned dat |
| 10 | design complex | technology. Airplanes don' | NYT_cleaned da |
| 11 | with computer | technology, and a | CD_cleaned dat |
| 12 | with computer | technology, and a | CD_cleaned dat |
| 13 | with computer | technology, and a | CD_cleaned dat |
| 14 | with computer | technology, and a | CD_cleaned dat |
| 15 | with computer | technology, and a | CD_cleaned dat |
| 16 | featuring the | technology and advancement | CD_cleaned dat |
| 17 | featuring the | technology and advancement | CD_cleaned dat |
| 18 | with advanced | technology and also | NYT_cleaned da |
| 19 | theory, method, | technology and application | CD_cleaned dat |
| 20 | ficial intelligence | technology and application | CD_cleaned dat |
| 21 | counterculture on | technology and computing. | NYT_cleaned da |
| 22 | , and other | technologies, and develop | CD_cleaned dat |
| 23 | , and other | technologies, and develop | CD_cleaned dat |
| 24 | the basic | technology and how, | NYT_cleaned da |
| 25 | global AI | technology and industries | CD_cleaned dat |
| 26 | Guangzhou Science | Technology and Innovation | CD_cleaned dat |
| 27 | material science | technology, and it | NYT_cleaned da |
| 28 | 's AI | technology and its | CD_cleaned dat |
| 29 | 2025 in both | technology and its | CD_cleaned dat |
| 30 | Institute of | Technology and John | NYT_cleaned da |

Search Term ☑ Words ☐ Case ☐ Regex
technology Advanced
Start Stop Sort Show Every Nth Row 1
Kwic Sort
☑ Level 1 1R ☑ Level 2 2R ☑ Level 3 3R

Search Window Size
20

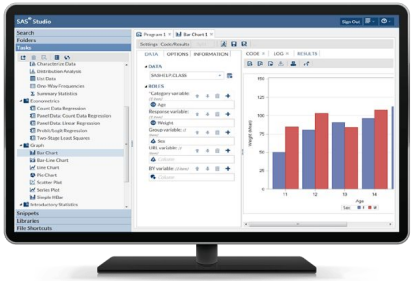
Clone Results

Total No.
2
Files Processed

# Programming

Nupoor Ranade | NC State University

# Programming software for corpus analysis

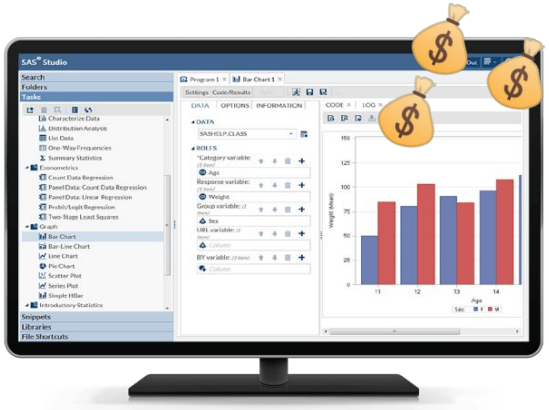


Programming using
Commercial
Software


Programming using
Open source
languages

# Commercial software

Programming software generally provide a graphical point-and-click user interface for non-technical users.

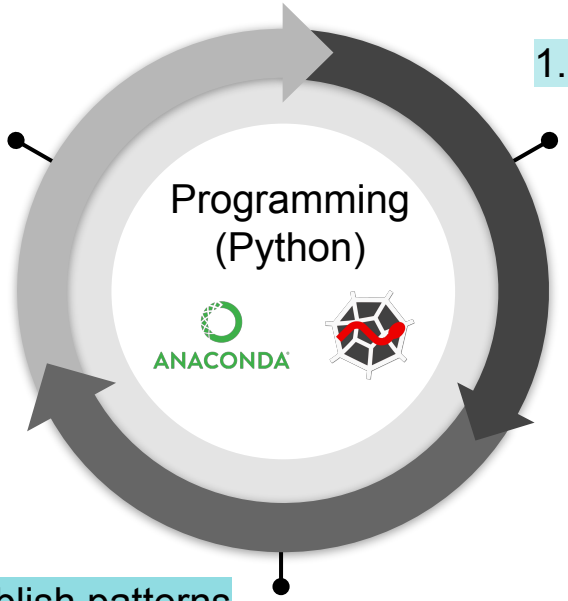
# Open source languages

Programming languages let users write custom programs suitable for their data set as well as need.


# Case Study: Corpus analysis using Python to locate String Citations

3. Code patterns to retrieve results

   The code for parsing combined with the regexes was run to identify all the string citations in the corpus




1. Determine a dataset and clean it for parsing

   Journal articles (777) in PDF format were converted to .TXT and cleaned up to see if formatting was accurate

2. Establish patterns

   Patterns for string citations were established by manual analysis for each of the journal types and regexes were created

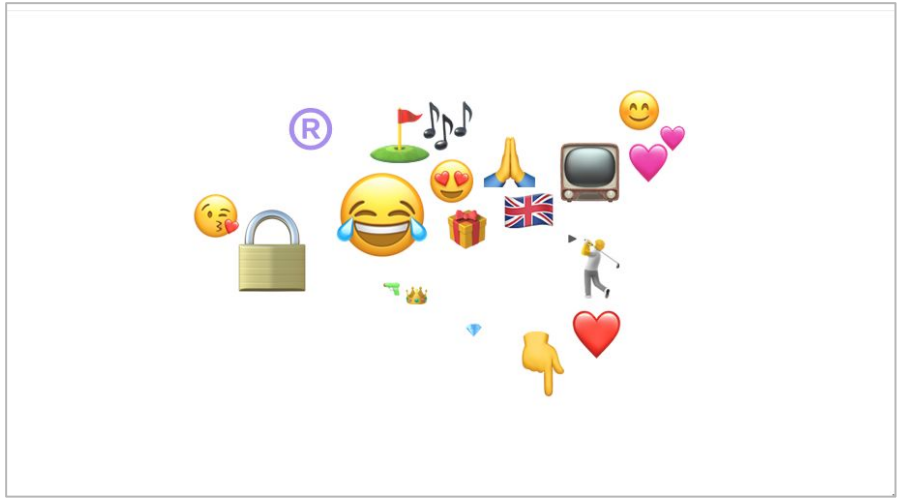
# Sentiment Analysis

Missy Hannah | NC State University

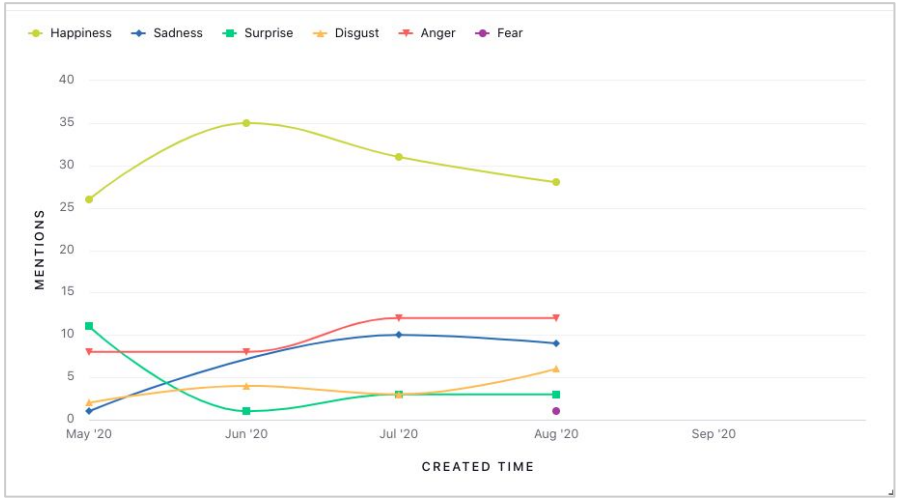
# What is Sentiment analysis?

Sentiment analysis: “the use of text analytics and computational linguistics to study and quantify affective states of participants.”

# Types of sentiment analysis



Different Emoticons Used around Brand/Topic




Emotional Trends for your Brand/Topic

# Software programs for sentiment analysis

LIWC

Academic version available for purchase.

SentiWordNet

Open-source and available on Github.

SentiStrength

Free for academic research and available on their website.

BRAND24

Free, point & click only!

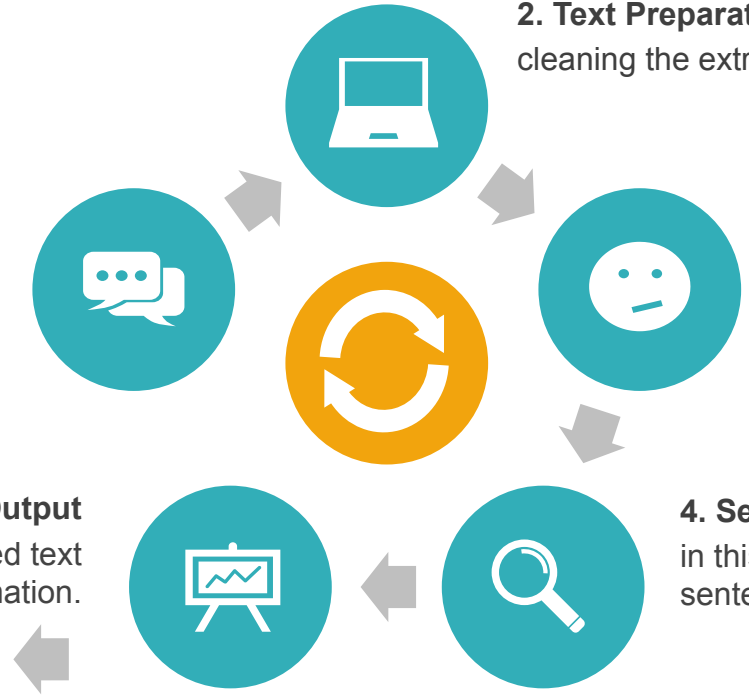
Repustate

17 languages, no coding, and part-of-speech tagging.

ParallelDots

14 languages, no coding, free lifetime version (but watered down).

# Sentiment Analysis Workflow



**1. Data Collection**

collect data from user generated content contained in blogs, forums, social networks

**2. Text Preparation**

cleaning the extracted data before analysis

**3. Sentiment detection**

the extracted sentences of the reviews and opinions are examined.

**4. Sentiment Classification**

in this step, subjective sentences are classified

**5. Presentation of Output**

convert the unstructured text into meaningful information.

# Factors to consider

when choosing a suitable approach

Datasets

Skills

Tools

# Factor 1: Datasets

| Methods | Dataset Type | Data Size |
|---|---|---|
| **Verbal data analysis** | Good for smaller size of data sets (can be used for bigger using automation codes) |  |
| **Corpus software** | Good for varying size of data sets; works well for multiple corpora |  |
| **Programming** | Good for large data sets |  |
| **Sentiment analysis** | Good for large or big data sets |  |

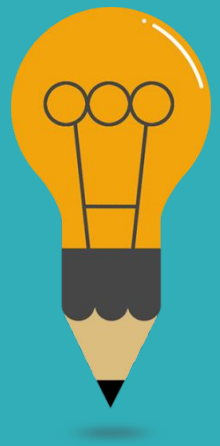
# Factor 2: Skills

| Methods | Knowledge needed | Learning curve |
|---|---|---|
| **Verbal data analysis** | discourse analysis, grounded theory, text segment and coding skills | |
| **Corpus software** | basic corpus linguistic concepts & how to apply them in context | |
| **Programming** | syntax, variables, data structures, algorithms, and debugging | |
| **Sentiment analysis** | sentiment detection, sentiment classification, presentation of output, validation | |

# Factor 3: Tools

| Methods | Tools needed | Usability |
|---|---|---|
| **Verbal data analysis** | Microsoft Excel,  MAXQDA, or Dedoose, |  |
| **Corpus software** | AntConc, Sketch Engine, WordSmith, Wmatrix, or LancsBox |  |
| **Programming** | Python, or R supplemented by a text editor |  |
| **Sentiment analysis** | Emoticons, SentiStrengh, Happiness Index, and Sentiment 140 |  |

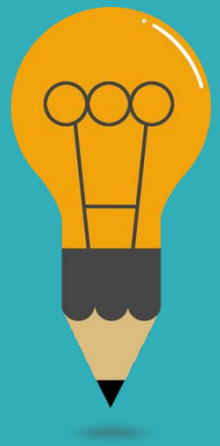
# Conclusion



1. Methods cannot be fully determined in advance.
2. Careful qualitative analysis based on contextual knowledge is also important.
3. Mixed-methods approach and triangulation are beneficial.

# References


- McEnery, T., & Wilson, A. (2001). Corpus linguistics: An introduction. Edinburgh University Press.
- Wiedemann, G. (2013). Opening up to big data: Computer-assisted analysis of textual data in social sciences. Historical Social Research/Historische Sozialforschung, 332-357.
- Geisler, C., & Swarts, J. (2019). Coding streams of language techniques for the systematic coding of text, talk, and other verbal data. Boulder: University Press of Colorado.
- Kennedy, G. (2014). An introduction to corpus linguistics. Routledge.
- Baker, P. (2006). Using corpora in discourse analysis. A&C Black.
- Sabatovych, I. (2019). Do social media create revolutions? Using Twitter sentiment analysis for predicting the Maidan Revolution in Ukraine. Global Media and Communication, 15(3), 275-283.
- Alessia, D., Ferri, F., Grifoni, P., & Guzzo, T. (2015). Approaches, tools and applications for sentiment analysis implementation. International Journal of Computer Applications, 125(3).
- Geisler, C. (2018). Coding for language complexity: The interplay among methodological commitments, tools, and workflow in writing research. Written Communication, 35(2), 215-249.
- DeLyser, D., & Sui, D. (2013). Crossing the qualitative-quantitative divide II: Inventive approaches to big data, mobile methods, and rhythmanalysis. Progress in Human Geography, 37(2), 293-305.
- Egbert, J., & Baker, P. (Eds.). (2019). Using corpus methods to triangulate linguistic analysis. Routledge.

ykong2@ncsu.edu
nupoor.ranade@ncsu.edu
chen3421@purdue.edu
melissa_hannah@ncsu.edu

# Thank you

@YeqingKong
@nupoorwriting
@sugejianfen
@heylittlemissy_