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 of variety of language (Sinclair, 1991).

Two approaches & RQs

Programming

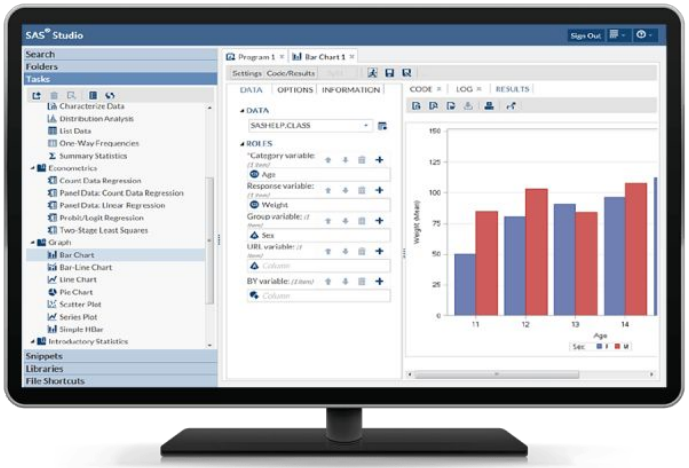
OR

Corpus Software

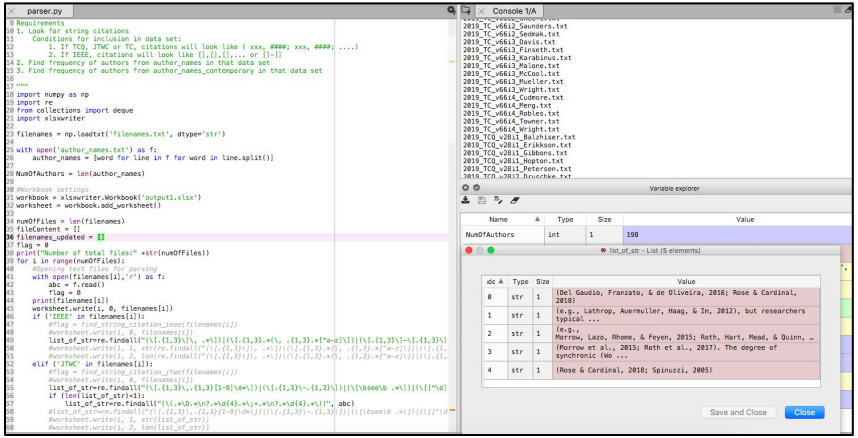
RQ1: What are the affordances and limitations of each approach?
RQ2: How should researchers make appropriate choices?

Programming

Commercial Software



Open-source languages



Programming basics for corpus analysis

1. Determine a dataset and clean it for parsing

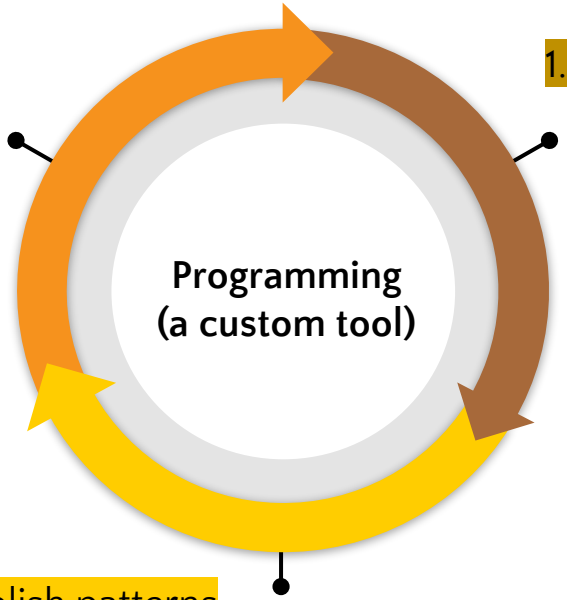
   Journal articles (777) in PDF format were converted to .TXT and cleaned up to see if formatting was accurate

2. Establish patterns

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

3. Code patterns to retrieve results

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

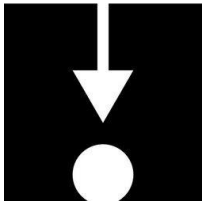


Corpus software packages

AntConc

WordSmith

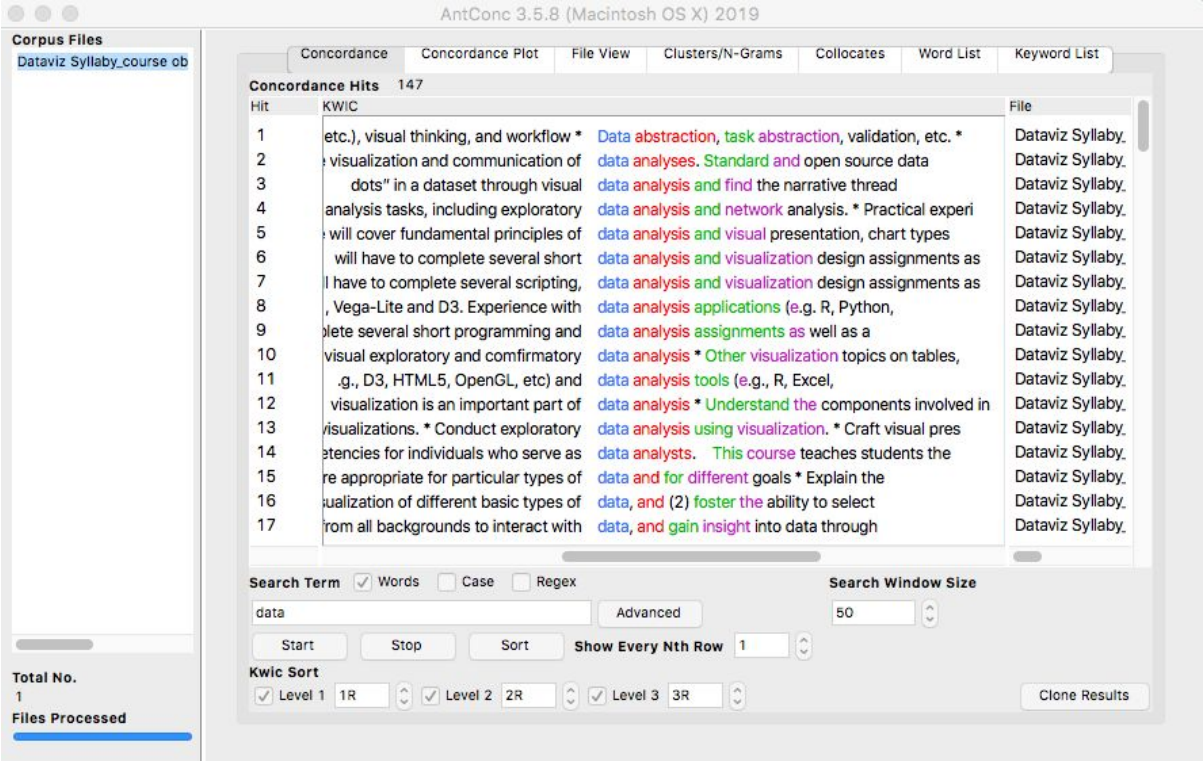
Wmatrix



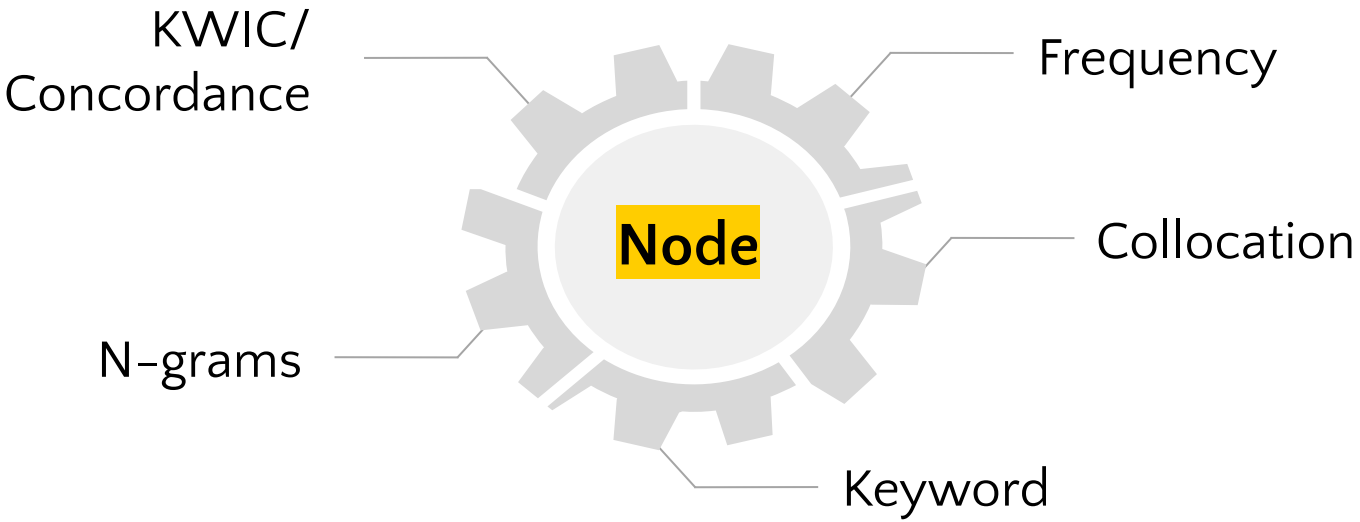
LancsBox

Sketch Engine

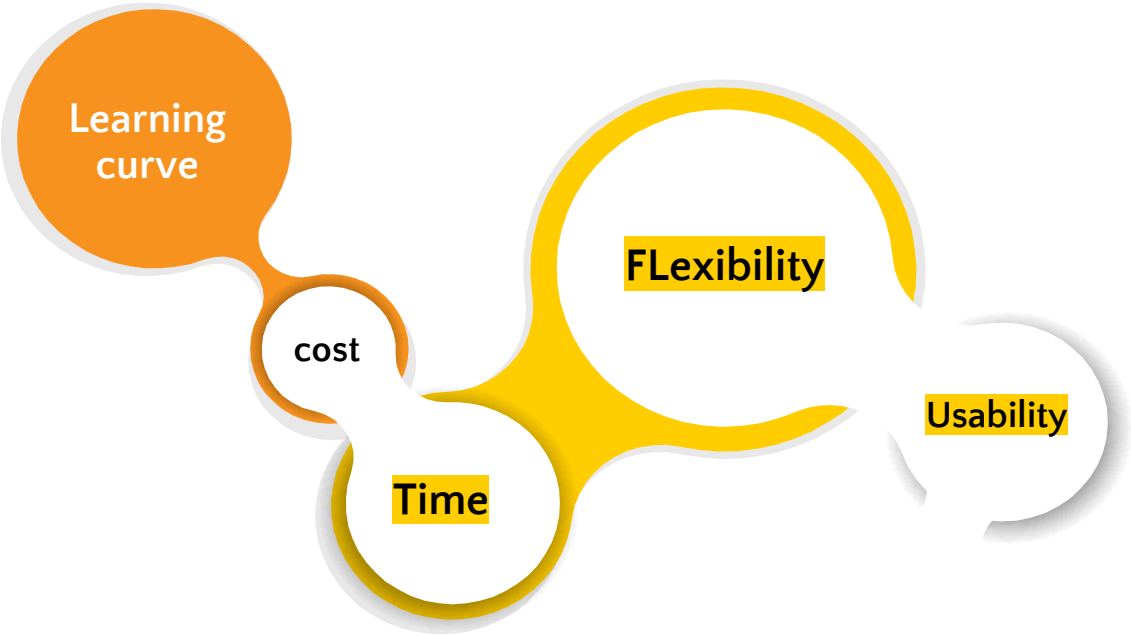
Corpus software interface



Key corpus linguistics concepts



Comparison: programming vs. software



Differences: learning curve

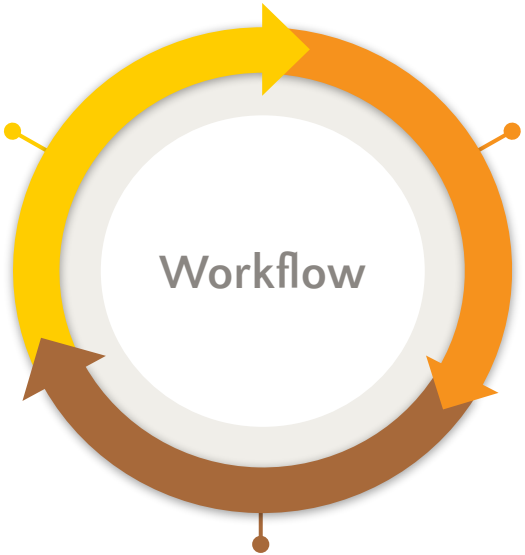
	Knowledge needed
Programming	syntax, variables, data structures, algorithms, and debugging
Corpus software	basic corpus linguistic concepts & how to apply them in context

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. Kilgarriff 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