Fall 2017

# Part-of-Speech Tagged Corpus Search

Jacob Tootalian
*University of South Florida*, jactootalian@gmail.com

# Part-of-Speech Tagged Corpus Search

Jacob Tootalian

[**NOTE:** This assignment can be retooled to look at other part-of-speech-tagged corpora available on CQPWeb (https://cqpweb.lancs.ac.uk) or BYU Corpus (https://corpus.byu.edu), including translations of the *One Thousand and One Nights*, Shakespeare's plays, the early English books online corpus, as well as various archives of periodicals.]

## Assignment Prompt:

Students will identify a phrase they find interesting in one of the course readings, discern the basic linguistic formula of that phrase, and try to find an article from the scientific journal *Philosophical Transactions*—accessed through the CQPWeb search interface (http://corpora.clarin-d.uni-saarland.de/cqpweb)—that uses the same language pattern. They will then respond to a series of questions, prompting them to analyze the linguistic pattern they have identified and to argue for what is significant about its use in both the literary work and the scientific article (**1.5-3 pages, single-spaced**).

The goal of this assignment is to encourage you to develop your analytical and argumentative skills by exploring not just the diction but also the syntax used in both literary and scientific texts. We'll be using the CQPweb interface to perform part-of-speech tagged searches of the scientific journal, the *Philosophical Transactions.* You may need to try a few different options before you find a language pattern that appears in the corpus.

Use the outline below to guide your research and analysis. Write up the assignment using the section headings below and for each of them include at least a paragraph. Make sure that your paragraphs would be understandable to a reader that hasn't read the questions that prompted them. Even though this is set up as a questionnaire, make sure you revise your responses. Your writing should be both analytical and persuasive. Here's a rubric to give you a sense of how these projects will be evaluated. Also, make sure to consult the notes at the bottom of this prompt for more information about effective writing strategies.

### Key Phrase:

Select a phrase or expression from one of the literary works we have read that you find interesting. Why do you find this phrase interesting? What is its function in the passage it appears in? How is it relevant to the larger themes of the text?

### Part-of-Speech Translation:

Translate your selected phrase into a searchable linguistic pattern with at least one variable. For instance, "that great Leviathan called a Commonwealth" could be translated to "that [adjective] [noun] called a [noun]." Make sure to double-check that you have correctly identified the part of speech for the words you're translating into variables. Explain why you have selected the variables you have and what you expect to find in the *Philosophical Transactions*.

### Search the Corpus:

Consult the exercise from our digital workday [**included below**] for details on actually performing the search. How many search hits do you get? Without clicking on any individual search result, survey what you find. Do you notice any patterns? What questions do you have before digging deeper into your results?

### Scientific Article:

Select one result that looks interesting to you. Click on both the passage and the metadata for the article you selected. Also click on the JSTOR link on the metadata page to access a scanned version of the article itself. Explain any relevant information you learn that might help us understand this article (author, year, title, etc.) What is this article about?

**Analysis:**

What does the selected phrase add to the scientific article you found? Why is the underlying pattern of syntax significant? Imagine some alternative ways of phrasing this expression to see what this pattern adds. How does this example from a scientific article affect your understanding of the original phrase you selected from a literary text. Are they using the pattern in the same way? What does this phrasing add to the work of literature?

**Conclusion:**

What does the comparison between these phrases suggest to you about the relationship between science and literature?

## IN-CLASS PREPARATION

### Step 1: Set Up Your Account

Create an account on Saarland University's website. Saarland University hosts the Royal Society Corpus, which is accessible through the CQPweb interface, a platform created by Andrew Hardie. You should receive an email prompt to activate your account. After you click through that prompt, make sure that you are able to log in to the interface.

### Step 2: Some Background

I'll be presenting some background on corpus search and the *Philosophical Transactions*, illustrating the possibilities of this technique with an example from my research. [**I have a Powerpoint presentation on this material that I can share**].

### Step 3: From 17th-Century Journal to POS-Searchable Corpus

Case study: Isaac Newton's "Letter ... Concerning His New Theory about Light and Colors"

This letter, along with the other articles from the *Philosophical Transactions* that researchers at Saarland included in the Royal Society Corpus, has gone through several versions in order to make it accessible to the kinds of searches we'll be performing. First, the original editions of the *Philosophical Transactions*, from 1664 onward, were digitally scanned as facsimile images by *JSTOR*. Here's what the facsimile of Newton's letter looks like. The team at Saarland then transcribed these articles into digital texts. Here's a textual transcription of Newton's letter from another project. What is special about the transcriptions created by Saarland is that each word was tagged with part-of-speech codes. Here's a section of Newton's letter with those tags.

A Note of Caution: Because we are dealing with texts that have been filtered through different phases--each step requiring a certain degree of human interpretation and technical translation--the texts we are searching are not perfect. Be aware that typos and other kinds of mistranscriptions might prevent us from finding relevant results. The part-of-speech tags were applied with an automated tagging tool so there might also be errors (or just differences of interpretation) that affect the way that words are categorized. Also, language is complicated.

### Step 4: How to Search

We are accessing the Royal Society Corpus through CQPweb interface, which allows us to perform more complicated search queries. Because of that it takes a few more steps than a normal search would take. The part-of-speech tags follow a system called the PenTreebank Tagset. Consult that key to figure out which codes to use for different categories of words.

- For direct word searches, make sure to put each word in its own quotations marks. (**Note**: Remember that words are case sensitive; this can get complicated with early modern printed texts, which followed looser rules for when words should or shouldn't be capitalized):

  *"the" "received" "laws" "of" "Refraction"*

- For searches that use part of speech tag variables, use the following formula for each term:

  *[pos="XX*"]*

- For XX fill in the appropriate part-of-speech code from the PennTreebank Tagset. For instance, if I wanted to search for any words tagged as singular nouns, I would type this into the search box:

  *[pos="NN*"]*

- Putting these two methods together, we can string together complex search queries that allow us to find patterns of language without anticipating the precise vocabulary that they use. So, if I wanted to find phrases similar to Newton's reference to "the received laws of Refraction," I could put together a search formula that uses variables for "received" (a past participle verb) and "Refraction" (a singular noun):

  *"the" [pos="VVN*"] "laws" "of" [pos="NN*"]*

**Step 5: Searching!**

1. Once you are logged into the interface on Saarland's website, click on the most recent version of the Royal Society Corpus (V3.6.0).

2. Then, try the search pattern we formulated above to find other articles in the *Philosophical Transactions* that use a similar phrasing. (Don't just copy/paste it; try to type it out yourself so that you get used to the formulas for searching).

3. Take notes as you go through this process. You'll be discussing your findings in your group.

4. Before clicking on any of the results, look over the list of search hits. What do you notice about these phrases? What words fill in the blanks in your query? Are these phrases similar to or different from Newton's expression? How?

5. Pick one of the hits, and click on the phrase itself to see it in its fuller context. Read over the passage, and try to figure out what this article might be about. How does the phrase you found seem to be contributing to the meaning of the passage or the goals of the article?

6. Go back to the list of search hits. Now click on the numerical code in the "Filename" column for the article you selected. Look over the metadata for that article, and try to figure out what some of the rows mean. Does any of that information help you understand the passage you just read? How?

7. Talk about what you found with your peers. Do you notice any patterns between the search results?

**Step 6: Try Your Own Search**

1. Come up with a search query of your own that uses at least one part-of-speech tag variable. You can look through Newton's letter for another phrase you find interesting or that you have questions about, or you can come up with another expression.

2. If you don't get any hits (but you're sure you have your query properly formulated), why do you think that might be?

3. If you do get hits, go through the same steps above to see what you find.

4. Share your findings with your peers.