Using ElasticSearch, applied graph theory, and the EEBO and ECCO corpora to identify the sources of Samuel Johnson's dictionary quotations

Beth Rapp Young
University of Central Florida
byoung@ucf.edu
https://johnsonsdictionaryonline.com

NOT FINAL – PLEASE DO NOT CIRCULATE OR QUOTE WITHOUT PERMISSION

SLIDE 1 – TITLE PLUS EXAMPLE ENTRY (FOR FACILE)

Samuel Johnson's *A Dictionary of the English Language* was hailed as a major literary accomplishment when it was first published in 1755, and it still remains an important literary and reference text today. Our NEH-funded project, johnsonsdictionaryonline.com, is making this dictionary more accessible by creating an online, searchable database of the 1755 and 1773 folio editions, with a search interface comparable to other modern dictionaries.

Johnson famously supplied evidence for his definitions in the form of quotations (as you can see here in his 1755 entry for *facile*). Each folio edition contains more than 110,000 of these quotations, drawn from respected writers in diverse fields (agriculture, religion, ship-building, literature, politics, medicine, print-making, history, etc.) The collection of quotations was perceived as one of the more valuable aspects of the folio editions, and they illustrate shades of meaning that are difficult to capture in definitions. On occasion (e.g., *self*, sense 8), Johnson doesn't even provide a definition; he just says, essentially, "You can figure it out from the quotations." ["8. It is much used in composition, which it is proper to explain by a train of examples." and then more than a folio page of examples]

However, these quotations are not as useful to modern readers, partly because of language change, partly because they have no idea who the authors are or what these titles refer to.

SLIDE 2 – XML FROM BIBLIOGRAPHY PLUS FUNCTIONS IT MAKES POSSIBLE

In order to make this text more accessible to modern readers, one of our goals is to identify as many of the possible sources of Johnson's quotations as possible, and link them to additional information.

When we have identified an author/title, we insert the information into the xml markup. On the bottom right you can see a small slice of our bibliography that shows the XML tags we insert for title/author. These tags link to additional

information behind the scenes, information that readers can choose to display.

On the top of the slide you see additional information about authors—most of that information is already there. Clicking the little circle with an "i" in it will pop that information up. Behind the scenes, we're researching the titles also, with the goal of providing information such as the English Short Title Catalog number, Library of Congress record, and potentially an e-copy of the text.

We are also using this information to build an author / title browse (coming soon). The author browse will solve the problem of inconsistent abbreviations (e.g., Shakespeare) and missing information (e.g., Macbeth).

Note that we are working to identify a possible source, which is not the same thing as Johnson's exact source. We

knew from the start that we would have trouble identifying Johnson's exact source, so our goal is to just find *a* source, so that someone who is interested has a decent starting point for more research. (Brian Grimes, an independent scholar, has been working to identify Johnson's source: sjdictionarysources.org)

But in order to supply this information, we need to investigate these sources ourselves. Johnson never provides what we would consider to be "complete" bibliographic information. Often he lists an author without a title or a title without an author, or he abbreviates one or both so that we aren't sure where the text comes from. For example, "Brown" could be Sir Thomas Browne, Edward Browne, Isaac Hawkins Browne, or Moses Browne. "Cla" could be "Clarendon" or

Richardson's Clarissa. And occasionally, the printed citation is not correct.

So, we are searching for quotation sources ourselves. There's a long tradition of investigating Johnson's sources—as far as I'm aware, no one has succeeded in locating all of them. Part of the reason is the sheer number of quotations.

SLIDE 3 – QUOTATION FROM JOHNSON'S DICTIONARY "PREFACE"

Another part of the reason is the fact that Johnson rarely quoted his sources exactly. He explains why, in his Preface:

"When first I collected these authorities, I was desirous that every quotation should be useful to some other end than the illustration of a word; I therefore extracted from philosophers principles of science; from historians remarkable facts; from chymists complete processes; from divines striking

exhortations; and from poets beautiful descriptions. Such is design, while it is yet at a distance from execution. When the time called upon me to range this accumulation of elegance and wisdom into an alphabetical series, I soon discovered that the bulk of my volumes would fright away the student, and was forced to depart from my scheme of including all that was pleasing or useful in English literature, and reduce my transcripts very often to clusters of words, in which scarcely any meaning is retained; thus to the weariness of copying, I was condemned to add the vexation of expunging. Some passages I have yet spared, which may relieve the labour of verbal searches, and intersperse with verdure and flowers the dusty desarts of barren philology.

"The examples, thus mutilated, are no longer to be considered as conveying the sentiments or doctrine of their

authours; the word for the sake of which they are inserted, with all its appendant clauses, has been carefully preserved; but it may sometimes happen, by hasty detruncation, that the general tendency of the sentence may be changed: the divine may desert his tenets, or the philosopher his system."

Given that there are more than 220,000 quotations in the two editions we're putting online, finding these quotations has been a not-inconsiderable challenge.

We've tried several different strategies to research them:

First, we enlisted human beings to look for the quotations one at a time. Here we had the option of searching author-by-author, such as all the quotations marked as being from Milton, which was reasonably efficient but about the most boring way to do it. Or we could search word-by-word, such as researching all the quotations in the entries in the

letter K, which was slightly more interesting but still very, very slow. To do this, we had to guess at which keywords to type into a search engine to find the quotation, and often it took several attempts. We located about 50,000 matches that way, but the process was so laborious and slow we knew it wasn't enough.

Second, a team of senior computer science majors programmed a way to do a "fuzzy" search, using a linear search with a sliding window technique to identify likely sources in a corpus of about 28,000 texts, and to give each match a likelihood score. This process worked, but it was very difficult to work with the results, and the search it took more than 5 months to run. (This search method was very helpful for limited datasets, though—for example, matching all

quotations attributed to "Dryden" to just the one piece that Dryden's son wrote.)

We're now on our third attempt, another computer science senior design project. This application uses elasticsearch, a search engine that is extremely fast and can run searches in parallel; it's > 1200x faster than the linear search, taking 2.5 hours to run on a corpus that is more than twice as big. Elasticsearch powers giant projects like LinkedIn, Wikipedia—much larger projects than ours! [ADDED IN RESPONSE TO A POST-TALK QUESTION: ElasticSearch is open source software from the Apache foundation. Additional levels of service/storage are not free, but we just used the open source version: https://www.elastic.co/about/free-and-open. The UI we used

to interact with ElasticSearch is Kibana

https://www.elastic.co/kibana/]

SLIDE 4 - HOMEPAGE OF QUOTATION MATCHING TOOL

This is the homepage of our quotation matcher tool; it's not public so I'm not giving the URL but the interface is online so that multiple people can login and work remotely.

This tool interfaces with the elasticsearch results that compared Johnson's quotations to texts in the Early English Books Online (EEBO) and Eighteenth Century Collections Online (ECCO) databases. It has a lot of features that work more or less well.

I'm going to briefly explain how it works, then talk about challenges that we still face.

When you click the "Start Matching" button, you go to the Match screen.

SLIDE 5 – MATCHING SCREEN FROM QUOTATION MATCHING TOOL

Here, you see a randomly-chosen quotation on the left, and a list of possible matches on the right.

To select the best match, you click “Mark Canonical” to indicate that the quotation is drawn from this text. You can also “skip” (which means the quotation will show itself again later) or “no match” (which means none of the texts is a match and the quotation never shows up again.

In this image, you can see a match that has already been made.

If the work you select as match has already been tagged, then those tags (the author and title ID) get attached to your quotation.

This quotation has been tagged to match Shakespeare's Macbeth.

SLIDE 6/11 – VIDEO/IMAGE OF TAGGING SCREEN

If the work that we select in the match screen has not already been tagged, we get to this screen that lets us make a tag. The screen is interactive so that when you type in the boxes, the suggested tags change, and if you don't select a tag, this screen will make a new one. The tag is saved with the work, so any other quotations matched to the work will automatically get the same tag. There's a box we can click if the work is an anthology, so that we can narrow down the work title further later if we like.

[This video clip comes from the design team's final presentation video: https://youtu.be/AHJp_-Vx18E. Members

of the design team: Ethan Berry, Zayne Dyal, Jesse Gingold, Matthew Pena, Taylor Richards]

SLIDE 7 – IMAGE OF QUOTATION GRAPH SCREEN

After we have chosen a work and (if necessary) created a tag, the matcher then shows us all of the places in both editions where it thinks Johnson used this same quotation.

Quotations are used on average 2.4 times each across both editions—this particular quotation is used more times than average. The software recognizes that these quotations are matched even when the different instances have been edited differently, as you can see here.

On this screen, we can select the quotations that really do match each other and give them all the same tags at the same time, which speeds up the matching process.

Making this part of the tool was a challenge for the software team. An early effort had trouble disambiguating, for example, Bible verses, the many quotations that describe plants, etc.

The method that they used was to use elasticsearch to find similar quotations among all of the quotations in both editions. They generated scores depending on how close the match is, so a perfect match would be a 1. Whenever quotations score close enough, an edge is added between the two nodes in an enormous quote graph. When the graph is constructed, they searched for strongly connected nodes (which represent quote groups) with each member of the group having a bidirectional relationship with at least one other member of the group. Any time one member of the group is selected for a match, the whole group is displayed

and then we confirm that each member of the group belongs there.

This screen is also useful for discovering typos—where the word "sought" in one quotation becomes "fought" in another—and also those times that an author attribution changes—when Johnson uses the same quotation but attributes it to different people.

SLIDE 8 – MATCHING CHALLENGES

We hoped that this software would make short work of the quotation identifying process, and it has been very helpful . . . but there have been some challenges.

I'd hoped to be able to sit students down at the quotation matcher and turn them loose, but it turns out they need to be prepared to recognize and work with. Students expect EXACT

quotations, and once the quotations are not exact, identifying the matches is trickier. And which match is "best"?

Some corpus-specific challenges are listed here. We try to be very careful about concluding that Johnson got a match "wrong" (though on rare occasions he undoubtedly did). Any time our match doesn't agree with Johnson's attribution, we spend extra time scrutinizing it to make sure we are correct.

I regularly feel like I don't really know enough to choose the "best" match, but this is where I refocus on our goal: to find *a* match that can be useful to readers who need more context, even if our match is not THE match that Johnson worked from. In fact, Johnson's text might not even be IN this corpus.

Fortunately we never promised to locate ALL quotation sources. We'll get as close as we can, but it's quite likely that some of them will elude us.

SLIDE 9 – TAGGING CHALLENGES

We also have challenges with the tagging.

One great thing about this tool is that tagging is easy! And making new tags is also easy! Once a new tag has been made, it lodges itself into the database and presents itself in the choices for additional works in the future. Which is terrific—IF it is a good tag. But if it's a bad tag—if it duplicates existing tags, or if it entangles work that should have disambiguating tags—that is a problem that we need to sort out when we add the information to the dictionary. Which means we have created extra work for ourselves.

Examples: There are TWO John Drydens, "THE" John Dryden and his son, also named John Dryden. We have disambiguated them by the year in which they died. But a student "helpfully" created a new John Dryden tag that did NOT specify dad or son, and now we need to disambiguate them all over again.

Another easy duplicate to make is one that changes the capitalization. Or the spelling (which is very tempting for students who don't realize that the same author name could be spelled in different ways).

We're addressing these challenges in a few ways:

1. **Eyeballing the tags before we add any to the corpus, then find/replace**
2. **"layering" the info into our xml rather than replacing it. If we have already identified an author in one of our previous attempts, we don't replace it, we just add the new attribute. Later we can pull all the works that have >1 tag and decide which one to**

keep. This means our earlier efforts are not wasted! They are helpful!

SLIDE 11 – TAKEAWAYS + GRAPH

(The background image here is all the quotations in both editions, graphed)

Takeaways:

- **Don't be afraid to keep trying and changing your strategies**
- **Student work is student work (even this fabulous quotation matching tool has rough edges and glitches)**
- **You'll never find the perfect technology solution but that helps make the work interesting**
- **I have been pleasantly surprised about how eager students have been to work on this project. I've had to limit the quotation matching because it is too easy for things to go horribly wrong, but I have been fortunate to have between 20 – 100 volunteers (from all over, not just students) working on this every semester. Numbers are highest at the beginning of the semester, then people fall away, but that's ok, because every little bit brings us closer.**

And sometimes people who have fallen away wind up returning.

- **Network! I'm not naturally good at that, but it has been really helpful – that's how I got hooked up with the computer science students and also how I've located volunteers.**

Johnson's work is challenging to digitize; it isn't at all systematic. It's almost like he didn't foresee that we would be trying to wedge his dictionary into a database!

So for the project, I have often found myself echoing what he said in his preface:

Such is design, while it is yet at a distance from execution.

But with technology tools ever improving, we can get closer to achieving our design.