

Link to Slides:

<https://tinyurl.com/y7cqb6q8>

Link to Transcript:

<https://tinyurl.com/y9mx92kz>

Custom OR Off-the-shelf Software: Making the Right Choice for Corpus Analysis



IEEE ProComm | July 20-21, 2020
Nupoor Ranade & Yeqing Kong
NC State University



AGENDA

- Introduction
- Corpus Analysis Approaches
 - Programming
 - Corpus Software
- Comparison: Programming vs. Corpus Software
- Conclusion



What is a **corpus**?



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



Two approaches & RQs

Programming

OR

**Corpus
Software**

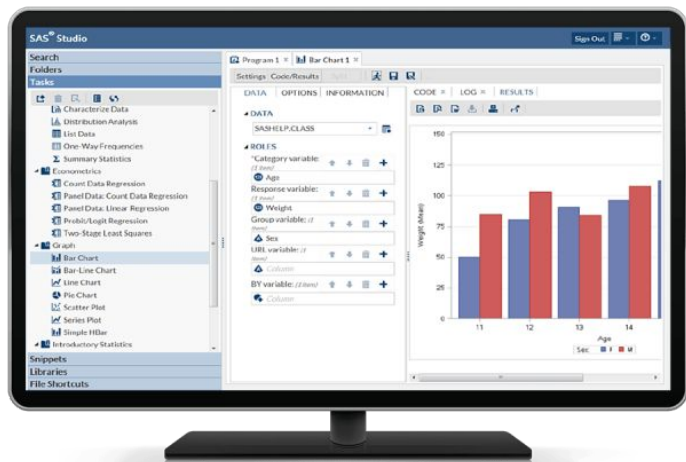
RQ1: What are the affordances and limitations of each approach?

RQ2: How should researchers make appropriate choices?



Programming

Commercial Software



Open-source languages

The screenshot shows a Python script in a code editor and its output in a console window. The script is named 'parser.py' and contains the following code:

```
1 # Requirements
2 list 1. Look for string citations
3 Conditions for inclusion in data set:
4 1. If TOL, JMC or TC, citations will look like f xxx, ###, xxx, ###, ...
5 2. If IEEE, citations will look like [1], [1], ... or [1]-[1]
6 3. Find frequency of authors from author_names in that data set
7 4. Find frequency of authors from author_names_contemporary in that data set
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
```

The console window shows the output of the script, which is a list of authors and their frequency of occurrence. The output is as follows:

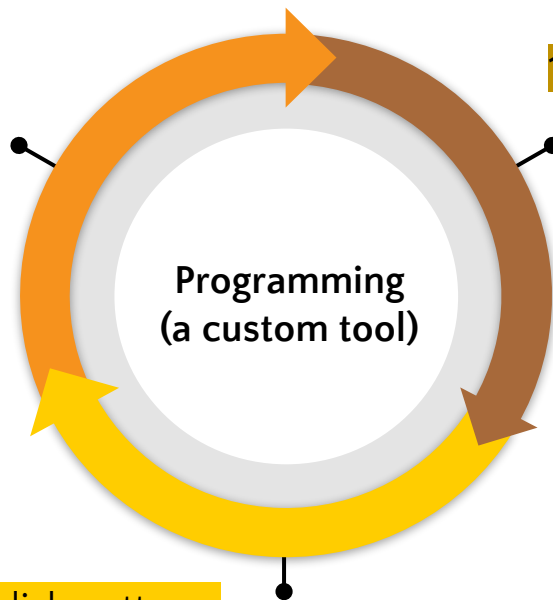
id	Type	Size	Value
0	str	1	[De] Gaudio, Franzato, & de Oliveira, 2016; Rose & Cardinal, 2013
1	str	1	[e.g., Lathrop, Auermuller, Haag, & In, 2012], but researchers typically ...
2	str	1	[e.g., Morrow, Lees, Brown, & Feyer, 2015; Roth, Hart, Mead, & Quinn, ...
3	str	1	[Morrow et al., 2015; Roth et al., 2017]. The degree of synchronic (Mo ...
4	str	1	[Rose & Cardinal, 2016; Spinuzzi, 2005]



Programming basics for corpus analysis

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

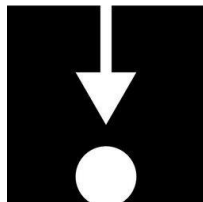
Patterns for string citations were established by manual analysis for each of the journal types and regexes were created



Corpus software packages



AntConc



LancsBox



WordSmith



Wmatrix



Sketch Engine



Corpus software interface

AntConc 3.5.8 (Macintosh OS X) 2019

Corpus Files
Dataviz Syllaby_course ob

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

Concordance Hits 147

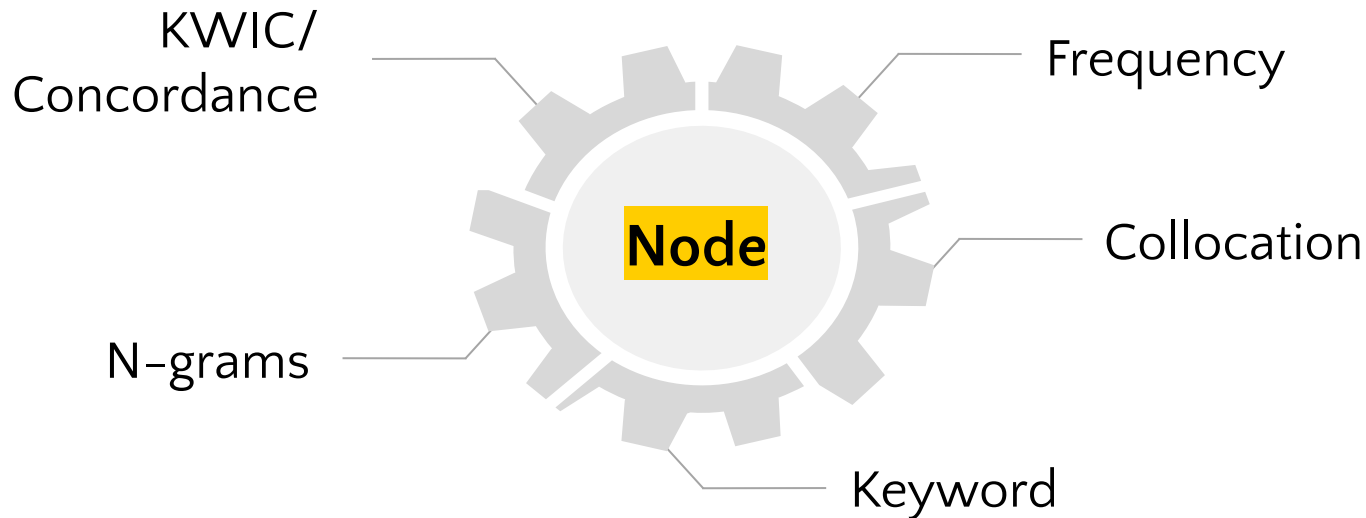
Hit	KWIC	File
1	etc.), visual thinking, and workflow * Data abstraction, task abstraction, validation, etc. *	Dataviz Syllaby.
2	visualization and communication of data analyses. Standard and open source data	Dataviz Syllaby.
3	dots" in a dataset through visual data analysis and find the narrative thread	Dataviz Syllaby.
4	analysis tasks, including exploratory data analysis and network analysis. * Practical experi	Dataviz Syllaby.
5	will cover fundamental principles of data analysis and visual presentation, chart types	Dataviz Syllaby.
6	will have to complete several short data analysis and visualization design assignments as	Dataviz Syllaby.
7	will have to complete several scripting, data analysis and visualization design assignments as	Dataviz Syllaby.
8	, Vega-Lite and D3. Experience with data analysis applications (e.g. R, Python,	Dataviz Syllaby.
9	plete several short programming and data analysis assignments as well as a	Dataviz Syllaby.
10	visual exploratory and confirmatory data analysis * Other visualization topics on tables,	Dataviz Syllaby.
11	.g., D3, HTML5, OpenGL, etc) and data analysis tools (e.g., R, Excel,	Dataviz Syllaby.
12	visualization is an important part of data analysis * Understand the components involved in	Dataviz Syllaby.
13	visualizations. * Conduct exploratory data analysis using visualization. * Craft visual pres	Dataviz Syllaby.
14	encies for individuals who serve as data analysts. This course teaches students the	Dataviz Syllaby.
15	re appropriate for particular types of data and for different goals * Explain the	Dataviz Syllaby.
16	ualization of different basic types of data, and (2) foster the ability to select	Dataviz Syllaby.
17	from all backgrounds to interact with data, and gain insight into data through	Dataviz Syllaby.

Search Term ☒ Words ☐ Case ☐ Regex
data Advanced Search Window Size 50
Start Stop Sort Show Every Nth Row 1
Kwic Sort
☒ Level 1 1R ☒ Level 2 2R ☒ Level 3 3R
Clone Results

Total No.
1
Files Processed

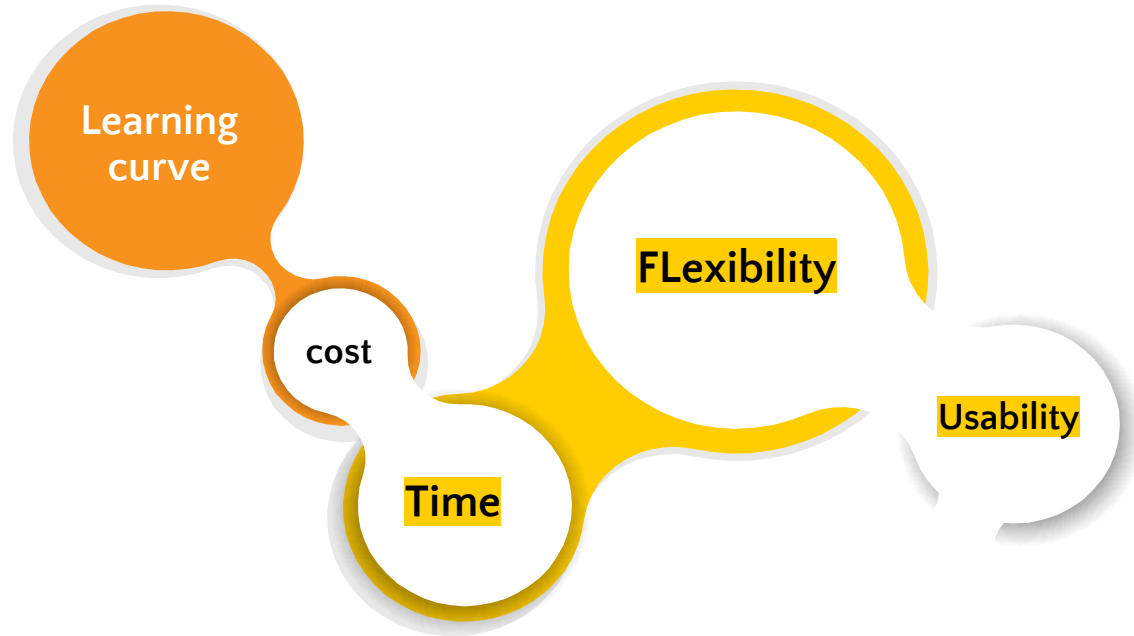


Key corpus linguistics concepts





Comparison: programming vs. software





Differences: learning curve

	Knowledge needed
Programming	syntax, variables, data structures, algorithms, and debugging
Corpus software	basic corpus linguistic concepts & how to apply them in context



Differences: cost

	Free	Subscription needed
Programming	Python, R	SAS, MATLAB
Corpus software	AntConc, LancsBox	Wmatrix, Sketch Engine, WordSmith tools

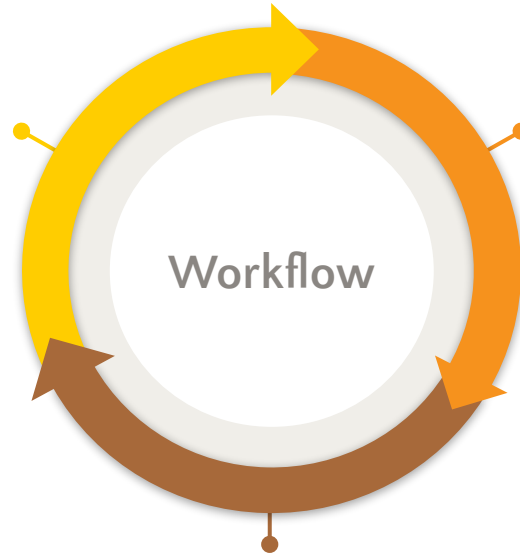


Differences: time/labor

1. Pre-processing data



2. Computational operations

3. Contextual analysis







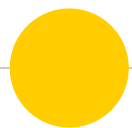
Differences: flexibility

	Ratings
Programming	
Corpus software	



Differences: usability/accessibility

	Ratings
Programming	
Corpus software	



Conclusion

- ◉ Each approach has its own merits and demerits when used for computational social science research.
- ◉ Analyzing the differences will help scholars make informed decisions about choosing appropriate approaches in corpus research.



References

- ◉ P. Baker, *Using Corpora in Discourse Analysis*. A&C Black, 2006.
- ◉ O. Mason, “Developing software for corpus research,” *Int. J. English Studies*, vol. 8, no. 1, pp. 141–156, 2008.
- ◉ W. G. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. (2nd. ed.). O’Reilly Media, Inc, 2017.
- ◉ F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” *J. Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- ◉ W. G. McKinney, “Data structures for statistical computing in Python,” in *Proc. 9th Python in Science Conf. (SCIPY)*, 2010, pp. 51–56.
- ◉ A. Kilgariff et al., “The Sketch Engine: ten years on,” *Lexicography*, vol. 1, no. 1, pp. 7–36, 2014.
- ◉ M. Scott. WordSmith Tools. (version 8). [Online]. Available: <https://lexically.net/downloads/version8/HTML/index.html>
- ◉ L. Anthony. AntConc. (version 3.5.8). Waseda University. [Online]. Available: <https://www.laurenceanthony.net/software>
- ◉ T. McEnery et al., *Corpus-based Language Studies: An Advanced Resource Book*. Taylor & Francis, 2006.
- ◉ V. Brezina et al. #LancsBox v. 4.x. [Online]. Available: <http://corpora.lancs.ac.uk/lancsbox>
- ◉ P. Rayson, “From key words to key semantic domains,” *Int. J. Corpus Linguistics*, vol. 13, no. 4, pp. 519–549, 2008.



Thanks!

Any questions ?

You can find us at

- nupoor.ranade@ncsu.edu | Twitter: @nupoorwriting
- ykong2@ncsu.edu | Twitter: @YeqingKong