Piloting a Workflow for Extracting Author Citations from Samuel Johnson's *Dictionary of the English Language*

Jasmine Wong[1] and Ryan Dubnicek[1]

[1] University of Illinois, Urbana-Champaign
jmwong5@illinois.edu

Abstract. Since the 18th century, English-language dictionaries have used quotations from written works to illustrate a word's use in context. These quotations form a link between language authority and literary authority. In this paper we pilot a workflow for identifying, extracting, and counting author citations in Samuel Johnson's *Dictionary of the English Language* to investigate how authors in a defined corpus are represented. We consider how these authors are distributed across the text and compare our results to past studies that used different methodologies. We find a consistency that encourages the broader application of our workflow on other dictionary texts, enabling further study of author citations in dictionaries across time.

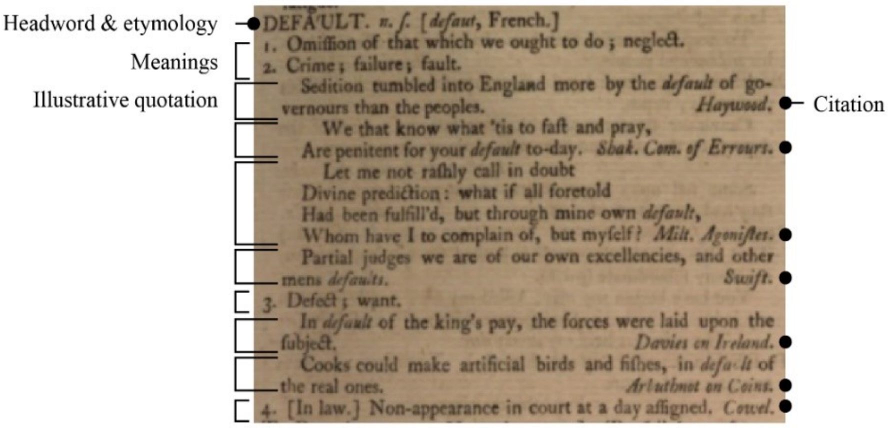
Keywords: Text Mining, Samuel Johnson, Dictionaries.

1 Introduction

Contemporary English-language dictionaries use quotations from written sources to demonstrate the language's use in context. Although Greek and Latin dictionaries have maintained this tradition as early as the 16th century [1], it was not widely adopted by English lexicographers until Samuel Johnson published his *Dictionary of the English Language* in 1755 [2]. Given the popularity of Johnson's dictionary, these quotations may have influenced societal views of literary importance and influenced the canon of Western literature that persists today. This study looks closely at the attributed authors in Johnson's quotations and pilots a text mining workflow for extracting and analyzing this author corpus.

1.1 Samuel Johnson's *Dictionary of the English Language*

Each entry features the headword, etymology, and a numbered list of meanings with their respective illustrative quotation(s) (see Fig. 1). A complete citation includes the author's surname and the title of the cited work.



DEFAULT. *n. ſ.* [*defaut*, French.]

1. Omiſſion of that which we ought to do; neglect.
2. Crime; failure; fault.

Sedition tumbled into England more by the *default* of governours than the peoples. *Haywood.*

We that know what 'tis to faſt and pray,
Are penitent for your *default* to-day. *Shak. Com. of Errours.*

Let me not raſhly call in doubt
Divine prediction: what if all foretold
Had been fulfill'd, but through mine own *default*,
Whom have I to complain of, but myſelf? *Milt. Agoniſtes.*

Partial judges we are of our own excellencies, and other mens *defaults.* *Swift.*

3. Defect; want.

In *default* of the king's pay, the forces were laid upon the ſubject. *Davies on Ireland.*

Cooks could make artificial birds and fiſhes, in *default* of the real ones. *Arbuthnot on Coins.*

4. [In law.] Non-appearance in court at a day aſſigned. *Cowel.*

Fig. 1. Entry for *DEFAULT*, scanned from Johnson's *Dictionary of the English Language* (1755), with parts of the entry labeled.

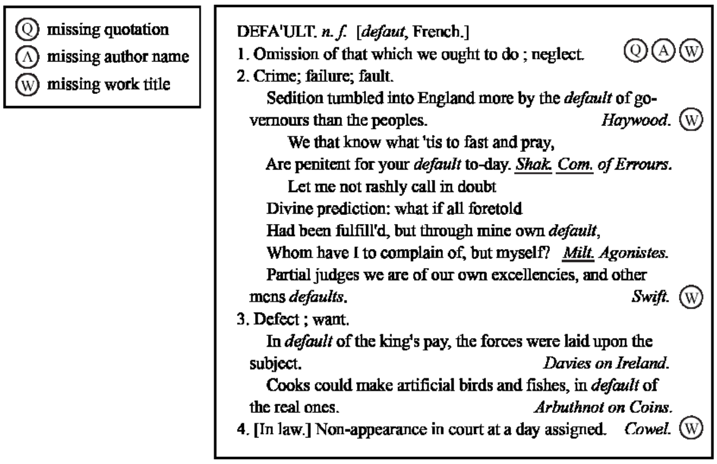
Johnson's Authorities. Johnson's cited authorities reflect his view of how language should be used and by whom it was best used – he explicitly gives preference to "writers of the first reputation to those of an inferior rank," of which he was the sole judge [3]. The final text constructs a specific view of literature, defined both by the works he chose and those he excluded. Although previous research has found that Johnson relies heavily on a relatively small group of authors to support his dictionary of roughly 43,000 words and 114,000 quotations [4], the exact distribution of authorship is yet unclear. Without a reliable method of automating the time-intensive citation and author extraction process, answering these questions will be an ongoing challenge.

2 Johnson's *Dictionary of the English Language*

We used a 1979 facsimile reprint of the first edition of Johnson's *Dictionary*, available for text analysis via the HathiTrust Research Center (HTRC) Data Capsule [5]. This dataset was selected based on the superior accuracy of its machine-encoded text relative to other manuscripts we evaluated. To pilot an extraction workflow, ten commonly-cited authors were selected, and their citations were extracted from the text in a two-stage process of programmatic and manual evaluation. The final citation counts were counted and grouped by letter section for analysis.

2.1 Formatting and Text Encoding

Due to the age and formatting of the original manuscript, our dataset presented a number of challenges. At the structural level, Johnson was highly inconsistent in his citation notation. He used a variety of abbreviations for author names and work titles (e.g. “Shakespeare” is variably cited with *Shak*, *Shaks*, *Shake,* and *Shakes*, among others*)* or frequently omits the author or work entirely (see Fig. 2).

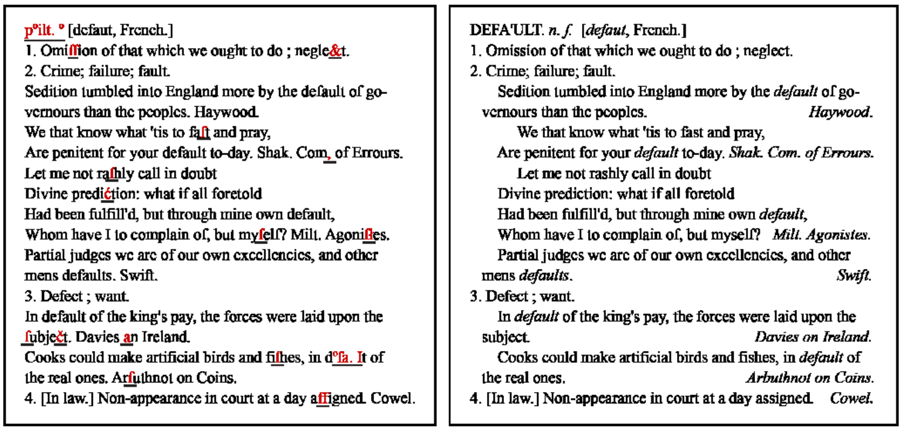


Ⓠ missing quotation
Ⓐ missing author name
Ⓦ missing work title

DEFA'ULT. *n. ſ.* [*defaut*, French.]
1. Omission of that which we ought to do ; neglect. Ⓠ Ⓐ Ⓦ
2. Crime; failure; fault.
Sedition tumbled into England more by the *default* of go-
vernours than the peoples. *Haywood.* Ⓦ
We that know what 'tis to fast and pray,
Are penitent for your *default* to-day. *Shak. Com. of Errours.*
Let me not rashly call in doubt
Divine prediction: what if all foretold
Had been fulfill'd, but through mine own *default,*
Whom have I to complain of, but myself? *Milt. Agonistes.*
Partial judges we are of our own excellencies, and other
mens *defaults.* *Swift.* Ⓦ
3. Defect ; want.
In *default* of the king's pay, the forces were laid upon the
subject. *Davies on Ireland.*
Cooks could make artificial birds and fishes, in *default* of
the real ones. *Arbuthnot on Coins.*
4. [In law.] Non-appearance in court at a day assigned. *Cowel.* Ⓦ

Fig. 2. Entry for DEFAULT from the first edition of Johnson’s *Dictionary of the English Language* (1755), manually encoded by the authors. Black underlines indicate abbreviations.

Close analysis of the encoded text also revealed high frequency of error in the form of character substitutions or omissions (see Fig. 3).



p°ilt. ° [defaut, French.]
1. Omiſſion of that which we ought to do ; negle&t.
2. Crime; failure; fault.
Sedition tumbled into England more by the default of go-
vernours than the peoples. Haywood.
We that know what 'tis to faſt and pray,
Are penitent for your default to-day. Shak. Com, of Errours.
Let me not raſhly call in doubt
Divine predićtion: what if all foretold
Had been fulfill'd, but through mine own default,
Whom have I to complain of, but myſelf? Milt. Agoniſtes.
Partial judges we are of our own excellencies, and other
mens defaults. Swift.
3. Defect ; want.
In default of the king's pay, the forces were laid upon the
ſubječt. Davies an Ireland.
Cooks could make artificial birds and fiſhes, in d°ſa. It of
the real ones. Arſuthnot on Coins.
4. [In law.] Non-appearance in court at a day aſſigned. Cowel.

DEFA'ULT. *n. ſ.* [*defaut*, French.]
1. Omission of that which we ought to do ; neglect.
2. Crime; failure; fault.
Sedition tumbled into England more by the *default* of go-
vernours than the peoples. *Haywood.*
We that know what 'tis to fast and pray,
Are penitent for your *default* to-day. *Shak. Com. of Errours.*
Let me not rashly call in doubt
Divine prediction: what if all foretold
Had been fulfill'd, but through mine own *default,*
Whom have I to complain of, but myself? *Milt. Agonistes.*
Partial judges we are of our own excellencies, and other
mens *defaults.* *Swift.*
3. Defect ; want.
In *default* of the king's pay, the forces were laid upon the
subject. *Davies on Ireland.*
Cooks could make artificial birds and fishes, in *default* of
the real ones. *Arbuthnot on Coins.*
4. [In law.] Non-appearance in court at a day assigned. *Cowel.*

Fig. 3. Machine-encoded entry for DEFAULT from the first edition of Johnson’s *Dictionary of the English Language* (1755) (left), compared to the same entry manually re-encoded by the authors to demonstrate more accurate formatting and character encoding (right). Character encoding errors on the left are indicated with red text and black underlines.

While the citation inconsistencies are characteristic to the manuscript itself, the severity of the text encoding issues varies slightly between versions. To select our dataset, we compared three copies of Johnson's *Dictionary* available in the HathiTrust Digital Library: the first edition (1755) [6], the fourth edition (1773) [7], and a facsimile reprint of the first edition (1979) [8]. We eliminated the fourth edition based on our desire to evaluate the dictionary in its original form. After considering the significant age difference between the two remaining texts and performing manual spot-checks for encoding accuracy, we chose the 1979 reprint for processing.

2.2 Defining the Author Corpus

Due to the challenges discussed above, we determined that extracting all authors from the text would not be feasible without first piloting and evaluating a workflow on a smaller test set. Ten commonly-cited authors were selected based on past research on this text, primarily Rüdiger Schreyer's work with *A Dictionary of the English Language on CD-ROM* [9]. To position these authors in the larger history of illustrative quotations, we also cross-referenced Schreyer's list with the Oxford English Dictionary's Top 1000 Sources [10]. We formed our author corpus based on the following criteria:

- **Timeliness:** Johnson endeavored to source authors from the Elizabethan era [11] but made exceptions for a number of Restoration and Augustan authors. Our corpus is limited to writers from those periods.
- **Identifiability:** Some citations are recorded under abbreviations, alternate names, and non-standard spellings too numerous to encapsulate. Our corpus is limited to authors whose citation variations we could fully identify and incorporate into our search program.

2.3 Processing the Dataset

The data was processed programmatically and manually to form a corpus of citation strings labeled with their identified author name, separated by alphabetic section. The cleaning, extraction, and analysis workflow had five general steps:

Text Cleaning. In the encoded text, author names are contained within a single line. Following this structure, we split the text line-by-line into a list of strings using Python. All characters were lowercased and special characters were converted to their English alphabet equivalents to standardize the text, resolving both encoding error (e.g. letters with diacritics converted to equivalents without diacritics) and historical alphabet discrepancies (e.g. long s "ſ" converted to modern "s"). As each string was cleaned, it was added to a new list of line strings for processing.

During our formatting analysis, we found the text lacks consistent differentiators between alphabet sections. We manually inserted the unique string "ALPHASECTEND" between each alphabet section and used it in our program to automatically organize our results and facilitate finer granularity in data analysis.

RegEx Patterns. Johnson’s author name variations impede broad programmatic extraction of citations without significant manual reprocessing of the text. We first considered using the HathiTrust extracted features dataset [12] but found that the proper noun (NNP) tagged token results for our volume contained significant noise and failed to capture all author names. We also concluded that using this dataset would limit applicability of our methods to HathiTrust volumes registered with the extracted features dataset. To maximize accuracy and broader application, we developed flexible regular expression (RegEx) search patterns to capture a range of author name strings while ignoring non-author strings (see Table 1).

Table 1. This table lists the ten selected authors, common in-text representations of those authors, and the RegEx patterns used to identify them.

AUTHOR	COMMON STRINGS	REGEX PATTERN(S)
Joseph Addison	addis., addison, adison, aoison’s, 4ddison, 4dison	[\sadl][da]ison[^a-z] 4d[ido][nis]
John Dryden	dryd., dryden, drydon, orydon, dryder, dryan, dryzon, dryaen, jdryden, dryden’s	ry[^i^\s^n] dry[\s][dl][eo][hnr]
Roger L'Estrange	l'estrange, l'estr., l'esir, l'e/irange, l'e//range, l'e/irange, l’estrange’s	[^a-z^\s]e[si\/][tir\/][r\.]
John Locke	locłe, lecke , locke, locke's	[ldzij\s\.1]oc[klł][eo][^a-z]
Alexander Pope	pope, p pe, p.pe	[^a-z]p[^i]p[ec][^a-z]
Matthew Prior	prior, prier, pricr, prior’s	[\s]pri[zoec]r[^a-z]
William Shakespeare	shak, shakes, shakspeare, shakespeare, shakso, shakoff, shakespeare’s	[\s]sha[okfli][^l][^n]
Philip Sidney	sid, sidney, sidnj, sidn, sidro, sidny, sidhy, sidney’s	[\s]sid[\.nrh]
Edmund Spenser	spensor, spenser, speoser, spessor, sponser’s, spenser's	[\s]sp[eo\s][nos]s[oe\s]
Jonathan Swift	swift, swift’s	[^a-z]sw[li][sft][tf][^a-z^\-^;]

Initial Extraction. Using Python’s standard RegEx library [13], each of our RegEx patterns was compared against each cleaned line string. If a match was detected, the string and its matched author name were added to a CSV file labeled with the letter section. When the program detected the alphabet section differentiator string, a new CSV file was created for the next alphabet section and the evaluation process began again. This process was repeated until every line string was evaluated.

Cleaning False-Positives. Preliminary review of the extracted line strings and their author labels revealed a significant number of false-positive citations, or text erroneously identified by the program as an author name. These false-positives were particularly prevalent among Dryden citations (see Table 2), where our RegEx matched almost any string containing "ry", including material from definitions or quotations (e.g. *country*, *everything*), other authors (e.g. *Atterbury*), and work titles (e.g. Bacon's *Natural History*, Wiseman's *Surgery*).

Table 2. This table illustrates the initial citation counts for John Dryden, final counts after false-positives were removed, and the difference between the two. Letter X is not shown because it contains no definitions, quotations, or citations. I/J and U/V are combined, respectively, as they are in the original manuscript.

	INITIAL	FINAL	DIFF.		INITIAL	FINAL	DIFF.
A	1193	475	-718	**N**	277	137	-140
B	1283	720	-563	**O**	457	267	-190
C	2184	882	-1302	**P**	1573	773	-800
D	1298	623	-675	**Q**	111	45	-66
E	773	313	-460	**R**	909	468	-441
F	1242	645	-597	**S**	2553	1348	-1205
G	785	417	-368	**T**	939	554	-385
H	954	483	-471	**U/V**	950	533	-417
I/J	856	390	-466	**W**	670	389	-281
K	122	68	-54	**Y**	51	30	-21
L	804	448	-356	**Z**	11	6	-5
M	951	481	-470	**TOTAL**	20946	10495	-10451

To filter out these errors, new RegEx patterns were created to sort the extracted author-string pairs into two sub-groups: "Positive" pairs that were certainly correct, and "Review" pairs that required further analysis to determine accuracy (see Table 3) . Each "Review" pair was then manually evaluated and labeled "true" (for correct citations) or "false" (for false-positives). The "Review" author-string pairs labeled "true" were then extracted and combined with the "Positive" author-string pairs to form the final author-string pair corpus.

Final Tally. The authors in the final author-citation pair corpus were tallied and organized by alphabet section. Results were printed to a text file and formatted for readability.

Table 3. This table lists the ten selected authors, the new RegEx patterns used to identify them in the second stage of evaluation, and which sub-group the matching strings were added to. "OTHER" indicates which group non-matching strings were added to. Due to the high volume of Dryden false-positive citations, a RegEx pattern was developed to classify obviously erroneous Dryden citations into an additional "Dud" classification group.

AUTHOR	REGEX PATTERN	CLASSIFICATION
J. Addison	[^a][\sadl][da]ison[^a-z]	Review
	OTHER	Positive
J. Dryden	[^a-z]dryd	Positive
	ry[^d]	Dud
	OTHER	Review
R. L'Estrange	l[^a-z^\s]e	Positive
	OTHER	Review
J. Locke	[^\s][ldzij\s\.1]oc[klł][eo][^a-z]	Review
	OTHER	Positive
A. Pope	pape[^a-z^\s]	Review
	[fea][^a-z]p[^i]p[ec][^a-z]	Review
	OTHER	Positive
M. Prior	[a-z][\s]pri[zoec]r[\.]	Review
	[\s]pri[zoec]r[^\s^\.^\']	Review
	[\.][\s]pri[zoec]r[\.]	Positive
	OTHER	Positive
W. Shakespeare	shak[^a-z]	Positive
	shak[es][sp][peh]	Positive
	shakef	Positive
	OTHER	Review
P. Sidney	sidro	Review
	OTHER	Positive
E. Spenser	[\s]spens	Positive
	OTHER	Review
J. Swift	[\.][\s]sw[li][sft][tf][\.]	Positive
	[a-z][\s]sw[li][sft][tf][\.]	Review
	[\s]sw[li][sft][tf][\s]	Review
	OTHER	Review

2.4 Data Analysis

Our analysis focused on how our authors were represented in the dictionary as a whole and between letter sections. We found that Shakespeare and Dryden together are cited more than all the other authors in our corpus combined. These results are consistent with results produced by Schreyer using the search feature within the *Dictionary of the English Language on CD-ROM* [9]. In addition to analyzing the broader breakdown of authors in our corpus, we separated citations by letter to analyze how the distributions vary between sections (see Fig. 4).

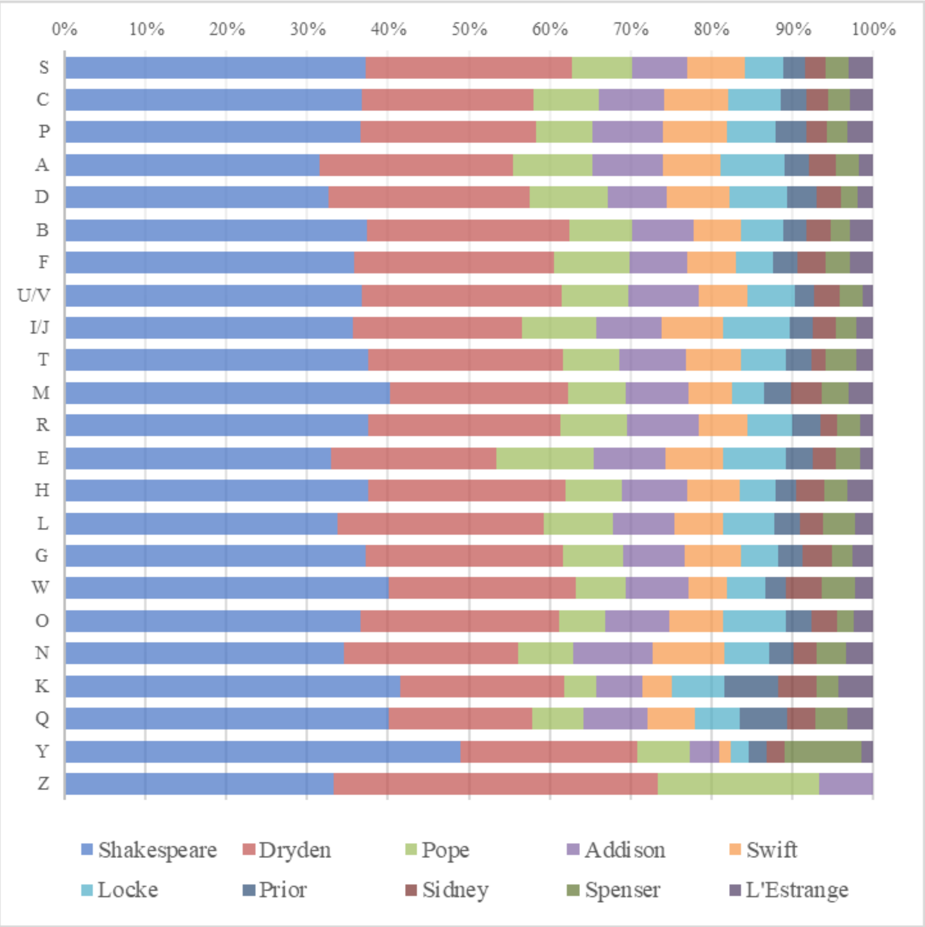


Fig. 4. Proportions of author citations, by letter section. Sections are arranged from most pages (S) to least pages (Z). Letter X is not shown.

We found that the broader proportion trend is consistent across sections – Shakespeare and Dryden make up more than half of the citation counts, and the remaining citations are distributed variably among the other eight authors. However, we also found that fluctuations in each author's proportions are not consistent with the number of pages in each section (see Fig. 5).

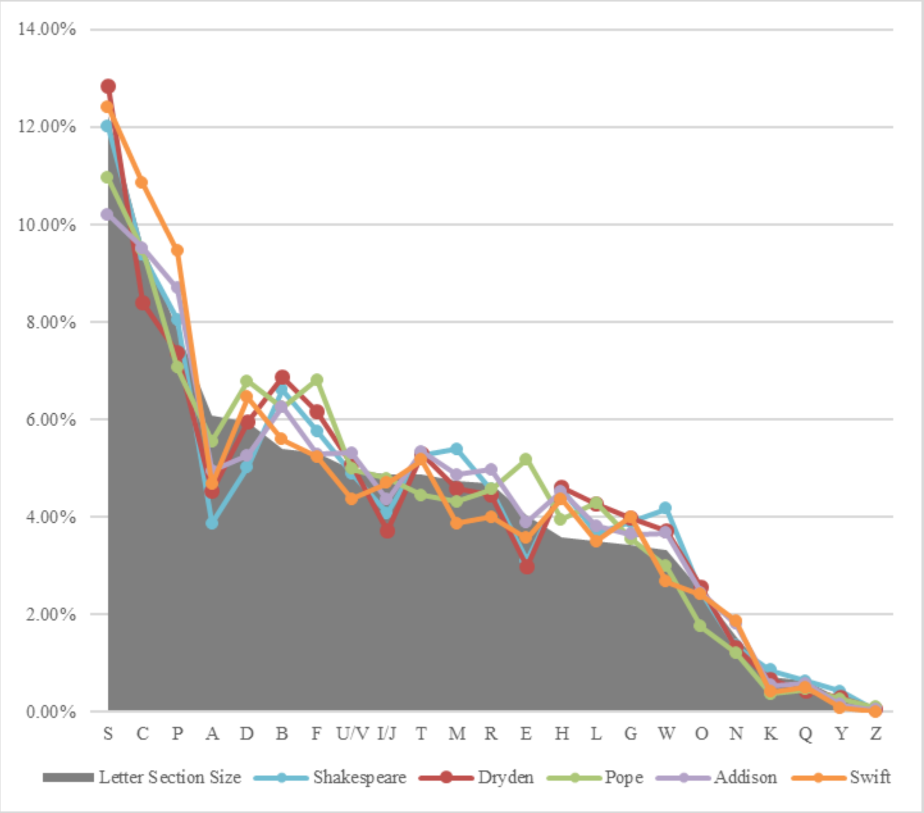


Fig. 5. Citation proportions of top five authors compared to letter section proportions. Author citation proportions are the percentage of that author's citations in that section out of the total number of citations by that author. Letter section size is the percentage of the number of pages in that section out of the total number of pages in all letter sections. Letter sections are arranged from largest (S) to smallest (Z). Letter X is not shown.

For example, letter B makes up 5.4% of the dictionary (122 pages) but contains a disproportionate 6.61% of Shakespeare's collected corpus (1075 citations). Conversely, letter A takes up slightly more space (137 pages, or 6.07% of the dictionary) but holds little more than half the number of Shakespeare citations as letter B (629 citations, or 3.87% of Shakespeare's collected corpus). Further study is needed to determine what other authors make up the other citations, and why the selected corpus was cited less in these sections.

3 Conclusions and Future Work

Our study illustrates that Samuel Johnson showed a significant preference for particular authors when compiling his dictionary, which is supported by other studies exploring the same topic [9]. This positive benchmarking supports the potential for broader application of our workflow to other machine-encoded dictionary texts (the code developed for this project is available openly on GitHub: https://github.com/asnowmenjig/OIDLPP), and indicates that further investigation into Johnson's motivations for citation selections is warranted.

In the absence of corrected OCR and a comprehensive tagged features dataset, we chose to identify authors from a defined corpus rather than build a corpus from the text. Our workflow maximizes precision and recall for the authors in our selected corpus but is not generalizable for high recall of all authors in the dictionary. However, the RegEx patterns we developed are portable and applicable for author identification in other historical dictionaries.

In the future, we hope to improve this workflow to refine accuracy, broaden its applicability to a larger corpus of authors, and enhance extraction to include full citations. Though manual identification and extraction of authors and citations has a high level of accuracy, it is time-intensive and reliant on consumptive access to the full dictionary volumes. A programmatic approach as piloted in this project helps to overcome both of these non-trivial challenges with the added utility of machine-readable outputs that enable further computational and manual analysis. With a comprehensive list of authors and citations, questions about representation, bias, and impact on the Western literary canon could be posed.

References

1. Green, J.: Chasing the Sun: Dictionary Makers and the Dictionaries They Made. Jonathan Cape, London (1996).
2. Brewer, C.: Treasure-House of the Language: The Living OED. Yale University Press, New Haven (2007).
3. The Plan of a Dictionary of the English Language (1747), https://johnsonsdictionaryonline.com/history-of-johnsons-dictionary/the-plan-of-a-dictionary-of-the-english-language-1747, last accessed 2019/9/8.
4. 1755 - Johnson's Dictionary, http://www.bl.uk/learning/langlit/dic/johnson/1755johnsonsdictionary.html, last accessed 2019/9/8.
5. HTRC Data Capsule - Documentation - HTRC Docs, https://wiki.htrc.illinois.edu/display/COM/HTRC+Data+Capsule, last accessed 2019/9/6.
6. Catalog Record: A dictionary of the English language: in which the words are deduced from their originals, and illustrated in their different significations by examples from the best writers, to which are prefixed, a history of the language, and an English grammar | HathiTrust Digital Library, https://catalog.hathitrust.org/Record/009310086, last accessed 2019/9/6.
7. Catalog Record: A dictionary of the English language: in which the words are deduced from their originals, and illustrated in their different significations by examples from the

best writers. To which are prefixed, a history of the language, and an English grammar | HathiTrust Digital Library, https://catalog.hathitrust.org/Record/100218658, last accessed 2019/9/6.

8. Catalog Record: A dictionary of the English language | HathiTrust Digital Library, https://catalog.hathitrust.org/Record/100153694, last accessed 2019/9/6.
9. Schreyer, R.: Illustrations of Authority Quotations in Samuel Johnson's Dictionary of the English Language (1755). Lexicographica 16, 58-103 (2000).
10. Top 1000 sources in the OED, http://www.oed.com/sources, last accessed 2019/9/8.
11. Preface, https://johnsonsdictionaryonline.com/preface, last accessed 2019/9/8.
12. Extracted Features Dataset - Documentation - HTRC Docs, https://wiki.htrc.illinois.edu/display/COM/Extracted+Features+Dataset, last accessed 2019/12/15.
13. re - Regular expression operations - Python 3.7.4 documentation, https://docs.python.org/3/library/re.html#module-re, last accessed 2019/9/6.