The Lexometer: A Shiny Application for Exploratory Analysis and Visualization of Corpus Data

Oufan Hai, Matthew Sundberg, Katherine Trice, Rebecca Friedman, Scott Grimm

University of Rochester, Department of Linguistics, USA

`scott.grimm@rochester.edu`
`{ohai, msundbe2, ktrice, rfried11}@u.rochester.edu`

Abstract

Often performing even simple data science tasks with corpus data requires significant expertise in data science and programming languages like R and Python. With the aim of making quantitative research more accessible for researchers in the language sciences, we present the Lexometer, a Shiny application that integrates numerous data analysis and visualization functions into an easy-to-use graphical user interface. Some functions of the Lexometer are: filtering large databases to generate subsets of the data and variables of interest, providing a range of graphing techniques for both single and multiple variable analysis, and providing the data in a table format which can further be filtered as well as provide methods for cleaning the data. The Lexometer aims to be useful to language researchers with differing levels of programming expertise and to aid in broadening the inclusion of corpus-based empirical evidence in the language sciences.

Introduction

The Lexometer is capable of cleaning and filtering data and supports a range of common analyses and visualizations of corpus data. The application provides an intuitive graphic user interface (GUI) which allows to greatly accelerate corpus analysis for supported tasks. The goal of the app is to reduce the the amount of time researchers need to spend to write complex programs to perform basic data science tasks in data exploration and visualization. The Lexometer is able to accept all corpus databases in .csv (comma separated values) format and thus can be quite generally applied. At present, our work has focused on using the Contemporary Corpus of American English (COCA) (Davies, 2009) subsequent to processing in an NLP pipeline, which this paper describes.

Annotated Database for Lexical Investigation

We have designed the application to handle two primary types of dataframes:

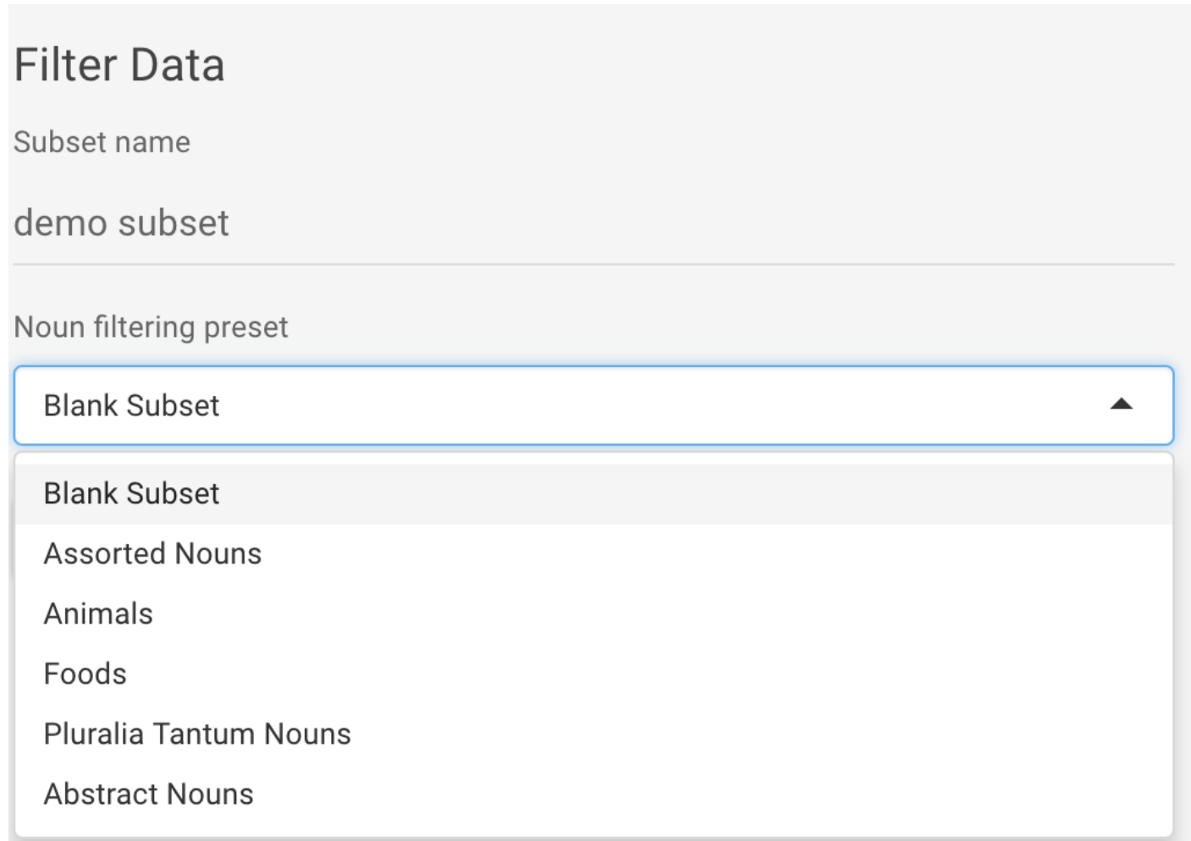
1. Dataframes containing corpus occurrences of a lexical item along with annotations of properties of those occurrences. (*Individual Databases*)
2. Dataframes containing aggregate statistics of a lexical item's distributional (or other) properties (*Global Databases*).

Lexometer: a Shiny Application

Data Filtering and Subset Generation

The Lexometer offers options for users to build subsets of corpora based on user-defined constraints on the database.

- The constraints can be listed values, values included or excluded by a grep pattern, or numerical filtering with greater-than or less-than values.
- For global noun databases, we have included the ability to pre-define multiple noun subsets to streamline the process of generating noun sets repeatedly examined by the researcher.



Filter Data

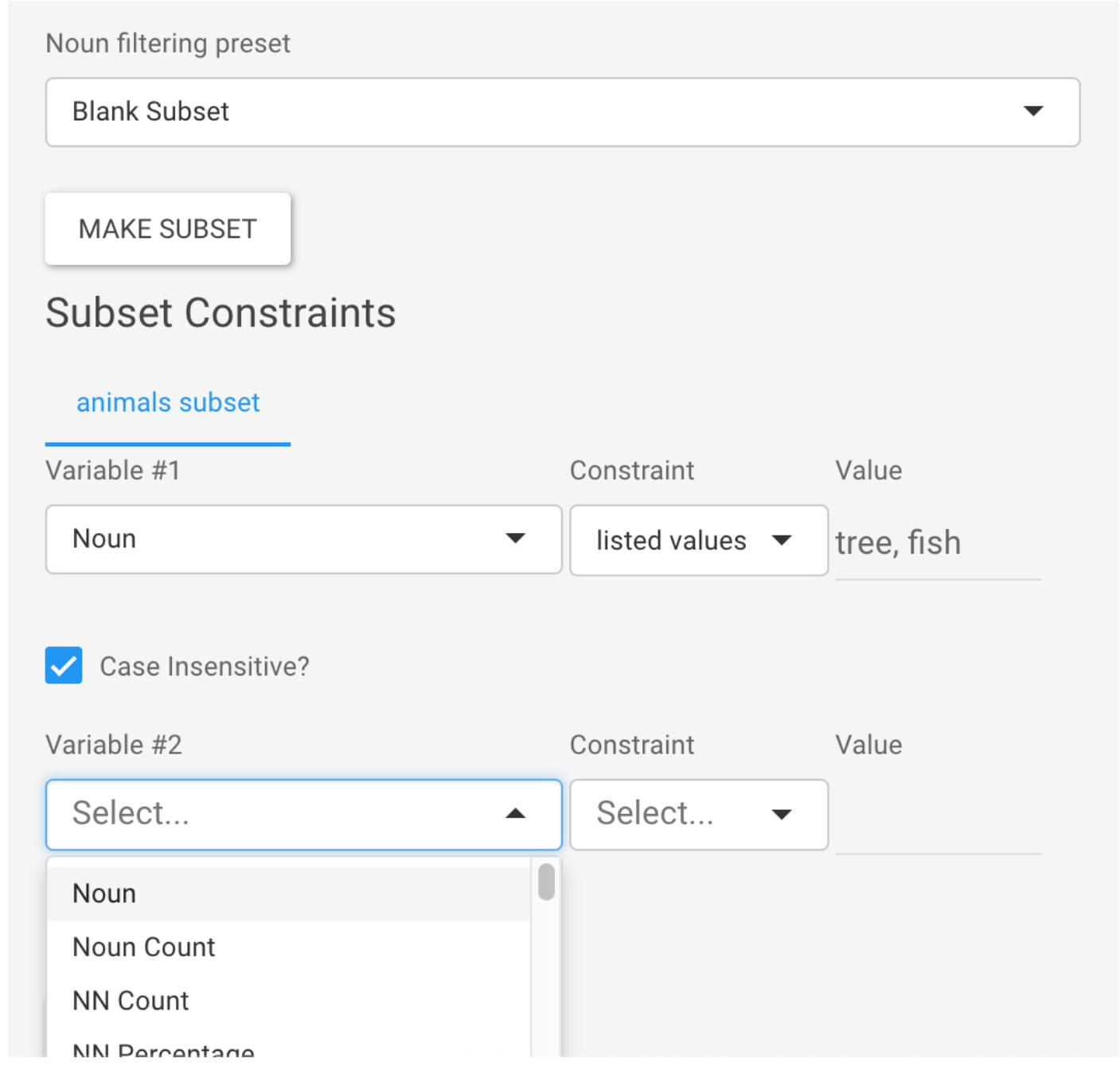
Subset name

demo subset

Noun filtering preset

Blank Subset

Figure 1: Subset presets for Data Selection



Noun filtering preset

Blank Subset

Subset Constraints

Variable #1: Noun — Constraint: listed values — Value: tree, fish

☑ Case Insensitive?

Variable #2: Select... — Constraint: Select... — Value:

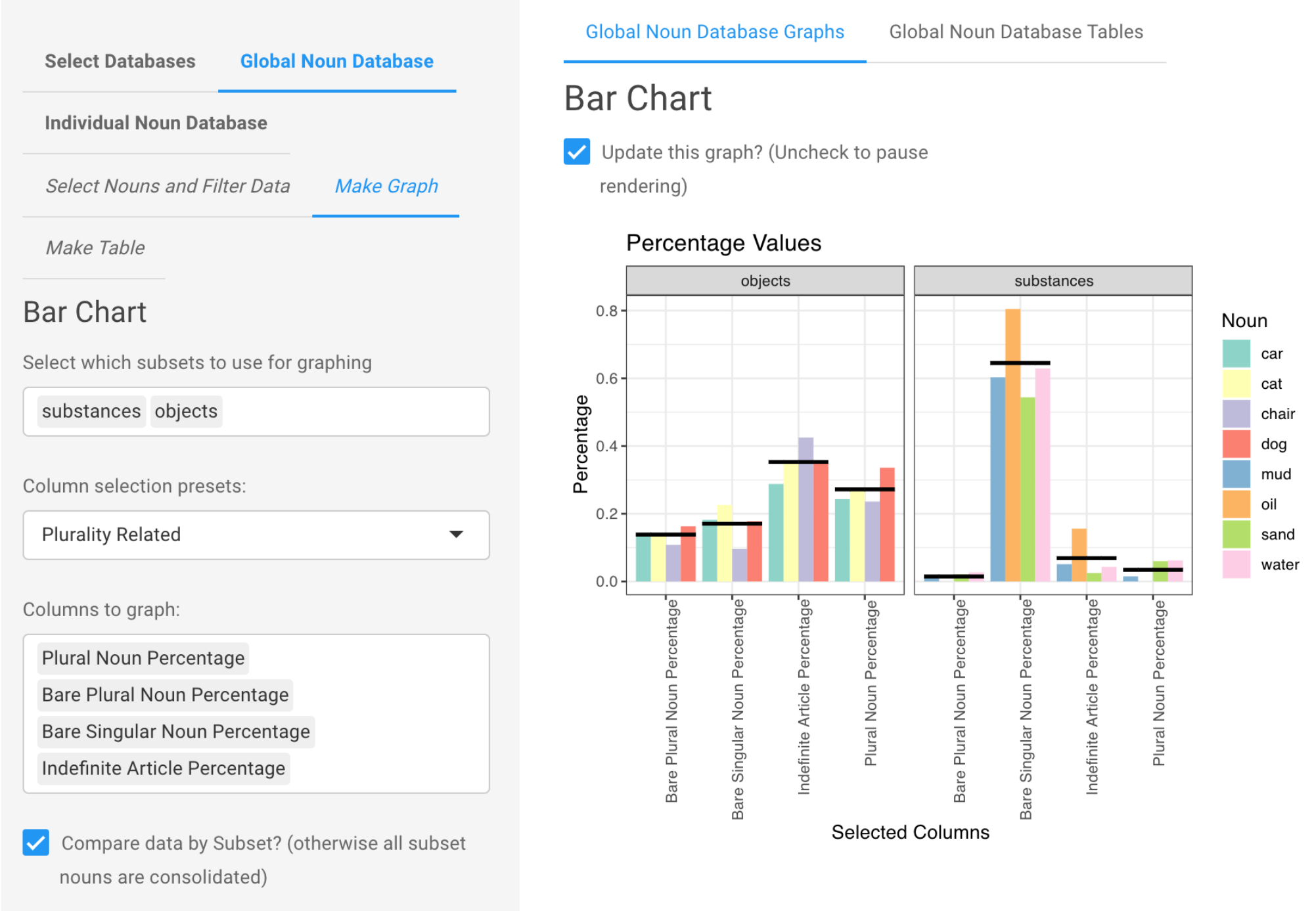
Figure 2: Data selection for Blank Subsets

Data Visualization

- The Lexometer provides quick access to common data visualization methods such as a bar graph, group-point graph, co-occurrence graph, and two variable comparison graph.
- A wide range of options are available for users to adjust their graphing settings to suit their needs.
- The user can download the graph as a .pdf or .png file with the click of a button.

Creating and Contrasting Noun Groups

- Users define subsets (via the Select Nouns and Filter Data Tab).
- Users choose the columns they wish to graph. The columns are read from the Global Noun Database input.
- Users are able to preset constellations of variables that the user may wish to repeatedly graph to streamline the process.
- Figure 3 is an example of a graph generated following the above procedure to display characteristics of 'count' and 'mass' nouns.



Select Databases
Global Noun Database
Individual Noun Database
Select Nouns and Filter Data
Make Graph
Make Table
Global Noun Database Graphs
Global Noun Database Tables

Bar Chart

Select which subsets to use for graphing: substances objects

Column selection presets: Plurality Related

Columns to graph: Plural Noun Percentage, Bare Plural Noun Percentage, Bare Singular Noun Percentage, Indefinite Article Percentage

☑ Compare data by Subset? (otherwise all subset nouns are consolidated)

Bar Chart

☑ Update this graph? (Uncheck to pause rendering)

Figure 3: A bar plot contrasting distributional features of count and mass nouns

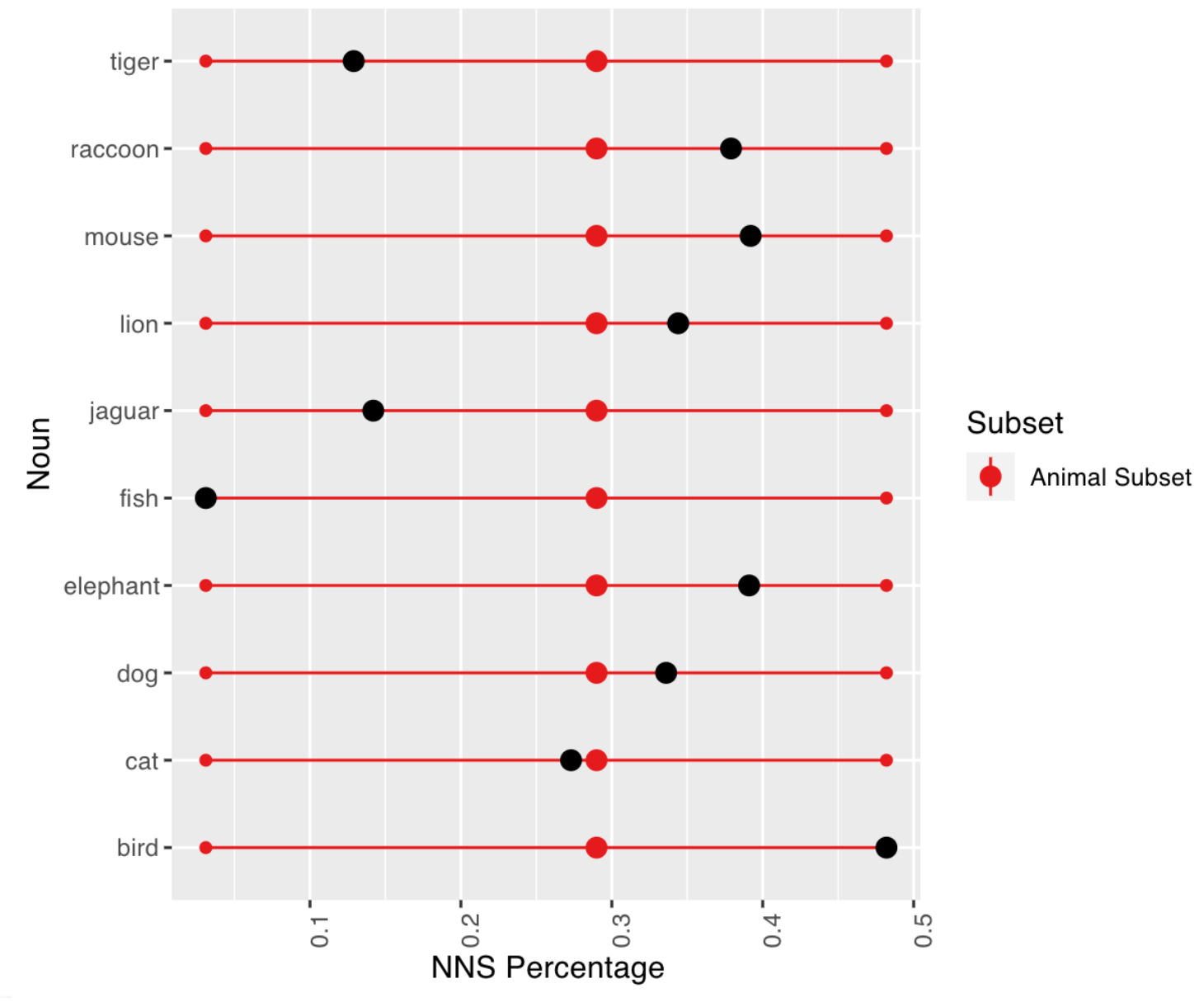
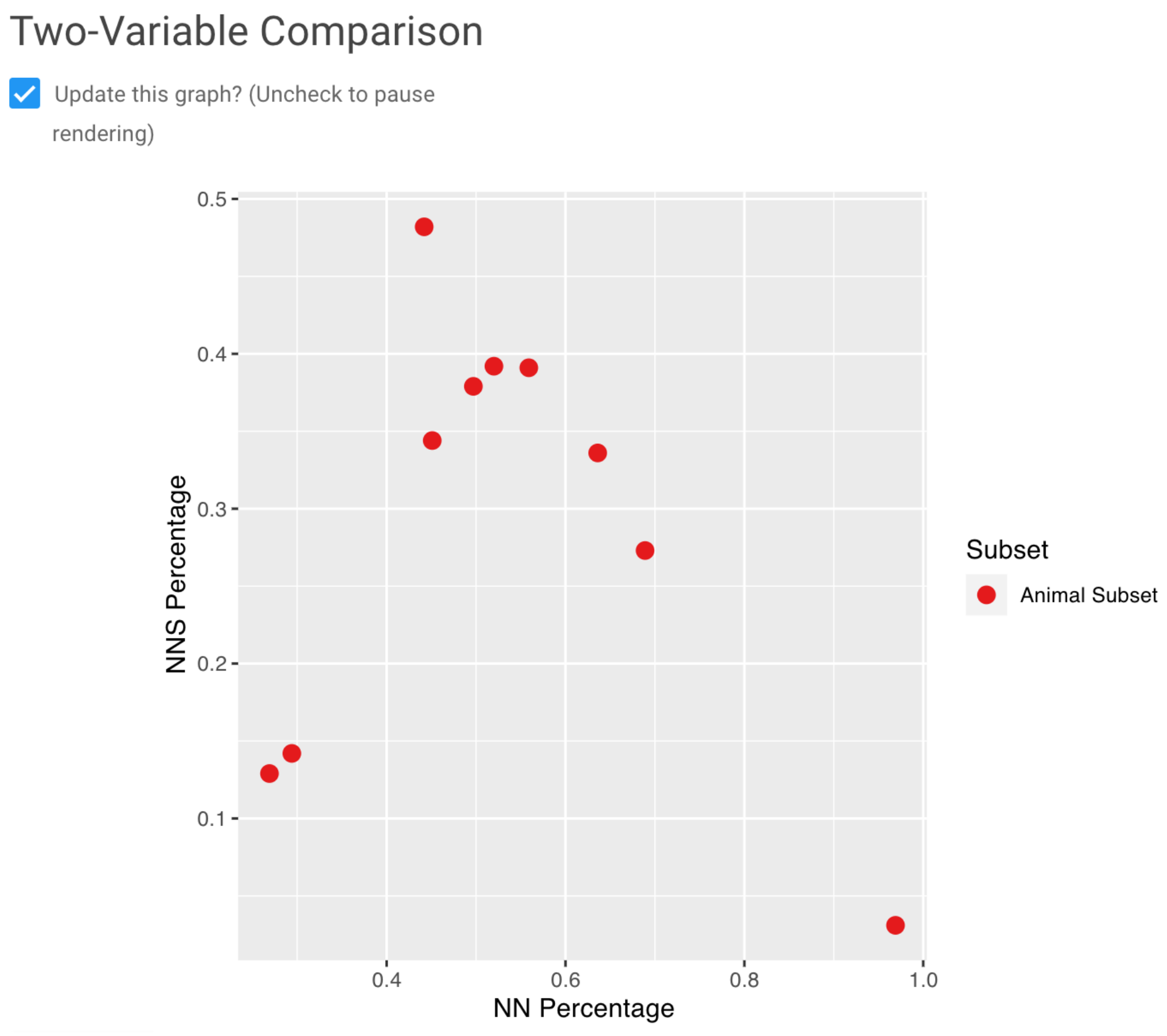


Figure 4: Group-point graph



Two-Variable Comparison

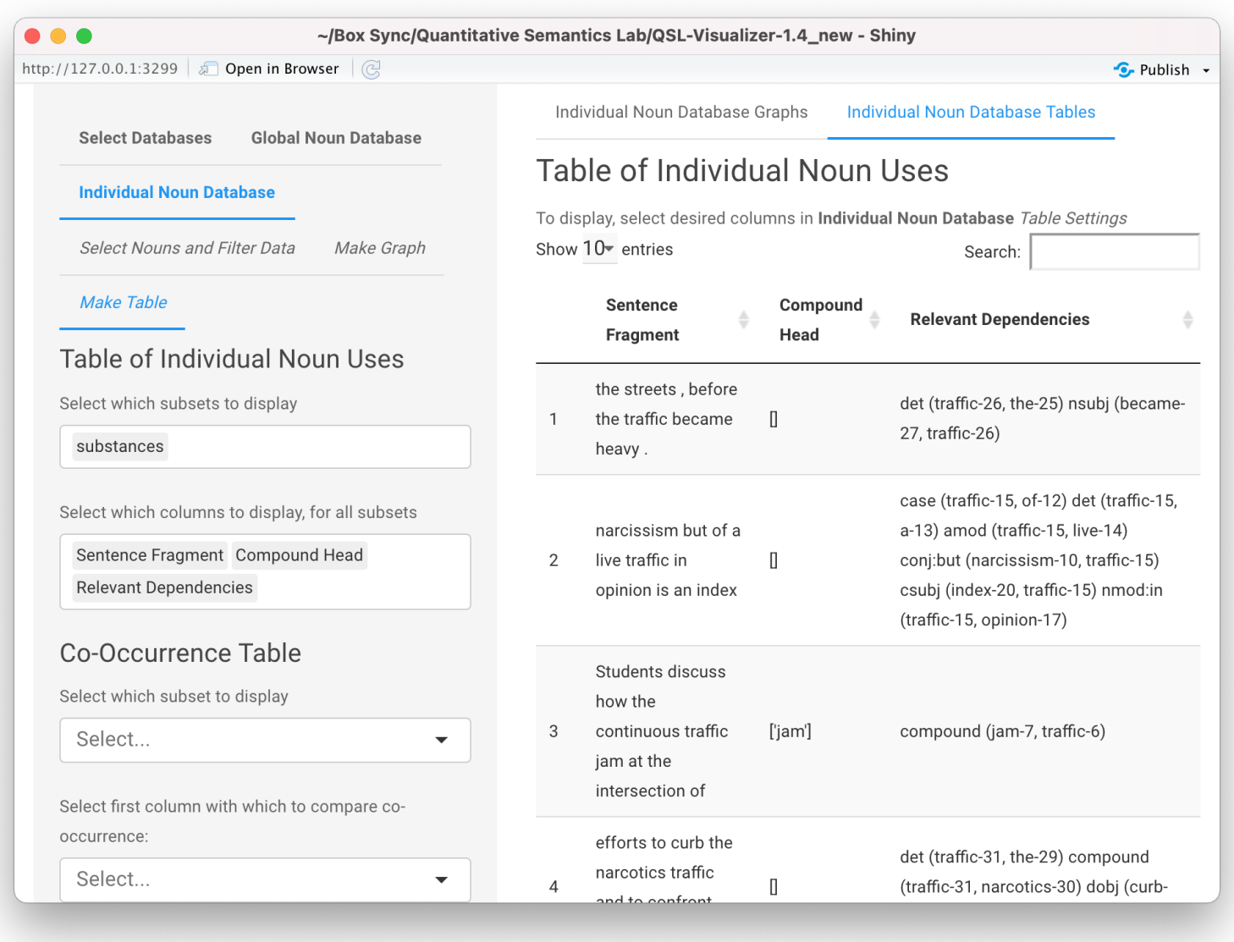
☑ Update this graph? (Uncheck to pause rendering)

Figure 5: Two variable comparison graph

Table Generation

The Lexometer is also able to generate tables based on selected columns of the noun subset created by the user.

- For the Global Noun database, users can create table views of different slices of the aggregate statistics.
- In the case of the Individual Noun databases, users can examine individual corpus occurrences and their annotated properties.
- All tables generated within the Lexometer can be downloaded as .csv files by clicking on a "Download CSV" button (in parallel to how PDFs of graphs can be downloaded).



~/Box Sync/Quantitative Semantics Lab/QSL-Visualizer-1.4_new - Shiny

http://127.0.0.1:3299 Open in Browser Publish

Select Databases
Global Noun Database
Individual Noun Database
Select Nouns and Filter Data
Make Graph
Make Table
Individual Noun Database Graphs
Individual Noun Database Tables

Table of Individual Noun Uses

Select which subsets to display: substances

Select which columns to display, for all subsets: Sentence Fragment, Compound Head, Relevant Dependencies

Co-Occurrence Table

Select which subset to display: Select...

Select first column with which to compare co-occurrence: Select...

Table of Individual Noun Uses

To display, select desired columns in **Individual Noun Database** *Table Settings*

Show 10 entries — Search:

	Sentence Fragment	Compound Head	Relevant Dependencies
1	the streets , before the traffic became heavy .	[]	det (traffic-26, the-25) nsubj (became-27, traffic-26)
2	narcissism but of a live traffic in opinion is an index	[]	case (traffic-15, of-12) det (traffic-15, a-13) amod (traffic-15, live-14) conj:but (narcissism-10, traffic-15) csubj (index-20, traffic-15) nmod:in (traffic-15, opinion-17)
3	Students discuss how the continuous traffic jam at the intersection of	['jam']	compound (jam-7, traffic-6)
4	efforts to curb the narcotics traffic and to confront	[]	det (traffic-31, the-29) compound (traffic-31, narcotics-30) dobj (curb-

Figure 6: Table of Selection of Individual Noun Aggregate Data

Conclusions

The Lexometer makes the exploration of quantitative linguistic data more accessible and accelerates the use of many common functionalities.

- Our goal is to be able to involve a greater portion of the language science community in corpus and quantitative work even if they have not yet developed sophisticated programming skills in, e.g., R and Python.
- The Lexometer simplifies a range of tasks, from cleaning data to making plots, which we hope will spur greater participation in the broader language community to adopt quantitative methods as a part of their research portfolio.
- In future work, we expect that adapting the Lexometer to databases beyond those developed in our research will aid us to generalize and improve many of the functionalities as well as the UI.
- Available at `https://quantitativesemanticslab.github.io`

References

[1] Mark Davies. The Corpus of Contemporary American English: 450 million words, 1990-present. Available online at http://corpus.byu.edu/coca/, 2008.