OXFORD

Women's Words and the Words of Women in the *Oxford English Dictionary*

David-Antoine Williams (iD)

This article explores datasets curated from the citation evidence in successive editions and revisions of the *Oxford English Dictionary* (1884–2022), which have been annotated to reflect the gender of the authors and other bibliographical metadata. This exploration aims both to supplement the historical account of the dictionary's uses of female-authored quotation sources, correcting and elaborating some figures which have previously been reported, and to provide a contemporary account of women's representation in *OED Online*, using the revision published in June 2022. In seeking to establish a more objective and empirical basis for judging 'representativeness', I treat the OED both as a self-contained bibliographical and lexicographical work, and comparatively, against other comprehensive or very large bibliographical corpora, namely the Garside et al. surveys of early English novels, the Library of Congress Catalog, and the HathiTrust Digital Library. The OED data studied here represents a significant (if restricted) subset, rather than a representative sample, of the OED corpus as a whole: modern (post-1700) quotations from books appearing with their author's name in the OED evidence are considered. While this approach does not claim to make an objectively complete tally of every woman-authored quotation collected in the OED, it does enable a more detailed and accurate account than has previously been possible, and allows for a number of consistent cross-comparisons. A companion document of Supplementary Data & Notes, available at *The Review of English Studies* online, describes in technical terms how the data was compiled and the processes and principles by which it was annotated.

Pip Williams's popular 2020 novel *The Dictionary of Lost Words* tells the story of Esme Nicoll, the fictional daughter of an (also invented) sub-editor of the *New English Dictionary* (OED1, 10 vols, 1888–1928, later the *Oxford English Dictionary* [1933]). A child of James A. H. Murray's Scriptorium in many ways—her sixth birthday is celebrated at the Murray residence at 78 Banbury Road, on the same day as the publication party for Volume I (A–B)—Esme's coming-of-age involves the realization that certain words and senses are being left out of Dr Murray's great work, not as a result of sound editorial judgement, or even chance oversight, but because of the social and intellectual biases of its compilers. In response, she goes about recording examples of the language of persons she encounters about town, principally illiterate or uneducated women of

Supplementary data for this article are available on *The Review of English Studies* website.

the English working classes. This is everyday language, sometimes coarse, pertaining to the family, social, and occupational lives of these women, ranging from words about the body, sex, and pregnancy (among them 'clitoris', 'quim', 'cunt', and 'fuck'; the latter two controversially excluded from OED1 on grounds of obscenity) to words referring to gendered social roles and statuses, including 'dollymop', 'game', 'trade' (all three terms related to prostitution), 'latch-keyed', 'sisterhood', 'suffragette', and 'suffragent'. Also in this latter group is 'bondmaid', a key word in the novel, as Williams says, signifying duty and devotion (to a person, or a cause, or even a dictionary) but also sacrifice, servitude, and even subjugation.[1]

Beyond the prejudices of certain men who inhabit this story—two of them tell Esme point blank that her project is 'of no importance', or 'of no scholarly importance'—Esme's gathered usage evidence would have been inadmissible to the OED, because it was sourced in the field (as it were), from *viva voce* interactions, rather than from printed texts.[2] Indeed, a central concern of the novel is the application of OED inclusion standards, the fictionalized Edith Thompson complaining in a letter to Esme that 'literately' will likely not appear in the dictionary because its 'lady author has not proved herself a "literata"—an abomination of a word coined by Samuel Taylor Coleridge'.[3] The inclusion of Coleridge's condescending neologism ('The young lady is said to be the most literary of the beautiful, and the most beautiful of the literatæ') is assured because of who coined it, it is implied, but Thompson doubts the word will gain widespread currency, surmising that 'the number of literary ladies in the world is surely so great as to render them ordinary and deserving members of the literati'.[4]

In the personage of Esme Nicoll, Williams's novel dramatizes the author's sense that 'A lack of representation' of women editors, compilers, and quoted authors 'might mean that the first edition of the *Oxford English Dictionary* was biased in favour of the experiences and sensibilities of … Older, white, Victorian-era men'.[5] This is a sense reflected in, and indeed partially informed by, current scholarship in lexicography and dictionary studies. A number of examples of Esme's 'lost words'—including 'bondmaid', 'literata', and 'literately'—are taken from Lynda Mugglestone's chapter of that name, in *Lost for Words: The Hidden History of the Oxford English Dictionary* (2005).[6] There Mugglestone discerns a 'certain pattern of selection' in the pruning of entries in later proofs of OED1, where, for example, the idiosyncratic usages of Elizabeth Griffiths (author of 'literately') and Adeline Whitney are discarded, while those of Coleridge and Ralph Waldo Emerson are retained.[7]

The cultural prestige of these authors cannot be disentwined from the question of gender, a 'fault line in language and culture as reflected – and endorsed – within the *OED*', as Mugglestone says, which filters its imperialist accounts of English through 'distinctly male-as-norm ideologies'.[8] More generally, Lindsay Rose Russell has described a 'scholarly consensus … that mainstream lexicography, past and present, is shot with sexism and androcentrism'.[9] With respect to OED1 specifically, Russell records the un- and underacknowledged work of women who contributed to the project in various official and unofficial capacities, while also describing

[1] Pip Williams, *The Dictionary of Lost Words* (New York, NY, 2021 [2020]), 363.
[2] Williams, *Lost Words*, 242, 338.
[3] Williams, *Lost Words*, 51–2. Written into the novel as Esme's godmother, the real Edith Thompson (1848–1929), author of a *History of England* (1873), worked on several aspects of the OED. She is the only woman among the eight major contributors to the First Edition with a biography on the 'Contributors' page at *OED Online* (<https://web.archive.org/web/20211119205958/https://public.oed.com/history/oed-editions/contributors/>), and one of forty-odd women in an expanded list of over 300 contributors (<https://web.archive.org/web/20210509083210/https://public.oed.com/history/oed-editions/contributors/biographical-information/>).
[4] Williams, *Lost Words*, 52.
[5] Williams, *Lost Words* ('Author's Note'), 362.
[6] See Lynda Mugglestone, *Lost for Words: The Hidden History of the Oxford English Dictionary* (New Haven, CT, 2005), 82, 97; and more generally Chapter 3, 'Lost Words', 70–109.
[7] Mugglestone, *Lost for Words*, 97.
[8] Mugglestone, *Lost for Words*, 166.
[9] Lindsay Rose Russell, *Women and Dictionary Making: Gender, Genre, and English Language Lexicography* (Cambridge, 2018), 15.

ways in which, with its combination of 'andro- and ethnocentrism, the OED reflects trenchant ideologies about Englishwomen'.[10] Amid his broad critique of the ideologies underwriting the OED enterprise, John Willinsky argues that the 'expressly masculine citational authority of the OED' exemplifies the way in which 'the selective traditions of canons and dictionaries, especially citational dictionaries, can exacerbate prejudices, giving these prejudices a greater veracity by further restricting access to what is already an uneven playing field'.[11] The exclusion from the OED's lexicographical record of writings by women 'appears to be the dictionary's largest sin of omission', he writes, with 'roots in the general regard of the masculine hold over artistic creation, philosophical speculation, scientific inquiry, political theory, and so on'.[12]

A large part of the research in this area, especially as it pertains to later supplements and editions, is by Charlotte Brewer. In a number of articles, she has pressed points of statistical 'representation', writing that, 'certainly there is no question but that the lexicographers favoured male over female sources, and this bias will undoubtedly have affected the nature of the language witnessed in the quotations'.[13] Compared to men, Brewer later writes, '*OED* quotes comparatively few women authors (even from the last decades of the eighteenth century, when a third of the novels published were written by women)', and when women were to be quoted, the OED 'often favoured distinctive usages in female-authored texts—innovative, eccentric or domestic vocabulary—rather than usage which exemplified linguistic norms'.[14] This occurred, Brewer writes, because 'the first edition … generally preferred [sources] written by men to those written by women', with the 'degree of bias … still clearly evident in the OED we consult online today'.[15]

Indeed, perhaps provoked in part by such arguments, contemporary OED staff have themselves been cognizant of the treatment of female sources, usually at the hands of their predecessors. Today's *OED Online*, the web portal to the dictionary, describes a modern Reading Programme (to source usage evidence) that specifically 'covers women's writing and non-literary texts which have been published in recent times, such as wills, probate inventories, account books, diaries, and letters'.[16] In an editorial note to a recent *OED Blog* post by Pip Williams herself, OED staff seek to distance current views and practices from those Williams fictionalizes in *The Dictionary of Lost Words*:

> In the years that followed the conclusion of the first edition of the OED, the position of women in society has changed … And the OED has sought to reflect that. Definitions and examples have been updated. New words have been added. Stereotypes and generalizations have been questioned, challenged, and corrected. The dictionary is, and always will be, a living document, reflecting the way in which language is used and the biases that exist around us but, if Esme were to see us now, perhaps she would have fewer words to capture for her own dictionary.[17]

It is an upbeat self-evaluation which, however measured (one notes the tempering 'fewer' and 'perhaps'), cries out to be tested.

---

[10] Russell, *Women and Dictionary Making*, 150–66, 5.
[11] John Willinsky, *Empire of Words: The Reign of the OED* (Princeton, NJ, 1994), 183.
[12] Willinsky, *Empire of Words*, 186. See Willinsky's case study of the citational basis for OED1's definition of WOMAN, for him a prime example of how 'the dictionary lends its weight to creating a natural history of the English language, a natural history constituted by the inequity and exclusionary nature of the citations' (187).
[13] Elizabeth Baigent, Charlotte Brewer, and Vivienne Larminie, 'Gender in the Archive: Women in the *Oxford Dictionary of National Biography* and the *Oxford English Dictionary*', *Archives: The Journal of the British Records Association*, 113 (2005), 13–35, 27–8.
[14] Charlotte Brewer, 'The Use of Literary Quotations in the *Oxford English Dictionary*', *RES*, 61 (2010), 93–125, 105; Charlotte Brewer, '"Happy Copiousness"? OED's Recording of Female Authors of the Eighteenth Century', *RES*, 63 (2012), 86–117, 86.
[15] Charlotte Brewer, '"That Reliance on the Ordinary": Jane Austen and the *Oxford English Dictionary*', *RES*, 66 (2015), 745–65, 747.
[16] OED Online, *Reading Programme* (n.d.) <https://web.archive.org/web/20210307232923/https://www.oed.com/page/reading/Reading$0020Programme/>.
[17] Editorial note to Pip Williams, *Women of Words: A Journey Through the Archives of the OED*, in *OED Blog* (12 April 2021) <https://web.archive.org/web/20210421193112/https://public.oed.com/blog/women-of-words/>.

Regrettably, testing claims about gender rigorously is not something even the adept manipulator of *OED Online* can do. This is because the gender of OED authors is neither evident in the citations themselves (typically initials are used for given names, and the gendered honorifics often employed in previous editions are, rightly, no longer used), nor recorded in underlying or associated data, a situation rued by Brewer and others, even as they found other ways of approaching the question.[18] These approaches typically analysed small subsets of OED quotations, compiled from individual highly cited authors; or from tranches of these, such as the 'Top 1000 sources in the OED' page at *OED Online*; or from narrow time periods; or by random or haphazard sampling; or by variously motivated case studies.[19] In each case 'representativeness' is primarily an impressionistic rather than a statistical category, however numerical the presentation of the discussion may be. It is a feature that Brewer repeatedly acknowledges amid her calls for future research and improved access to OED data.[20]

In this article, I explore datasets curated directly from the encoded texts of three editions of the OED, which have been annotated for the gender of the quotation author and, in a smaller subset of cases, for the Library of Congress subject classification of the work.[21] To focus the analysis on questions of author 'representation' (in the various senses elaborated below), this data is restricted to modern (post-1700) quotations from books that appear with their author's name represented in the OED evidence. It will be evident that such parameters exclude significant numbers of quotations from a variety of genres and publication types—from early anonymous novels and periodical fiction (when cited as such), to articles in scientific and other learned journals, to newspaper and magazine articles—on the basis that to some degree such works have (or are represented in the dictionary as having) either no identifiable authorship, or a sort of corporate authorship not always readily identifiable with one or a few individuals. How the 'real' gender of authorship is distributed within OED's citations of these genres and types can only be a matter of speculation, influenced perhaps by those quotations for which an author is identified (the subject of this study).[22]

The data studied here is also therefore not a 'representative sample', but rather a specific subset of the OED quotation corpus as a whole, stretching across its three editions and various revisions. It is nonetheless a large subset (0.72, 0.99, and 1.41 million quotations in OED1, OED2, and OED3, respectively) corresponding to a defined idea of what an authored source is, and is within its chronological and generic parameters essentially complete, with 97 to >99 per cent of quotations marked up with author gender or genders (depending on the edition or revision in

[18]　As early as 2005 Brewer and her co-authors foresaw that it would be 'very difficult' to judge the effects of OED3's revised policies on source materials on gender representation 'unless the authors are to be tagged by gender' (Baigent et al., 'Gender in the Archive', 31). Though they reported at the time that this 'is something the editors are now actively considering for future entries', to date no such work has appeared at *OED Online*. Brewer noted as much in 2010, complaining how unsatisfactory were the raw numbers that could be derived from 'individual random searching' (Brewer, 'Literary Quotations', 116), and again in 2012, where she further noted two additional technical limitations of *OED Online*: the inability (at that time) to search for first attestations of senses (Brewer, 'Female Authors', 93), and the impossibility (which persists today) of determining systematically the edition in which a quotation or definition was added or amended (Brewer, 'Female Authors', 89).

[19]　OED Online, *Top 1000 sources in the OED* <https://www.oed.com/sources/> accessed 5 May 2023. This list, which changes periodically, is behind the *OED Online* paywall, and so cannot be archived as its public pages can. The Baigent et al. study ('Gender in the Archive') demonstrates among many useful things the pitfalls of manually parsing unsystematic OED data, as the authors fail to register among their top twelve Ann Radcliffe (at 1,123 quotations she is the eighth most quoted woman in OED2), and seriously undercount Charlotte Brontë's quotations, presumably because they did not include the 314 OED2 citations of 'C. Bronte' with their 698 references to 'C. Brontë'. Also left out is 'Ouida', aka Maria Louise Ramé, aka Marie Louise de la Ramée, who at 797 OED2 quotations ought to have made their list.

[20]　Recently Oxford Dictionaries developed a prototype application programming interface (API) to query OED data, which as of this writing can be accessed, with registration, at <https://languages.oup.com/research/oed-researcher-api/> accessed 5 May 2023. The API allows for some author filtering according to gender, based on curated data for about 15,000 highly cited authors, plus automatically inferred gender data of variable quality for an unspecified number of others. The present article does not draw on this source.

[21]　All OED data is published by Oxford University Press.

[22]　Women's writing published anonymously is per se a topic of ongoing scholarship. For discussions of several authors featuring here (though cited by name in OED), and periodicals important especially to OED1, see Alexis Easley, *First-Person Anonymous: Women Writers and Victorian Print Media, 1830–70* (London, 2004).

question), or as undeterminable (<3 per cent), by human researchers. A companion document of Supplementary Data & Notes, available at *The Review of English Studies* online, describes in technical terms how the data was compiled, the processes and principles by which it was annotated with this additional metadata, and the quantitative profiles of the resulting dataset and its relevant subsets.

In presenting analyses of this data here I aim both to supplement the historical account of the First Edition and Supplements' uses of female-authored quotation sources, correcting and elaborating some figures which have previously been reported, and to provide a contemporary account of women's representation in *OED Online*, using the revision published in June 2022 (OED3). In seeking to establish a more objective and empirical basis for judging 'representativeness', I treat the OED both as a self-contained bibliographical and lexicographical work, and comparatively, against other comprehensive or very large bibliographical corpora, namely the bibliographical surveys of early English novels by Garside et al. (ENBS), the Library of Congress Catalog (LOC), and the HathiTrust Digital Library (HATHI). As previous studies have, here I do measure the OED's treatment of frequently cited female-authored sources against its treatment of frequently cited male-authored sources, but in the main I look across the entire population, from the most- to the least-cited, dividing and grouping it in different ways to identify dominant trends and their constituent sub-patterns.

A BIGGER PICTURE

This comprehensive approach enlarges the scope of attention by orders of magnitude, due to the uneven distribution of quotations and authors in the dictionary. In OED3, for instance, 17,600 women authors (59 per cent) are quoted only once, together accounting for 10 per cent of female-authored quotations. Conversely, the ten most-cited female authors in OED3, each quoted between 1,200 and 3,340 times, would similarly account for 10 per cent of female-authored quotations. However, these ten authors would make up just four one-hundredths of one per cent of all female authors, leaving out more than 29,600 names.

A salient example of the effects of scope in making quantitative judgements is the idea quoted above that 'OED quotes comparatively few women authors (even from the last decades of the eighteenth century, when a third of the novels published were written by women)'. The figure for novels comes from a comprehensive bibliographical survey of 3,677 novels published in the British Isles between 1770 and 1829, compiled by Garside, Schöwerling, et al. (2000), later extended to 1836 (+610 works) by Garside, Mandel et al. (2016 [2003]), and revised and enlarged periodically in online reports (here referred to collectively, and with some modifications, as ENBS; see Supplementary Data & Notes §§2.7, 3.3 for further details). As the first print bibliography reports, in the four decades from 1780 to 1819, of the titles where an author gender could be identified or inferred, between 32 and 64 per cent of newly published novels in each decade were by women.[23]

By matching OED citations to these bibliographies, we can arrive at a complete quantitative picture of OED quotation gender in this specific genre during this limited early period. The graphs in Fig. 1a and b show the percentage of ENBS novels quoted in the 1989 Second Edition (OED2), and in newly added OED3 quotations (OED3n), broken down by date of publication and author gender, while Fig. 1c and d compare the rates of female authorship in ENBS, OED2 quotations, and OED3n quotations. From these figures we can assemble a much

---

[23] A significant proportion of early titles is not attributable to any gender: 30–55 per cent in 1770–1799, down to 17 per cent on average between 1800 and 1829, according to the original analysis in Peter Garside, Rainer Schöwerling, et al., *The English Novel, 1770–1829: A Bibliographical Survey of Prose Fiction Published in the British Isles*, 2 vols (Oxford, 2000) 1. 46–9, and 2. 72–6. Many of the unattributed titles in the early print editions were identified in subsequent updates, which have been incorporated into ENBS.
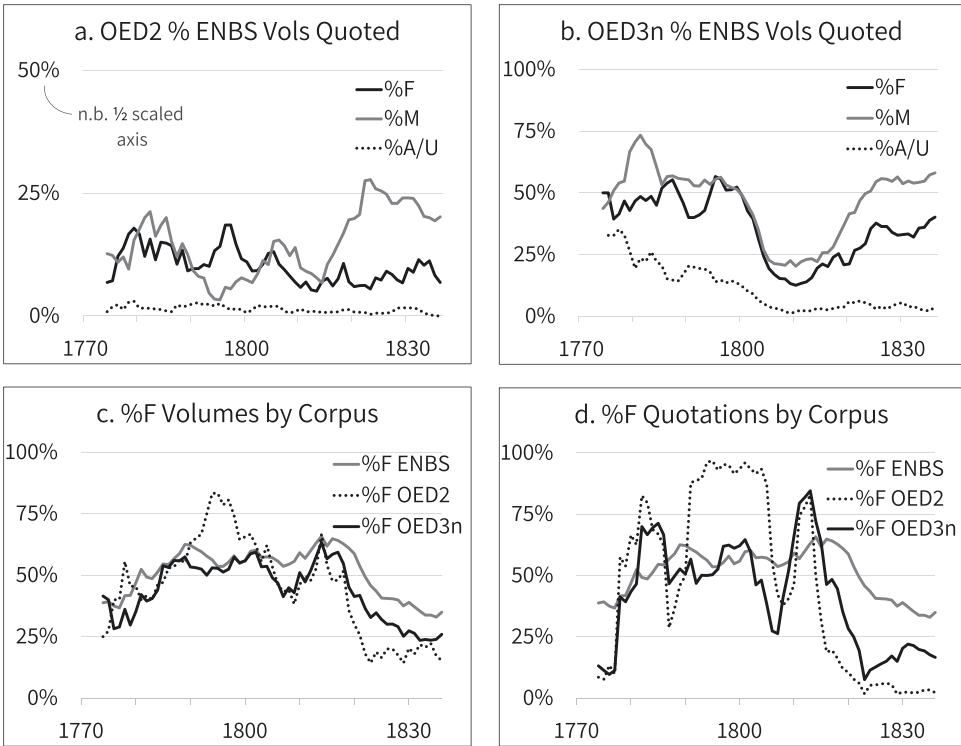
FIG. 1. Author gender of English novels 1774–1836, in OED quotation evidence vs ENBS (trailing 5-year average).

more detailed—and convoluted—idea of the OED's rates of citation, and by-the-by make a demonstration of the pitfalls of generalization.

That is, just because many of the OED's most heavily cited sources are novels does not mean that its citations of novels are at all typical of the quotation corpus as a whole: the 28,100 OED2 quotations from sources recorded in ENBS are only 10 per cent of all OED2 quotations 1770–1836; and only 7 per cent of quotations in OED3n from this period are from these novels. Moreover, these novels are only a fraction of those recorded in ENBS: 11 per cent overall, and never more than 28 per cent in any five-year period for OED2; and about two fifths overall (37 per cent) for OED3 (peaking at 56 per cent in 1776–1780, with 53 of 94 ENBS novels cited). Where along the overall gender distribution curves for any population—whether ENBS or OED, fiction or all texts—any of these subsets lies is impossible to know a priori, even if they can be assumed to include a large amount of Walter Scott, Jane Austen, and Maria Edgeworth.

What Fig. 1 illustrates above all is that a relatively small and substantially predetermined OED subcorpus such as 'quotations from early English novels' can be highly sensitive to individual publication events by a handful of greatly esteemed authors. In times when women novelists such as Austen or Edgeworth or Ann Radcliffe or Frances Burney publish, the OED evidence draws disproportionately on books by these authors (compared to other novels), and the ratio of female-authored quotations rises in consequence. The opposite occurs when John Galt, or Frederick Marryat, or especially Scott publish, mostly towards the end of the timeframe in question. In both cases this exaggerates the gendered proportion of OED quotations vis-à-vis that of actual

novels in publication. In OED2, for instance, over 90 per cent of attributable quotations from novels published between 1785 and 1805 are by women, and half of those are by Edgeworth, Burney, or Radcliffe; whereas after about 1820 this figure falls below 5 per cent (Fig. 1d), with 46 per cent of all quotations from novels in the period drawn from the works of Scott alone. The proportion of female-authored volumes quoted in OED3n cleaves more closely to ENBS (Fig. 1c), because OED3 in general sources quotations more broadly than OED2 did, but the exaggeration effect on the number of quotations from those volumes is still apparent, especially in the later part of the timeframe (Fig. 1d).

However, as this example illustrates, what is true of small subcorpora and narrow timeframes is not generalizable to the broad corpus of OED quotations. Because the data are more generally obtainable, quantitative analyses of OED evidence have in the past tended to centre its superlatives, discussing the most-quoted sources, the most words or senses coined, the most prolific authors and periods, and so on. While such a focus may tell us something about the OED's treatment of a particular genre or period, or a certain class of author—usually, the very well-known literary author—it can hardly be taken to characterize OED's attitudes towards a population as diverse as 'women authors writing in English'. Indeed, in concentrating on only the most relied-upon authors, such analyses may be amplifying the underlying androcentrism of the dictionary, while at the same time obscuring more subtle patterns of bias and their causes.

The top ten or twenty most-cited women authors in OED1, for instance, are likely to be familiar names to readers of a certain education. Although in OED1 they do represent significant shares of quotations by women—28 and 40 per cent, respectively—they make up less than 1 per cent of all women authors quoted there (see Supplementary Data & Notes §3.2). For later supplements and editions, the utility of a 'top-n most-cited authors' approach diminishes considerably at the quotation level as well: the top ten female authors added in SUP1, SUP2, and OED3n represent 21, 12, and 5 per cent of quotations by women in those corpora, respectively. Beyond raw rankings, comparing these figures to the corresponding top-n male-authored sources naturally leads to distorted outcomes, given that, for example, the twentieth most-cited woman in the current OED3 (Charlotte Smith, 1749–1806) is only about the 477th most-cited person overall, not counting periodicals and unattributed sources (thus she comes 885th on *OED Online*'s current list of 1,000 most-quoted sources, which does include these). While notable in and of itself, this discrepancy only refers to one extremity of the distribution, and cannot be generalized.

REPRESENTING THE WORDS OF WOMEN

Williams's implicit and Brewer's explicit critiques of the OED's treatment of women's writings may be seen as diametric but complementary approaches to the question of 'women's words': whereas Williams's Esme Nicoll finds common women's words and senses missing from the dictionary, Brewer finds the words of women missing from the dictionary's common words and senses. That is, there are two converse ways in which any distinct population of authors may be more or less 'represented' across OED quotation evidence. One is in the dictionary's treatment of restricted or domain-specific vocabulary—words which are disproportionately used by those authors in particular contexts—such as Esme's fictive recording of a midwife's use of 'lie-child', or Brewer's impression of a preponderance of 'innovative, eccentric or domestic vocabulary' in women-authored OED quotations.[24] The other is in the equitable use of the words of women to illustrate the unremarkable words and senses that make up the vast majority of language as it is usually used—Brewer's 'linguistic norms'. The distinction so put plays on the difference between

---

[24]   In fact, 'lie-child' was identified by OED1 staff, not in the spontaneous speech of a midwife but in an essay by the oft-quoted literato Charles Lamb (cited 1,812 times in OED1). It was written up, but cut during editing (see Mugglestone, *Lost for Words*, 108), and does not to this day appear anywhere in the dictionary.

'words' understood metonymically to mean a saying or utterance (OED3, 2.b; in this case of written statements) and 'words' as lexical items forming the 'basic units of meaningful speech' (OED3, 12) recorded by dictionaries and the like, and refers in statistical terms to the centre and edges of any distribution.

These two categories of 'representation'—that of the common and of the restricted vocabulary—naturally converge and overlap, beginning in the shadier realms of connotation, where OED's quotation evidence is most functional and powerful, and extending into the comparatively bounded territory of definition. This is because common words often have restricted or domain-specific senses, usages, shadings, and implications, which may or may not be recognized in the dictionary. The definition in SISTERHOOD that seems not-quite-right to Esme is OED1 sense 2.b—'Used loosely to denote a number of females having some common aim, characteristic or calling. Often in a bad sense.' It is supported by a number of quotations by men (and one by a woman) disapproving of 'Canting Females', who 'agitate questions they know nothing about' with 'boldness and effrontery', i.e., who make common cause outside the approved structure of the convent, against which this sort of sisterhood is being ironically compared. One notes that the emendation in SUP2—'Recently also *spec.* of feminists'—is a precision but not an improvement. There the latest quotation, from the *Times* (1981, unattributed; the writer is Irving Wardle), refers with mock foreboding to 'what the sisterhood is now brewing up'.

Is Williams's fictive use of 'sisterhood', in a positive sense in 1912, therefore no more than an anachronism introduced by a latter-day sister (in a book which, it may be admitted, contains a number of apparently inadvertent linguistic anachronisms)? It is a discrete lexicological point which I cannot settle here, though it is worth noting that OED3 revised the definition so as to be nonpejorative in June 2018. The larger issue raised by this example is one of authority and authoritativeness, and the direct effects of usage connotations on authoritative definition. On the authority of OED1 and OED2, based on the usages of Philip Massinger, Scott, Wardle et al., 'sisterhood' in the extended sense *is* depreciatory in this period. Even if the OED3 revision has added two neutral earlier and later quotations to this sense, the bulk of OED2 quotations remain firmly seated in between.

It should be said that of the two types of 'representation', the question of the inclusion of restricted vocabulary is more obviously and directly a matter of lexicographical policy, especially in a comprehensive historical dictionary such as the OED. I will discuss this matter further in the next section. The question of representation in the sense of equitable presence, visibility, or prominence among the general mass of OED quotations is perhaps not so obviously disciplinary, at least as the discipline has traditionally been construed. Indeed, responding to Brewer's claims of underrepresentation, John Considine has argued that 'representative sampling was never the business of the OED or of any similar dictionary'; whereas 'The makers of a corpus which was meant to provide a representative sampling of the written English of a given period would have to decide very carefully what proportion of poetry, or of female-authored texts, was to be included', Considine says, such 'is not a question for lexicographers'.[25]

I think this defence too categorical. Even if we own the underlying point that demographically proportional representation is not a lexicographical matter per se, it becomes lexicographically important insofar as quotations are the starting points to definition and construe connotative range in an inductive dictionary such as the OED, and inasmuch as it is plausible that distinct populations will employ vocabulary in contexts and with connotations sufficiently distinct to affect sense division and definition, which seems all the more possible if their private and public roles, occupations, and pastimes are broadly differentiated from those of other populations.

[25] John Considine, 'Literary Classics in *OED* Quotation Evidence', *RES*, 60 (2009), 620–38, 632.

Representation is thus a question both of the dictionary's construction as well as its presentation. But the further point, regarding the gendered impression given of the history of English as a written language, while not strictly lexicographical, is not therefore to be discounted.

Still, one is left with the practical difficulties raised by Considine, as well as the question in principle of 'how many quotations would need to be changed in order to give a balanced impression of the people who use English?'[26] It is a problem encountered by all who have thought seriously on the topic. Baigent et al. frame it this way: 'How are the lexicographers to determine the appropriate proportion in which women … should be quoted? Should the proportion reflect the number of women authors in print as compared to men? Or the number of women writers (in whatever genre), in print or not, compared to men? Or the number of women speakers compared to men?'[27] Clearly, approaching any of these benchmarks would entail a certain amount of additional work and verification, whether taken as a radical remedial measure for existing OED evidence—an Augean task, a quantified sense of which is given below—or as a standard or an ideal for future work on OED3, which is, now over twenty years into its complete revision of OED2 material, only at the half-way mark.[28]

In order to give a historical account of the representation of women in the OED's citation of books for the period 1700–2022, I raise in the background two curated English language book corpora: the Library of Congress Catalog (LOC) and the HathiTrust Digital Library (HATHI). These corpora, which are further described in the Supplementary Data & Notes §2.6, were selected as plausible models for the somewhat nebulous category of 'what books were available to Oxford lexicographers and their informants at the time'—a version of the least stringent of Baigent et al.'s suggested benchmarks of representativeness—because the corpora are very large (9.4 million and 17.1 million records, respectively), and include the kinds of specialized volumes important to the OED, which only large university or national libraries will hold; are historical, with collections in the eighteenth century, and large collections from the early nineteenth; are replete with high-quality metadata, having been catalogued by librarians over decades; and are available in full as open-access datasets.[29]

Fig. 2 presents historical graphs showing the percentage of female-authored items (quotations or sources), by year of publication, for the following corpora: the 1928 *New English Dictionary* (OED1, Fig. 2a); the 1972–1986 Second Supplement (SUP2, incorporating 1986–1989 additions, Fig. 2b), isolating newly sourced quotations (2b.i) and those retained from the 1933 First Supplement (SUP1, 2b.ii); and the June 2022 revision of *OED Online* (OED3, incorporating the 1994–1997 Additions Series) with newly sourced (OED3n) quotations shown in Fig. 2c, split between those added to revised (2c.i) and new (2c.ii) entries, and all quotations (OED3a) in Fig. 2d; plus volumes in LOC (Fig. 2e) and HATHI (Fig. 2f), each including subsets that correlate with literary texts. At first glance the various plots show largely similar curves, especially when smoothed to a 25-year average. They generally begin at or below 1 per cent female authorship in the early part of the eighteenth century, and slowly rise towards, and eventually above, 10 per cent by the early years of the twentieth. The principal curves with data beyond this point then tend to rise gradually towards 20 per cent until about the three-quarter mark of the twentieth century, and then more rapidly, to 30 per cent and somewhat beyond, in the final thirty to forty years.

---

[26] Considine, 'Literary Classics', 631.

[27] Baigent et al., 'Gender in the Archive', 30.

[28] In a series of blog posts in 2020 I provided various measures of OED3's progress and projected completion. See posts beginning with David-Antoine Williams, 'More Precisions on Revisions', *The Life of Words* (blog) (15 January 2020), <https://web.archive.org/web/20220705223941/https://thelifeofwords.uwaterloo.ca/precisions-on-revisions/>.

[29] One limitation of these two corpora is that they mostly contain books in libraries located in the United States. However, while including a collection from a British legal deposit library (none of which are accessible *holus bolus* at this time) might conceivably have yielded additional overlap with OED sources, this is substantially mitigated by the size and scope of the American collections (see Supplementary Data & Notes, §§2.6, 3.4).
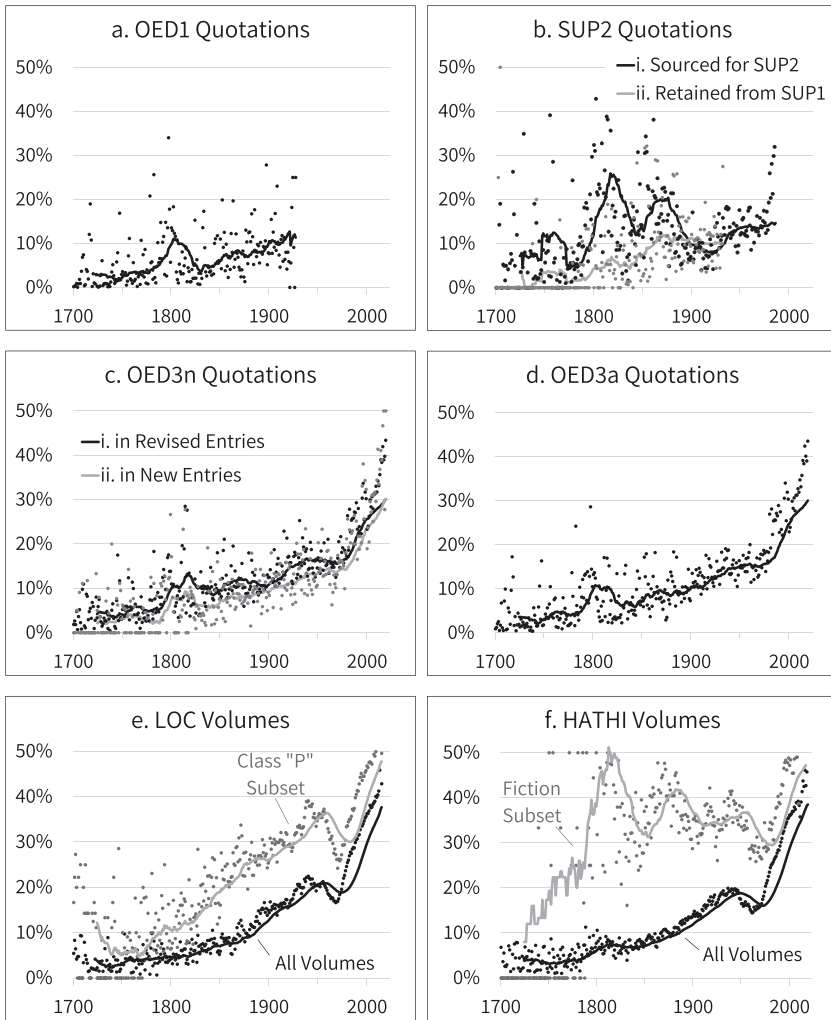
Fig. 2. Percentage female authorship in OED quotations and two comparator corpora, 1700–2022. *Dots represent figures calculated for each year, while lines represent a 25-year trailing average.*

In other words, each of the corpora in these graphs has always underrepresented women authors vis-à-vis the general population. The only corpus subsets that have ever approached the 50 per cent threshold on a 25-year aggregate basis are the Class 'P' books (Language and Literature) in LOC, and the 'Fiction' subset of HATHI. This occurs only very recently in the case of 'P', at the near end of the sharply rising, post-1970 part of the curve. In the HATHI 'Fiction' set, in addition to the recent climb towards 50 per cent, there is an earlier, short-lived moment of parity in the decades leading up to 1815, after which women authors recede somewhat, with the running average largely confined to a band between 30 and 40 per cent until the year 2000.

How then are we to read, for instance, the dismal-looking rate of 9 to 10 per cent female authorship among OED citations from the quarter century leading up to 1900 (in any edition or Supplement), when LOC and HATHI are returning figures of just 11 or 12 per cent? In Fig. 3, the OED averages from Fig. 2a–d are redrawn as the proportional excess or deficit compared to 2f (HATHI) (see Supplementary Data & Notes §§2.6, 3.4). This perspective is revelatory
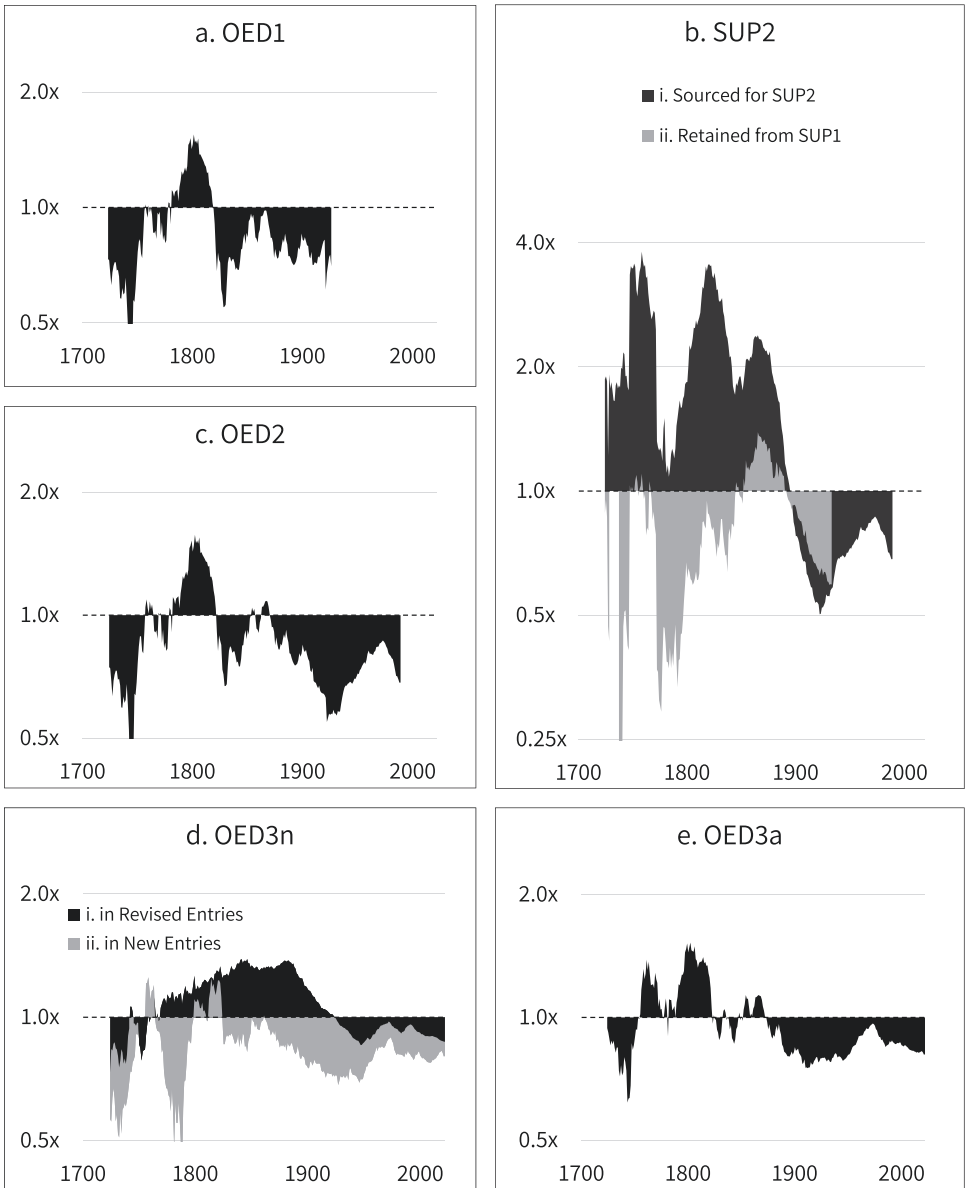
FIG. 3. Difference in percentage of OED quotations with a female author, vs Fig. 2f, HATHI Volumes, 1700–2022 (trailing 25-year average).

in several respects, none more salient than its reversal of the apparent direction of change in representation over time.

To look at all the quotations in *OED Online* today (Fig. 3e), for example, one would find, over-all, relatively even gender representation in eighteenth- and nineteenth-century quotations, with some periods of overrepresentation of women authors (centred around 1800) compensating for periods of underrepresentation (primarily before 1755, though the data here are sparse in both

corpora), compared to HATHI. The overrepresentation can be traced primarily to quotations added in SUP2 and in OED3's revision of existing entries (Fig. 3b and d), which together have broadened what was a briefer range of female overrepresentation in OED1 (Fig. 3a).

Outside this brief period around the turn of the nineteenth century (in which we find quoted prominently the women novelists discussed above), OED1 had, as has been previously assumed, significantly underrepresented women writers in all periods, vis-à-vis the benchmark set by HATHI. Whereas citation sourcing for SUP2, and later for OED3, mitigated and sometimes reversed these disparities in pre-1900 quotations, however, the picture is quite different for post-1900 quotations. Indeed it is in this timeframe, when the graphs in Fig. 2 begin to climb reliably above 10 per cent, that those in Fig. 3 all consistently underperform, for every revision, because the HATHI percentage is climbing faster, and in most cases accelerating at a higher rate as well. Perhaps most unexpected in this respect are the OED3n quotations added in new entries—i.e., for headwords not documented (as such) in OED2—which on the whole underrepresent women more than the quotations added to revised entries (Fig. 3d). This might be seen as an exaggeration effect of the pervasive deficit in post-1900 woman-authored sources, since the usage evidence for new headwords tends to be concentrated in the more recent past. It may also be that, whereas new evidence for revised entries could be quoted from previously quoted sources—among women authors OED3n relies most heavily on Austen, Burney, and George Eliot—these works are much less likely to produce evidence for previously undocumented headwords. Whatever the case, the picture repeated in each graph in Fig. 3 indicates that as women's writing becomes more common, the OED tends to underrepresent it more.

The corpus in Figs 2 and 3 that illustrates most sharply the pre/post-1900 divide is SUP2 (Figs 2b.i and 3b.i), incorporating quotations sourced for the Second Supplement to the OED, edited by Robert Burchfield from 1958 until the publication of its fourth and final volume in 1986, and additions made by John Simpson and Edmund Weiner during the preparation of OED2 (1989, effectively a merging of OED1 and SUP2). These graphs show three discrete periods before 1900 from which SUP2 sourced a greater proportion of quotations from women-authored sources than any edition or supplement before or since, and also overrepresented women authors compared to volumes in HATHI by a margin unseen in any other edition or revision ever. These three periods are indicated in Fig. 3b.i by two high peaks in 1750–1765 and 1820, and a smaller but still significant rise in the mid-1860s. They represent noteworthy figures: over the 25 years leading up to 1820, for example, just over a quarter of quotations added in SUP2 were authored by women, a proportion more than three and a half times the HATHI benchmark of 7 per cent, and more than double the highest levels ever reached in OED1 and SUP1 (unequalled in OED3n until 1997 in revised entries, and 2007 in new entries).

Unusually for a dataset the size of SUP2 (508,000 quotations), these highly visible macro effects can plausibly be attributed directly to the values, intentions, and decisions of individual actors, one in particular, who read widely both inside and outside the stated and implicit parameters of the project. That the three relatively aberrant periods in SUP2 occur before 1900 is not coincidental. The remit of the Second Supplement had been to replace the first, by bringing the OED up to date both lexicographically and historically. The focus was therefore primarily on lexicological development since OED1 began publication in 1884, and most especially on twentieth-century usages, supplementing existing coverage of earlier words and senses only when this was manifestly inadequate or incomplete (such as for those 'two ancient words', 'cunt' and 'fuck').[30]

---

[30] Robert Burchfield, 'Introduction' to SUP2, vol. 1 (1972), xii, xiv.

According to Burchfield the SUP2 Reading Programme collected 'about a million and a half quotations [i.e., slips] from works of all kinds written in the period from 1884 to the present day'.[31] About one third of submitted quotations (508,000) were eventually printed in SUP2, but more than 10 per cent of these were from outside the indicated timeframe. About 8,500 pre-1884 quotations in SUP2 (2 per cent) are drawn from posthumous publications and editions which would not have been available to previous editors, for example the letters of George Eliot (Yale, vols 1–8, 1954–1956), Charles Dickens (Pilgrim, vols 1–3, 1960–1974), Maria Edgeworth (Oxford, 1971 & 1979), Jane Austen (Oxford, 1952), and Elizabeth Cleghorn Gaskell (aka 'Mrs Gaskell'; Manchester, 1966), and the poems, journals, and letters of Gerard Manley Hopkins. Another 45,000 (9 per cent), however, are drawn from works published before 1884, many of them already quoted in the earlier edition. Prominent among these are Dickens (713 pre-1884 quotations in SUP2) and Mark Twain (588), whose corpus was made easier to scan by the publication of *A Mark Twain Lexicon* in 1938. Third-most quoted in this category is Charlotte Mary Yonge (1823–1901), whose 686 new SUP2 quotations include 498 published before 1884.

Yonge, in addition to being a prolific novelist, had herself been an early OED sub-editor (of the letter N) in the 1860s, under Frederick James Furnivall.[32] It was a latter-day follower of Yonge (in a variety of senses), however, who had the greatest impact on SUP2's representation of female authors. The journalist, novelist, and literary historian and biographer Marghanita Laski (1915–1988), herself quoted 209 times in SUP2 (including for a novel published under the pseudonym 'Sarah Russell'), began collecting quotations for the coming supplement soon after Burchfield's appointment, in the late 1950s. She is said to have submitted over 100,000 slips for the first (1972) volume of SUP2—'easily the most ambitious reading programme undertaken by any one reader', according to Burchfield—and perhaps another 40,000 for each of the three successive volumes.[33] Burchfield called her 'the most prolific and creative' of all the Supplement's readers, and paid tribute to her 'devotion to the Dictionary' in the first volume of SUP2 (1972).[34] Laski later undertook the proofreading of entries, a substantive task sometimes entrusted to domain experts outside the offices of the OUP.

Laski was a devotee of Yonge, advancing the academic study of her work with the foundation of the Charlotte Yonge Society, and an edited (with the biographer Georgina Battiscombe) collection of essays on her work, in 1965. But Laski was by no means a single-subject enthusiast: she later wrote critical biographies of George Eliot and Jane Austen, and in her academic and journalistic writings, including a series of articles for *Notes and Queries* published 1960–1962 (ed. Robert Burchfield), she discussed reading, specifically for the OED: Yonge, Eliot (503 new quotations sourced for SUP2), Austen (386), Gaskell (313), Edgeworth (262), and Burney (100), as well as Charlotte Schreiber (1812–1895; 135), Eliza Acton (1799–1859; 125), Dorothy Wordsworth (1771–1855; 75), Hannah Glasse (1708–1770; 67), Betsy Sheridan (1758–1837; 34), Cecilia Ridley (1819–1845; 21) and, among other twentieth-century sources, Virginia Woolf (1882–1941; 239).[35] In the front matter to SUP2 (vol. 1), Burchfield

31 Burchfield, 'Introduction', xii.
32 See Peter Gilliver, *The Making of the Oxford English Dictionary* (Oxford, 2016), 47.
33 Burchfield, 'Introduction', xiii. Laski's ODNB entry, authored by Burchfield, gives the round figure of 250,000 slips contributed by her between 1958 and 1986 (only a fraction of which would have been printed in the supplement). This figure, repeated widely in the literature, is on the very high end of the plausible range of inference from figures given in the individual volume prefaces and introductions.
34 Robert Burchfield, *Unlocking the English Language* (London, 1989), 8; Robert Burchfield, 'Preface' to SUP2, vol. 1 (1972), n.p.
35 See, by Marghanita Laski, in *Notes and Queries*, 7.5: 184–5; 7.6: 231–2, 232–3; 7.8: 312; 7.12: 459–65; 8.2: 64–7; 8.6: 229–30; 8.9: 339–41; 8.12: 468–9; 9.1: 27–30; 9.4: 147–50; 9.5: 167–71; 9.6: 223–6; 9.7: 269–70, 271–2; 9.8: 296–7; in *TLS* (11 January 1968), 37–9; and (31 October 1968), 1232; and in *LRB* 5.7 (1 April 1983). For a discussion of Laski's reading of Austen specifically, see Brewer, 'Jane Austen and the *Oxford English Dictionary*', 752–4.

would also mention Elizabeth Bowen (1899–1973; 353), as well as Laski's readings in 'numerous modern crime novels', and 'the general field of the domestic arts (old catalogues, books on gardening, cooking, embroidery, etc.)', naming specifically *The Guardian* (4,690 quotations in SUP2) and *Vogue* (300).[36]

Indeed, when writing about her monumental reading for the Second Supplement, Laski almost always discussed the writings of women, even if she did not generally address gender as a topic per se in these pieces. That Laski was therefore personally responsible for these and a great number more female-authored quotations in SUP2 is a natural inference, which has been drawn explicitly by Brewer and Willinsky, among others.[37] Further, we can say that, even if others also read and contributed quotations by, for example, George Eliot or Betsy Sheridan, Laski was almost certainly single-handedly responsible for the overrepresentation of women authors (vis-à-vis other editions, as well as general corpora such as HATHI) in the three periods prior to 1900 prominently visible in Figs 2b and 3b. To take the peaks of these hump (i.e., the 25-year period ending in each year), in each case around two-thirds or more of the female-authored quotations are by: (to 1770) Glasse (67 quotations, or 52 per cent) and Mary Wortley Montagu (18, 14 per cent); (to 1820) Austen (376, 40 per cent), Edgeworth (140, 15 per cent), Martha Wilmot (136, 15 per cent), Wordsworth (38, 4 per cent), and Ellen Weeton (30, 3 per cent); (to 1870) Eliot (342, 14 per cent), Yonge (303, 12 per cent), Gaskell (298, 12 per cent), Charlotte Brontë (155, 6 per cent), Acton (125, 5 per cent), Isabella Beeton (118, 5 per cent), Queen Victoria (65, 3 per cent), and Fanny Bury Palliser (55, 2 per cent).

Yonge, Eliot, Austen, Gaskell, Edgeworth, Acton, Wordsworth, and Glasse (2,411 quotations in total added in SUP2) are among those explicitly listed above as Laski's sources. Fitting within the 'general field of the domestic arts' are Glasse (*Art of Cookery*, 1747–1796), Acton (chiefly *Modern Cookery*, 1845), Beeton (*Book of Household Management*, 1861), Palliser (chiefly *History of Lace*, 1865), and Weeton (*Journal of a Governess*, 1807–1811, ed. 1969). Among modern works, 'crime fiction', as Burchfield calls it, is the most prominent genre among highly cited women authors, represented by four of the five most-quoted women of the twentieth century: Ngaio Marsh (492 SUP2 quotations; first among all twentieth-century women authors), Agatha Christie (449; second), Dorothy L. Sayers (446; third), and Margery Allingham (344; fifth).

Writing in the *TLS* in 1968, Laski recounted the remark of a male acquaintance who asked: 'Don't you sometimes … wonder how many of the words you collect are worth having?'[38] It is the sort of question—more of a comment—that one might expect from one of Esme Nicoll's encounters with the post-Victorian Oxford establishment, rather than a mid-century meeting of the British intellectual class, and one must wonder whether this particular gentleman would have ventured to say the same to Robert Burchfield. 'Naturally I told him off', Laski reports, and though in doing so she may have thought she was defending the honour of her beloved dictionary, her retort may also be taken in defence of the part she and others of her gender played in shaping it.[39]

Burchfield, who had nothing especial to say about gender beyond the grammatical sense of the word, made the most of the talent, industry, and expertise he recognized in Laski and other women. About half of the Editorial Staff credited in the front matter of each of the Supplement's four volumes are women, including prominently Lesley Susan Burnett, Senior Editor (General) of vols 3–4 and Sandra Raphael, full-time staff for vol. 1 and Senior Editor (Natural History and Library Research) for vols 2–4. Additionally, the first volume lists five female proof-readers (of eight), and twenty-nine (of eighty-six) 'important' contributors of slips, including Stephanyja

---

[36] Robert Burchfield, 'Contributors', in SUP2, vol. 1 (1972), x.
[37] See Willinsky, *Empire of Words*, 155, 185; Charlotte Brewer, *Treasure-House of the Language: The Living OED* (New Haven, CT, 2007), 162.
[38] Marghanita Laski, 'Words V', *TLS* (1 August 1968), 822.
[39] Laski, 'Words V', 822.

Ross, credited alongside Laski and two men with supplying together the greatest share (250,000 by the time vol. 1 was published). In the compiling of OED1 it had not been at all unusual for women to contribute large numbers of quotations and be credited for doing so—Russell estimates submissions upwards of 310,000 slips, by 258 named female contributors, based on credits published in OED1 front matter.[40] In neither edition, however—and this persists in the current edition—did labour force representation at this level result in rates of female quotation approaching the HATHI average after 1900.

REPRESENTING WOMEN'S WORDS

In addition to her exceptional work collecting quotation evidence for SUP2, Laski was among a small number of OED volunteer readers to write reflectively and in detail on this work for both scholarly communities (in *Notes and Queries*, for instance) and more general audiences, in the *LRB* and *TLS*. In the latter venue, she wrote in 1971 that 'one tends to get the impression, when reading the *OED*, that it was the giants of literature who formed our language. Any reading in trivia shows this impression to be wrong.'[41] It was a point she herself had demonstrated in that same publication the previous year: in 'Words from "The Times"', Laski reported whiling away some free hours to find seventy-odd antedatings and unregistered words in the 1 January 1785 issue of the *Daily Universal Register* (the name, until 1788, of what was after *The Times* of London). Most of these, she noted, came from the advertisements, and not the 'vigorous contemporary language of John Walter.'[42]

The citations Laski records, which OED1 and SUP1 had not found or did not use, begin with an attestation of 'Bath beaver' (from an advertisement for Mecklenburg Cloaks for Ladies), and proceed through a number of other terms related to clothing and fashion ('chain serge', 'habit-maker', 'hosiery', 'measure', 'south-willet', etc.), food, food preparation, and dining ('coffee-kitchen', 'fish-knife', 'kitchen-range', 'tea-kitchen', etc.), and sundry social and domestic activities ('Conversation card', 'fortune teller', 'sable iron', 'sentimental card'), along with some pertaining to journalism and newspapers ('column', 'morning paper', 'wire-dancing'). Laski probably was not looking consciously or with intent for women's words, but in the event, as this sample indicates, a large number belong to language associated with women's pastimes and occupations. Many are taken from advertisements directed specifically to 'Ladies', 'The Fair Sex', and so on. Only a few could be said to belong to a general vocabulary, and even fewer (bar those journalistic terms) to the kind of public language epitomized by the 'vigorous' John Walter (1738–1812), publisher of the *Daily Universal Register* and then *The Times*.

SUP2 eventually printed nineteen of the *Daily Universal Register* citations Laski collected that day.[43] OED3 has retained seventeen of these, suppressed two, and added a further five from Laski's list, which SUP2 had passed over.[44] In SUP2, as in OED1 and later OED2, senses with restricted uses might be found labelled as belonging to one of a number of assorted domain vocabularies, as Laski's sense of 'tenor' ('A name for the tenor violin') comes under a *Mus.* label (OED2, sub II.4). Such labelling is mostly unsystematic in OED2, however, and beyond a handful of very common labels, often haphazard and arbitrary. With refence to identified speaker groups (as distinct from knowledge domains), categories include such gender- and class-marked labels as *Workmen's slang* (four senses), *Schoolboys' slang* (ten), and even *Eton College slang* (fifty-two).

---

40 Totals here are derived from data in Table 4.2 in Russell, 154–7.
41 Marghanita Laski, 'Letter', *TLS* (13 October 1972), 1226.
42 Marghanita Laski, 'Words from "The Times"', *TLS* (2 April 1971), 403.
43 In SUP2: s.v. BOAT, COLUMN, CONVERSATION, DISH, JEAN(N)ETTE, KITCHEN (two quots), LITERARY, NEWSPAPER, OLD, PIERCED, PROTECTING, RETAIL, SABLE, SHIP, SUGAR, TABLE, TENOR, and WIRE.
44 The quotations for 'literary society' and 'wire-dancing' have been suppressed, with Laski's antedatings added s.v. MEASURE, MUSLINET, OPERA, PENCIL (sub 'pencil cedar'), and WHOLESALE (sub 'wholesale price'). A sixth quotation from that issue of the *Daily Universal Register*, which Laski does not mention, was also added, s.v. PRINTING HOUSE.

Reading through a list of OED2's subject and speaker-group labels (there are hundreds), one is reminded of the parochial opinion of that reputed lexicographer who wrote that 'most American slang is created and used by males … The majority of entries in [the *Dictionary of American Slang*] could be labelled "primarily masculine use".'[45] In OED2, there are no subject categories explicitly pertaining to women or girls, and only a handful for which a preponderance of female usage might be inferred based on received ideas about historically gendered occupations. Among these, the areas of textiles, fabrics, and fashion are represented by a small number of category labels, as is food preparation. The most common of these—*Cookery*—appears in 1,255 senses. It is a modestly high count, about as much as for *Angling*, though still less than a third as much as, for example, *Cricket* (3,995). *Needlework* (425) appears half as often as *Billiards* (860); *Knitting* (89) about one-seventieth (1.5 per cent) as often as *Heraldry* (6,119). One sense of one word is labelled *Midwifery*, the same number as *Lichenology*, and one-tenth that of *Oxford University slang*. There are no words with labels proper to nunship or nursing or domestic service or sales, clerical, or sex work, or nearly any of the myriad occupations and professions disproportionately filled by women at various times throughout history; nothing especial to do with household management, nor anything pertaining to women's social societies and organizations. As for the question of the masculine domain of jargon and slang speech, as Deborah Cameron put it to Jonathon Green, throughout history female institutions with specific vocabularies 'existed for women but they were not studied … men couldn't collect it.'[46]

Even when men and others did collect women's slang and jargon terms, their presence in the OED has not been assured, or they have not been identified as such. When the label *Hairdressing* registers in the 'definitive record of the English language' (as the OED has billed itself online) with the same incidence as, say, *Zoogeog.* (eight senses, or as good as never), and less often than, for example, in increasing orders of magnitude, *Craniometry* (65), or *Gunnery* (374), or *Phonetics* (1,134), or *Naut.* (15,321), it is difficult to avoid drawing inferences about the relative importance of these various knowledge sectors. Or perhaps even about their intrinsic importance: the famous OED labels *nonce-wd(s)* and *nonce-use(s)*, designating fleeting linguistic innovations— i.e., invented 'for the nonce', and so, in principle, belonging to no jargon or other shared lexicon, or any common lexicon—occur 19,968 times, or over 2,000 times as often as senses pertaining to an occupation routinely practiced on and by women since at least the late eighteenth century.[47]

What our inferences of importance attach to—to the English language, say, or its written corpora, or to volunteer readers or OED editors—will depend much on our values and perspectives as readers of the OED. It may well strike one (as it strikes me) as absurd to think that the language of hairdressing (to stay with this example) is or was of less relative importance to spoken English than, say, that of craniometry. And, with due account of Martin's point that 'Historical dictionaries can only be as inclusive as their sources', while it may or may not be the case that more words have been written explicating the latter topic than the former, as long as the lexical item itself has been committed to paper a small handful of times, in principle it should be recorded, and so labelled, with the same frequency as a word that has been written down hundreds or millions of

45   S. B. Flexner, 'Preface' to Flexner and H. Wentworth, *Dictionary of American Slang, Second Supplemented Edition* (New York, NY, 1975), xii. The patience with which Katherine Connor Martin (also an OED editor, and Oxford Languages manager) and Vivian de Klerk separately debunk this idea may itself be reflective of gender dynamics in the field. See Katherine Connor Martin, 'Gendered Aspects of Lexicographic Labeling', *Dictionaries: Journal of the Dictionary Society of North America*, 26 (2005), 160–73; Vivian de Klerk, 'Slang: A male domain?', *Sex Roles*, 22 (1990), 589–606.

46   Deborah Cameron, quoted in Jonathon Green, *Sound and Furies: The Love-Hate Relationship between Women and Slang* (London, 2019), 21–2.

47   That is, at least as known by that term: OED's first citation for HAIRDRESSING, unrevised since SUP2, is from Smollett (1771). The OED3 evidence as currently published, all of it held over from OED1/2, indicates moderate female gendering of 'hairdresser' and 'hairdressing', apparently more so in latter years; up-to-date American sources such as the *Merriam-Webster.com Dictionary* indicate an even stronger gender tendency (for both the hairdresser and the hairdressee), though this is rarely made explicit in the definition (cf. <https://web.archive.org/web/20220515163430/https://www.merriam-webster.com/dictionary/hairdresser>).

times.[48] We are here in the realm of low thresholds, rather than representation in the statistical sense.

It is a much more plausible inference that selection rather than production is the primary driver of this discrepancy, and this can occur at multiple junctures: hairdressing words may appear in publications typically overlooked by OED editors and volunteer readers; or readers may overlook those words in the publications they do trawl; or editors may put aside such terms after they have been submitted by readers; or they may include them and write them up, but fail to designate them as special uses particular to a recognized knowledge domain. The final scenario is evidently what occurred in SUP2 for all but one of the ninety-six terms with quotations drawn from J. S. Cox's *Illustrated Dictionary of Hairdressing and Wigmaking* (1966).[49] By way of contrast, the 1,500-odd OED2 citations taken from the prominent heraldic compendiums of Guillim (1610), Bossewell (1572), Cussans (1868–1882), Leigh (1562), Porny (1766), and Elvin (1889) draw the labels *Heraldry* or *Her.* over 60 per cent of the time (respectively, in descending order of total quotations: 49, 59, 83, 62, 81, and 87 per cent). One might be forgiven then for inferring that the composite compiler of OED2 thought the Kings of Arms to possess a significant established vocabulary of their own, whereas the hairdresser and her clients merely used the common tongue uncommonly.

Subject labelling in OED3 is substantively different from OED2 in a number of ways. Though the main text preserves the majority of OED2's italicized labels (replacing some offensive or insensitive terms there, and making further edits and revisions where appropriate), these labels have been largely superseded in *OED Online* by a system of subject 'Categories' derived from them, which have been populated algorithmically beyond the labelled text, based on textual keywords and other cues in the definitions.[50] Thus, to stay with the previous example, OED3 has fifty senses labelled *Hairdressing* in thirty-nine entries, but the Category 'Hairdressing' includes 1,025 senses in 800 entries. Already such a discrepancy points out a gap between the preponderance of hairdressery in the OED quotation corpus and the instance of *Hairdressing* as a label designating a knowledge domain, even allowing for a fair amount of algorithmic misclassification.

The examples of subject labels and Categories I have offered thus far have been selected to illustrate, as colourfully as possible but somewhat arbitrarily, some discrepancies in the handling of semantic categorization in the dictionary, and for that very reason they cannot be taken as dispositive in and of themselves. However, the Category structure employed in OED3 offers the opportunity of a more systematic analysis. Because Categories are structured hierarchically, they can be aggregated and disaggregated according to the specificity of the knowledge domain to be investigated, allowing for an analysis of the quotation evidence accumulated therein. Fig. 4 shows such an aggregation as a comparison of top-level OED3 subject Categories, counting the percentage of female-authored quotations in the citation evidence appearing within that domain (the total number of quotations for each Category is represented by dots, giving a sense of their relative preponderance). Because we know the gender of authorship to be correlated to the year of publication, Fig. 4 divides the corpus into the long period 1700–2022 (4a&b.i), and the modern sub-period 1950–2022 (4a&b.ii); and because the revision in which a quotation is added to the OED is also a known factor, newly sourced OED3n quotations (i.e., those added since

48 Martin, 'Gendered Aspects of Lexicographic Labeling', 164.
49 J. S. Cox, *Illustrated Dictionary of Hairdressing and Wigmaking* (London, 1966). To get a rough idea of the volume of specialist vocabulary in this area, the new, revised (1984) edition of Cox, which added 'over a thousand additional entries', lists over 8,150 terms A cursory analysis shows 463 of these entries to refer only to female-gendered terms (e.g., 'girl', 'woman', 'lady'), 375 only to male-gendered terms (e.g., 'boy', 'man', 'gentleman'), with another forty-eight referring to both male- and female-gendered terms.
50 Labels can still be indicated as a search parameter in the 'Advanced Search' section of *OED Online*, but the online interface strongly shepherds users towards the 'Category Browse' and 'Search' functions instead. I discuss the 'Category' system introduced in *OED Online* and some of its drawbacks and pitfalls in David-Antoine Williams, '"Alien" vs Editor: World English in the *Oxford English Dictionary*', Policies, Practices, and Outcomes 1884–2020', *International Journal of Lexicography*, 34 (2021), 39–65, 50–51.
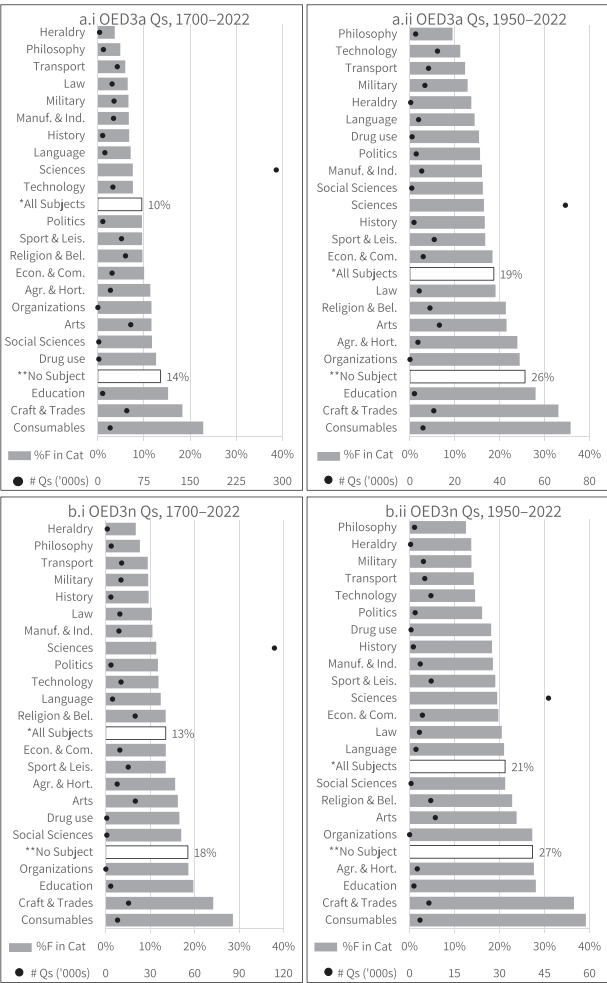
Fig. 4. Percentage of women-authored quotations in OED3 by top-level subject category. *Dots show total number of quotations in category, in thousands.*

the publication of OED2) are separated out (in 4b.i&ii) from all quotations displayed in the current *OED Online* (OED3a, in 4a.i&ii). For reference, figures are also given for all Category subject-labelled senses in each corpus, and for all senses with no Category subject label.

The graphs in Fig. 4 all show a strong male skew in the distribution of subject Categories, regardless of when the quotations analysed were published (period) or sourced (edition). This is especially apparent in the persistent gap, across all graphs, between aggregated un-categorized ('No Subject') quotations and aggregated categorized ('All Subjects') quotations. That is, for every period and edition, whatever male bias might apply to the corpus of quotations as a whole, it is exaggerated in (and therefore partially attributable to) subject-categorized senses. Neither is this an area in which recent OED work appears to be ameliorating matters very much: while the overall percentage of female-authored quotations increases when comparing the (a) graphs to the (b) graphs, and the (i) graphs to the (ii) graphs, subject-categorized female authorship is 6 percentage points lower in newly sourced modern quotations (21 vs 27 per cent, Fig. 4b.ii),

compared to 4 percentage points lower in female-authored quotations overall (10 vs 14 per cent, Fig. 4a.i), a disparity increase of 2 percentage points.

Part of the gender disparity in subject Categories is due to the very large number of quotations illustrating senses in a Sciences Category, which skew strongly male in all periods. But Sciences is far from the most male-skewing Category in the OED3 Categorization system: depending on the edition and period of the quotation, between seven and ten Categories contain even smaller proportions of women's writing, including History, Politics, Language, and especially Philosophy, which (along with Heraldry), has the lowest percentage of female authorships across all periods and editions. On the other side of things, across all graphs, the Categories most likely to skew female compared to uncategorized quotations are Education, Crafts & Trades, and Consumables, with Agriculture & Horticulture joining in 4b.ii (also joining in 4b.i, though trivially, is Organizations, a weirdo Category with only a handful of senses pertaining to either Scouting or Guiding, or Freemasonry). Subjects that tend to rely on more female-authored evidence than the average, but which still have a smaller proportion than uncategorized senses, include Religion & Belief and Arts, as well as Economics & Commerce and Social Sciences.

Equally apparent in Fig. 4 is that no top-level Category draws more than 25 per cent of its citation evidence from women writers overall (4a.i), and no subdivision of period or edition results in a top-level Category with as much as 40 per cent female representation. Subcategories are naturally less evenly distributed, but viewed together they too present an overwhelmingly male bias. The most granular level of Category is the Category endpoint (*n*-levels deep depending on the category and sense). Looking only at modern (1950–2022) quotations newly sourced in OED3n (i.e., those aggregated in 4b.ii), 335 Category endpoints (90 per cent of endpoints) contain 60-plus per cent male-authored quotations, whereas just 10 (3 per cent of endpoints) contain 60-plus per cent female-authored quotations. These are: Needlework (152 of 217 quotations [not figuring 11 with no gender attributed]), Lace (72 of 93 [2]), Ballet (34 of 56 [5]), Knitting (136 of 194 [17]), Tailoring (34 of 54 [1]), Silk (27 of 45 [2]), Dressmaking (22 of 33 [nil]), Sumo Wrestling (9 of 14 [6]), Mah-jong (11 of 17 [1]), and Enamelling (9 of 12 [nil]).

In discussing the gender of highly cited authors, Brewer concluded that, whatever the historical fact of the matter may be, 'we are left with the impression, in *OED3* as in previous editions, that male literary writers have contributed far more than female to the history and development of the language'.[51] Evidently this is as much the case, if not more so, for the history and development of knowledge domains in English-speaking places. Whether a dictionary user is exploring OED3 Category areas, as *OED Online* encourages one to do, or investigating directly a term of art in some familiar or unfamiliar field, that term is likely either to be overwhelmingly evidenced by male authors—even more so than the dictionary overall—or the field itself will not be appear as a distinct subject Category at all.

Two benchmarks, one internal and one external, might usefully be brought up against these figures and the impressions they give. The internal benchmark would utilize a separate system of semantic categorization embedded in OED3: the sense classification of the *Historical Thesaurus of the Oxford English Dictionary* (HTOED).[52] This semi-independent classification scheme, which begins with OED definitions (sourced mainly from OED1) but is compiled outside the OED operation (then to be re-integrated periodically with *OED Online*), assigns a semantic endpoint to (in theory all, though in practice not yet quite all) OED senses irrespective of usage restrictions or perceived significance. While HTOED's categories and taxonomy are in some ways as subjective and socially contingent as any effort at semantic organization must

[51] Brewer, 'Literary Quotations', 115.
[52] Christian Kay, Marc Alexander, et al., *The Historical Thesaurus of English*, University of Glasgow <https://ht.ac.uk/>. As HTE is periodically integrated with *OED Online* (referred to as such as HTOED), it can be accessed from *OED Online* entries, or at <http://www.oed.com/thesaurus/>, or independently from *OED Online* (as HTE) at <https://www.ht.ac.uk>, where sundry useful explanations and materials have also been collected.

be, the mere fact of its comprehensive application circumvents inclusion as the primary inflection point of such discriminations.[53] Its vast and intricate hierarchical construction, with nearly 50,000 endpoints in the focus period, thus provides another way of aggregating quotations according to domains and subdomains of human activity and thought. In the Supplementary Data & Notes §3.7, HTOED categories are presented in further detail according to the percentage of female-authored quotations they hold, for every category according to its depth in the hierarchy (for example, there are three top-level domains, namely THE WORLD, THE MIND, and SOCIETY), and for endpoint categories.

Reaggregated this way, at first glance the prevalence of the words of women in the OED3 quotation corpus may not seem dramatically different than in the subject Category aggregation. The percentage of those endpoints with more female than male authorship is still miniscule—less than 1 per cent overall, and about 4 per cent when counting only recent (1950–2022), newly added OED3n quotations. However, the systematic nature of HTOED's semantic categorization results in a much broader diversity of subject matter than the eight arts and textiles areas that make up the bulk of predominantly female-evidenced subject Categories. In the top-level domain THE WORLD, in addition to textile and clothing, there is evidence of longstanding strong female influence on senses pertaining to horticulture and gardening, food and food preparation, domestic animals and the breeding thereof, and sexuality and women's health; in THE MIND, on senses falling under the subcategories Mental Capacity (understanding, pedantry, doubt), Attention and Judgement (observation, esteem, beautification—including hairdressing), and Emotion and the Will (excitement, love, suffering, decision); in SOCIETY, on kinship, social attitudes, domestic service and household management, morality, faith, and education (see Supplementary Data & Notes, §3.7.2). Across all editions and periods, semantically gendered terms with a preponderance of evidence from female writers include those for womanly qualities, Black women, nurses, schoolgirls, female servants and governesses, and secretary work.

Still, these are but 265-odd distinct endpoints out of 50,000 in the HTOED taxonomy, using the union of all editions and periods in question, with any one permutation yielding at most 135-odd; and in the final analysis perhaps it is encouraging in only the mildest of ways to note that the current revision's editors have sourced evidence for, e.g., FEMALITY, FEMINACY, FEMINEITY, etc., something like two-thirds of the time, when, to take a counterexample, post-1700 quotations in the entry FEMALE (revised in 2012) are 79 per cent male-only authored.

Thus we are returned to the matter of the availability of evidence and the question of representation: to take the OED3 record at face value would imply not only that the share of female authorship grew more slowly after circa 1900 than HATHI shows it to have grown (Fig. 3); but also that in the twentieth century and after, documentation of new words and senses increasingly centred on areas in which male-authored texts predominated—on scientific terms, perhaps, or on sporting or military terms. Here a second external benchmark may be applied to contextualize the deficits shown in the latter parts of the graphs in Fig. 3. This is the catalogue of the Library of Congress (LOC), plotted in raw terms in Fig. 2e, which allows a further segmentation according to subject area, represented by the LC Classification system. Fig. 5 compares female authorship segmented by LC Class for two OED3 quotation subsets—all OED3 quotations (OED3a) 1700–1950 (Fig. 5a), and newly sourced quotations (OED3n) 1950–2022 (Fig. 5b)—to LOC volumes for each period, weighted according to the distribution of dates in the OED subcorpus (Fig. 5c.i&ii). The relative proportion of each Class in each subcorpus is also indicated (see Supplementary Data & Notes, §§2.5, 3.5).

---

[53] A pertinent illustration of the contingency of semantic organization is the modernizing reclassification of certain HTOED categories under a new, third-level 'sex and gender' category, applied to OED3 in the second half of 2022 (after the present analyses were carried out).
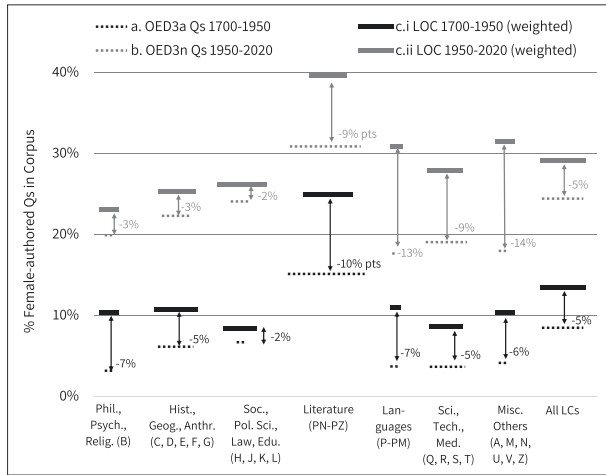
FIG. 5. Percentage of women-authored quotations in OED3 vs volumes in LOC (date-weighted), by LC class. *The width of each line along the x-axis is proportional to the relative size of the class subset within the relevant dataset.*

Where Fig. 3 showed fluctuating if pervasive deficits in OED corpora versus HATHI after 1900, the aggregation in Fig. 5 shows an overall proportional deficit of −37 per cent in the early broad corpus (8 vs 13 per cent, Fig. 5a), versus −16 per cent in the more recent OED3n corpus (24 vs 29 per cent, Fig. 5b). Quotations from books classed as Literature (Classes 'PN'–'PZ') are both the most prevalent (amounting to 35 and 37 per cent, respectively) in each OED corpus, and have the highest percentage of female authorship (15 and 31 per cent, respectively). But these high numbers compare disfavourably to those observable in the comparator corpus: female authorship in recent OED3n Literature quotations is −9 percentage points lower than in LOC, a proportional deficit of −22 per cent (Fig. 5b). In raw terms this is identical (−9 points) to the second-largest class of recent quotations, those from scientific categories (Classes 'Q', 'R', 'S', 'T', amounting to 22 per cent in Fig. 5b). But proportionally this means that quotations from scientific texts are quoted at a rate −32 per cent less in OED3n than their share in LOC. This suggests that a significant number of female authors are not passing from the available corpus into the OED record, and that this is especially pronounced in the two classes most heavily covered by the dictionary. In the scientific classes in particular, the OED record appears to exaggerate, and may be seen to authenticate, social stereotypes around female participation and achievement in science, technology, and medicine, even as these are being challenged in the general culture.

FUTURE OED3 REVISION

The ongoing OED project had, at the time of the June 2022 update, revised about half of pre-2000 entries, adding 1.49 million new quotations (39 per cent of all OED3 quotations) and retaining 904,400 existing quotations in 138,200 fully revised entries. A further 143,800 new quotations have been added in 18,100 new entries, and in draft additions and insertions in otherwise unrevised entries. 131,500 entries remain unrevised, containing 1.26 million quotations held over from OED2 (33 per cent of OED3 quotations). Taking stock now of the OED's representation of women's words (in both senses) offers the opportunity to reflect on ways in which English vocabulary may be better documented in these areas as the project moves ahead.

In the broadest terms, recently this has been a policy matter of some priority for OED and Oxford Languages management, which representatives have been minded to underline in public documents. As a precursor to the OED's 100th Anniversary celebrations, planned for 2028, the *OED Blog* announced the position of a Content Inclusivity Manager to aid in forming editorial guidelines, as well as upcoming editorial work on 'the representation of women, people of colour, the LGBTQ+ community, and those with disabilities, among others'. The same paragraph reveals a tension underlying this aim, however: 'At the heart of this process is research and evidence – always the life blood of the OED. We are working to ensure that this evidence is inclusive and represents multiple perspectives, so that not just mainstream opinion is reflected, but that marginalized voices are heard too'.[54] Research and evidence will supply the material which editorial policies will represent. But because the analyses conducted here on the whole indicate systemic effects rather than policy achievements, they suggest that improvements in this regard will largely stem from changes in systems and workflow, rather than, for instance, the revision of exclusionary language in definitions, however desirable this may be.

To that end, four system-wide measures might be undertaken now to remediate the record as it stands, improve future work, and allow for improved contextual presentation of the dictionary. They are:

1. A comprehensive scan of women's writing circa 1890–1950 to supplement existing entries, especially where significant gaps in the OED's semantic or chronological coverage exist, and to collect evidence for future revisions and new entries.
2. A selective reading program in women's writing after 1950, directed to: a) areas where there is a large amount of female-authored textual evidence available (for example, LC Literature and Sciences classes); and b) specialist works on women and women's work, for example in women's and gender studies, or specialist dictionaries and lexicons.
3. A new *OED Online* subject Category taxonomy, which while preserving old labels as a matter of historical interest, would be more systematic, less inductive, and less discriminating in its coverage.
4. The development of accurate and comprehensive gender metadata in the bibliography of OED sources, and the integration of this in *OED Online* search functions.

Some of these measures, especially the first and second, may be seen as costly in terms of the editorial, lexicographical, and technological resources they might require, and the benefit must be weighed against the other outcomes towards which these resources might be directed, principally the revision of pre-2000 material persisting in OED3. But in so doing, it must also be borne in mind that the benefit of more robust research in women's words will also accrue to the revision project generally; and the benefits of better search and categorization will accrue to the presentation and reception of the dictionary as a whole.

St Jerome's University, Canada

---

54 Charlotte Buxton, 'OED100: Repainting the dictionary'. *OED Blog* (18 March 2022), <https://web.archive.org/web/20220321135242/https://public.oed.com/blog/oed100-repainting-the-dictionary/>.