Investigating Curatorial Voice with Corpus Linguistic Techniques: the case of Dorothy George and applications in museological practice

Andrew Salway* and James Baker**

Abstract

We seek to demonstrate how corpus linguistic techniques can facilitate a comprehensive account of curatorial voice in a large digitized museum catalogue and hence leverage its value as a resource for generating new knowledge about: curatorial practice; the historical and cultural contexts of curation; and the content of collections. We worked with 1.1 million words written by the historian M. Dorothy George between 1930 and 1954 to describe 9,330 late-Georgian satirical prints. George's curatorial descriptions were analyzed in terms of their typical informational content and with regards to the extent George included interpretation and evaluation in her descriptions. We discuss how results from such analyses can provide a basis for addressing questions about George's curatorial voice and, more generally, suggest how this approach could benefit museological practice around the production of descriptions and the re-purposing of legacy catalogues for digital access and analysis of collections.

Key words: Curation, catalogues, corpus linguistics, British Museum, Dorothy George

1. Introduction

This paper concerns curatorial voice, the authorial voice of institutions and the contemporary role of legacy museum catalogue descriptions in indexing and understanding collections. Extensive digital and digitized sets of curatorial descriptions are increasingly available. The starting assumption for this paper is that these have the potential to be valuable resources for generating new knowledge about curatorial practice, the historical and cultural contexts of curation, and the content of collections at the level of individual items and at a macro level. However, digital and digitized catalogues have not yet been recognized as a form of 'big data' such that new and different kinds of questions can be asked about curation and the content of collections. In response, the 'Curatorial Voice' project is applying computational text analysis techniques to a large digitized catalogue of curatorial descriptions.[1] This work has two ambitions: to establish new directions in historical research – both into the content of collections and into institutional/cultural labour – and to enhance search and discovery functions that are based on legacy catalogue descriptions.

In 1995 Carol Duncan argued that 'to control a museum means precisely to control the representation of a community and its highest values and truths' (Duncan 1995: 8). Descriptions of objects curated by museums are one means by which a community's values and truths are controlled. For example, in the mid-twentieth century the historian M. Dorothy George catalogued over 12,500 late-Georgian satirical prints for the British Museum. Her descriptions of these prints – most of which were published in London – made a vital contribution to research on Georgian Britain. However, George's descriptions are far from straightforward verbal representations of visual representations. Rather, as exemplified by George's caution and squeamishness in the face of scatological humour, they are a product of a voice shaped by traditions, preferences, and values.[2] Today, curatorial descriptions – including those, like George's, made before the information age – are commonly subsumed into services that provide valuable public access to collections via textual search. However, this aggregation

comes at a price, for it masks curatorial idiosyncrasy with a datafied veneer of institutional authority (Putnam 2016) at the same time as scholars like Duncan – informed by queer and post-colonial turns – have sought to decentre the authorial voice of institutions in art historical and museological discourse (Çelik 1996; Greene 2016; Turner 2017).

Thus, we identify the need to be able to elucidate and foreground curatorial voice for a given set of collection descriptions, both as a precursor to studying their curatorial voice(s) and to enhance their use for search and discovery of collection items. Analyzing curatorial voice at scale requires the comprehensive articulation of curators' choices and preferences across what are often large bodies of text produced during decades of work. The challenge here is not only the scale of the material, but also the complexity arising from the multiple facets of objects that curators may refer to, and the variety of ways in which voice can manifest, e.g. through inclusion and omission, that is, what aspects of the item are referred to, and through varying degrees of description and interpretation/evaluation. We propose that the combination of corpus linguistic techniques such as word lists, keyness and sorted concordances,[3] alongside some qualitative close reading, is a scalable approach well suited to producing comprehensive accounts of the language used in museum catalogues.[4]

The aim of this paper is to demonstrate the value of corpus linguistic techniques for generating knowledge about curatorial voice. Section 2 describes the creation of the BMSatire Descriptions corpus comprising curatorial descriptions from Volumes 5 to 11 of the *Catalogue of Political and Personal Satires Preserved in the Department of Prints and Drawings in the British Museum* (those volumes written by George, and hereafter referred to as the '*Catalogue of Political and Personal Satires*' (George 1935-1954)), and outlines our approach to investigating curatorial voice. Section 3 reports a preliminary investigation of curatorial voice in this corpus with a frequency-led analysis of curatorial voice in terms of inclusion and omission (Section 3.1), and a variety of analyses focussing on the extent of interpretation/evaluation in George's curatorial descriptions (Section 3.2). The main objective here is to identify and critique ways in which corpus linguistic analysis may contribute to a comprehensive and systematic study of curatorial voice, rather than to complete such a study. In closing, Section 4 discusses how corpus linguistic analysis can contribute to investigating curatorial voice and, more generally, how our approach may support both the production of new catalogues and the re-purposing of legacy catalogues for accessing and analyzing collections.

2. Corpus building and approach

A corpus of George's curatorial descriptions was made from the contents of the 'Physical Description' field in a selection of records that we retrieved from the British Museum's Research Space SPARQL endpoint.[5] A query retrieved 23,932 records relating to the *Catalogue of Personal and Political Satires*. This set of records was refined iteratively with custom scripts and then the text content of the 'Physical Description' field in each record was taken and prepared for corpus linguistic analysis. The following two paragraphs describe the main points about the processes for selecting records and preparing the text data. More detailed documentation of the query and these processes, along with the resulting corpus and associated datasets, are available to download (Baker and Salway 2019a).

The selection of records was refined by discarding those relating to prints that were published outwith the period covered by Volumes 5-11, i.e. 1771-1832, and by discarding records relating to prints that were acquired by the British Museum after 1929 because George started her work in 1930. These criteria were chosen to minimize the chance of including descriptions that were not written by George, at the expense of missing out on some records containing descriptions that were written by her. The selection process resulted in a set of 9,330 records. The text from the 'Physical Description' field was taken from each of these records to form the BMSatire Descriptions corpus. Our assumption is that this field contains physical descriptions of collection items, in contrast to the 'Curatorial Comments' field, which we expect to contain George's writing about the historical contexts of prints and which hence seems less likely to reflect curatorial voice and more likely to have been edited by later curators. Spot checking suggests that data from the 'Physical Description' field matches the text printed between 1935 and 1954, and that any variations are unlikely to have a significant effect on the results of corpus linguistic analysis.

The selected text was processed in order to replace all text occurring in quotation marks with '*TRANSCRIBED*' and all text in brackets with '*BRACKETED*', see Table 1 for an example. We observed that most text in quotation marks is the transcription of words written in the prints such as speech, labels and signs. These words are not reflective of curatorial voice; however, the fact that they are transcribed is, so we leave a trace of the transcription in the text rather than deleting it entirely. Text in brackets comprises a variety of things, such as cross-references to other prints, curatorial notes, as well as further description of the scene and historical information. In fact, a lot of the bracketed text does contain valid curatorial description; however, because it is mixed with other kinds of information and the text is sometimes in note form, we chose to make the replacement as a convenience for subsequent corpus analysis.[6]

One of a set of prints on the marriage of the Prince of Wales, see BMSat 6924, &c. The Prince presides at a carouse round a circular table. His chair is above the level of the table on which he puts his left foot, crushing a wine-bottle; his right hand rests on the shoulder of (?) Sheridan, his left holds a wine-glass above his head; he says, \"Fall too Ye royal crew! Drink Drink! your bellies full! pray do! Ai treats I never winces\". Five men (right) are seated on his left: … On the left side of the table and on the Prince's right is first (?) Sheridan, then Weltje, then Topham (a Ministerialist journalist), then Lord Derby, leaning on the table and pointing to two coins. Next is a man in naval uniform, [Perhaps Admiral Hugh Pigot, see BMSat 5996, &c. (1782)] turning away from the table, resting his head on his arms which are on the back of his chair. On the table beside him is an open music-book inscribed, 'Catches Glees Which is the properest Day to Drink Saturday'. In the foreground sits a stout and jovial-looking man. [Perhaps Captain Morris (to whom the music-book may belong)] …

One of a set of prints on the marriage of the Prince of Wales, see BMSat 6924, &c. The Prince presides at a carouse round a circular table. His chair is above the level of the table on which he puts his left foot, crushing a wine-bottle; his right hand rests on the shoulder of *BRACKETED* Sheridan, his left holds a wine-glass above his head; he says, *TRANSCRIBED* . Five men *BRACKETED* are seated on his left: ... On the left side of the table and on the Prince's right is first *BRACKETED* Sheridan, then Weltje, then Topham *BRACKETED* , then Lord Derby, leaning on the table and pointing to two coins. Next is a man in naval uniform, *BRACKETED* turning away from the table, resting his head on his arms which are on the back of his chair. On the table beside him is an open music-book inscribed, *TRANSCRIBED* . In the foreground sits a stout and jovial-looking man. *BRACKETED* ...

Table 1: A sample of text before and after the text preparation process.

The resulting corpus contains about 1.1 million words from the 9,330 records, which is a substantial amount of text data for investigating curatorial voice. In broad terms, our methodological approach was to combine automated statistical analysis – counts of words, phrases and patterns, and keyword lists – with the close reading of concordance lines, but not yet with the close reading of complete descriptions and the prints to which they refer.

We began working in a data-driven way by concentrating on the most frequent words in the corpus to try and characterize the typical kinds of information given in descriptions, and by identifying negative keywords (words that occur less often in the corpus compared to a corpus of everyday language) to see what is not said, without relying on any preconceived notions about what to expect; see results in Section 3.1. Of course, concentrating on the most frequent words means missing rarer phenomena and more common phenomena that manifest in a wide variety of low frequency words. Thus we also took an hypothesis-driven approach to investigate the extent of interpretation/evaluation in the corpus because this was not apparent from inspection of the most frequent words. This involved testing preconceived ideas about how curatorial interpretation/evaluation might manifest by searching for and counting instances of words, phrases and patterns that were hypothesized as indicators of interpretation/evaluation; see results in Section 3.2. Most of the corpus linguistic analysis was carried out using AntConc, which is a freeware corpus analysis toolkit (Anthony 2018).

In most of the analyses presented below quantitative techniques were supported with the qualitative close reading of concordance lines, i.e. sets of text fragments around given words. In some cases we read samples of concordance lines to get an overall impression of how a word is used in the corpus, e.g. to check whether 'hand' is used mostly as a verb or as a noun. In other cases we scanned concordance lines to look for patterning in the words around the given word, e.g. to see if it is regularly used as part of certain phrases. Thus, for us in this paper, the reading of concordance lines both provides a check on how we interpret quantitative results and identifies phenomena for further, potentially quantitative, investigation. In several places below we identify the need for deeper reading of whole descriptions alongside the images they refer to, in order to draw on a broader context and domain knowledge, and to inform the interpretation of how words are being used.

3. Results

A summary of results is presented in the following two subsections. In 3.1 we show and discuss results from a frequency-led analysis intended to characterize the common kinds of words in the corpus and hence give a view on the typical kinds of information provided by George's curatorial descriptions. This may be thought of as an analysis of curatorial voice as inclusion/omission. In 3.2 we summarize several complementary analyses of curatorial voice on a descriptive-interpretive/evaluative scale by measuring the presence of words, phrases and patterns that are hypothesized to be indicative of interpretation/evaluation. We are only able to present and discuss a small fraction of the results that were generated in these analyses; however, pointers are given for the interested reader to download and examine complete results sets.

3.1 Common kinds of information, and words that are not there

Table 2 lists the 100 most frequent words in the BMSatire Descriptions corpus, in frequency order.[7] Note that capitalized words were counted separately, and some non-alphabetic characters were included in the token definition, e.g. '&' in order to count the word '&c'. At the very top of the list, in common with most English-language corpora, we see so-called function words such as *the*, *a* and *of*. However, moving down the list we soon encounter words that reflect the specialist nature of the text, including words referring to print processes – *inscribed*; the spatial arrangement of print content – *left* and *right*; commonly depicted entities – *man*, *hand*, *head*, *hat* and *woman*; and actions – *stands*, *says*, *holding* and *wearing*. Further down the list we see parts of names – *Fox*, *Lord* and *John*.

> the (67,434), a (64,436), of (32,450), and (32,020), is (24,050), in (22,078), his (20,520), with (18,164), on (17,568), to (15,775), The (11,659), are (11,161), A (8881), by (7721), inscribed (7458), from (7268), which (6971), left (6596), right (6516), an (6428), at (6366), He (5571), stands (5377), her (4973), says (4816), who (4815), On (4699), he (4683), man (4500), two (4466), him (4310), hand (4272), head (4223), holding (3885), one (3841), holds (3773), as (3545), In (3317), behind (3083), large (3079), No (3000), other (2960), profile (2949), wearing (2880), hat (2862), saying (2832), up (2820), wears (2809), Behind (2792), has (2641), back (2573), sits (2328), it (2270), for (2177), out (2156), over (2154), table (2099), woman (1937), three (1892), towards (1873), or (1844), their (1815), design (1754), small (1711), that (1661), paper (1656), arm (1655), but (1635), Fox (1631), hands (1631), each (1629), title (1625), BMSat (1592), round (1589), men (1565), dressed (1559), them (1558), background (1542), extreme (1526), long (1514), His (1495), its (1433), looks (1431), ground (1398), stand (1362), Lord (1352), under (1340), Two (1337), See (1315), wall (1311), John (1306), &c (1296), have (1276), arms (1262), seated (1258), above (1255), She (1244), beside (1244), I (1234), Below (1221)

Table 2: The 100 most frequent words in the BMSatire Descriptions corpus, with frequencies in brackets.

When inspecting a frequency list it is important to consider that, in general usage at least, most words are polysemous and so we should not be too hasty in making inferences about the informational content of a corpus based on a frequency list alone. For example, the words *man*, *hand* and *head* can all be used as verbs, as well as nouns, but given prior knowledge of the corpus here we can probably be confident that, for the most part, they are being used to refer to things depicted in prints rather than to actions. That said, we should inspect concordances before making any strong claims about how a word is used in a corpus. Take for example the word *hands*, which is another example of a word that can be used to refer to a part of the body and to describe an action. Figure 1a shows 15 consecutive lines from the concordance for *hands* in the BMSatire Descriptions corpus (out of 1,631 lines) which have been sorted alphabetically according to the words to the right of *hands*. This shows that there are some instances of *hands* being used to describe an action, e.g. 'a lawyer hands a bowl to Bute'. However, this is not a fair representation of the word's usage because it only shows lines in which *hands* is followed by the word *a*, so it is more likely that *hands* is being used as a verb. It is not feasible to check all the hundreds or thousands of concordance lines for each of the most frequent words in a corpus, but a reasonable sense of typical usage

Left context		Right context
pigtail queue, ruffled shirt and laced waistcoat. He clutches in both	hands	a bowl in which is a spoon, evidently the soup-meagre of
bowls of blood; another stoops to look at Britannia; a lawyer	hands	a bowl to Bute, seated *BRACKETED* in a high-backed chair. In
large spoon over the copper, while with his left hand he	hands	a bowl to a military officer, probably Conway. He wears an
ersation, two saying, *TRANSCRIBED* and *TRANSCRIBED* . Next, a woman	hands	a box to a military officer, saying, *TRANSCRIBED* . He answers:
to a child. Five little girls dance in a ring, holding	hands.	A boy stands in an apple-tree throwing down the fruit to
in judge's wig and gown, walking, right to left, carrying both	hands.	A bundle of rods resting on his left shoulder. He is
sword and jack-boots and holds his plumed cocked hat in both	hands.	A buxom lady stands beside him, pointing to the statue; she
clutches Fox's shoulder; behind them sits M. A. Taylor, clasping his	hands.	A caricature of the French admiral in profile to the right,
of the lash between the tips of the fingers of both	hands.	A coarse bedroom scene, Sir F. Standish with a courtesan. On
s, *TRANSCRIBED* . From the *TRANSCRIBED* . A lady in bed *BRACKETED*	hands	a coin to an official with a book under his arm,
chair, one hand on the tap of the urn, while she	hands	a cup to a footman who stands with a tray. Sam
a wisp of straw, for a wreath *BRACKETED* holding in his	hands	a Currycomb and Horse-shoe Armed; On his head a louse couchant
toes, which are talons, project through his top-boots. To Fox he	hands	a dice-box and dice, for which Fox, hat in hand, holds
hree country-people and a child gape in astonishment holding up their	hands;	a fat alderman in a furred gown does the same; from
seated on benches on each side of a dinner-table. A waiter	hands	a foaming tankard of beer to a woman with a child

Figure 1a: 15 consecutive lines from the complete concordance of 'hands', with lines sorted alphabetically according to words to the right of 'hands'.

Left context		Right context
Duenna a portly woman at left with wide open mouth and	hands	on chest, turning to Isaac who cowers from her; oval. Scene
pointed. His successor *BRACKETED* stands in back view, legs astride,	hands	thrust deep into his coat-pockets, a bludgeon projecting vertica
other in angry controversy. One *BRACKETED* stands with legs astride,	hands	on hips, registering contempt, his head turned in profile toward
inscribed *TRANSCRIBED* . Charles X bends excitedly towards him, both	hands	raised; he says: *TRANSCRIBED* On the wall a framed map of
which is out at elbows, holds a large bone in both	hands.	Ribs of beef project from the pocket of his coat. He
aved, *TRANSCRIBED* . A man *BRACKETED* holds a fighting cock in both	hands.	Above his head is printed *TRANSCRIBED* . A man *BRACKETED* is
decorated with the royal arms. One standing behind holds in both	hands	a smoking receptacle. The third *BRACKETED* kneels supporting Fo
. Fox, straddling across a puddle, stoops to collect mud in both	hands,	looking at Hastings. Behind him a little ragged chimney-sweep s
the roof. A stout man *BRACKETED* turns the press, using both	hands	and a knee and leaning back to pull at the lever.
the design is engraved: *TRANSCRIBED* . A young man stands with both	hands	in his breeches pockets, a bludgeon under his right arm. He
BRACKETED* , in place of *TRANSCRIBED* , and another with two clasped	hands	and the word *TRANSCRIBED* *BRACKETED* . The last two are above
one of a set of four dancing a reel with crossed	hands;	he capers vigorously, his outstretched arm pointing as if in de
, in a gaberdine, in profile to the right, stoops with both	hands,	saying, *TRANSCRIBED* . A SERVANT OF ALL WORK. *BRACKETED* Brou
in profile to the left, towards Perdita *BRACKETED* , who clasps her	hands	ecstatically; Lady Worsley sits on the right. In the centre gall
ith a twisted grimace of sour calculation; he says: *TRANSCRIBED* His	hands	are clasped round Burdett's neck. Behind the pair, on the edge
profile to the right with bent back before an altar, his	hands	together. His unpowdered hair is cropped. From his pocket projec
fat John Bull, holding his cudgel between his legs, claps his	hands	delightedly, shouting, *TRANSCRIBED* . Behind him stands a Jew *
on a wire. Below it is engraved *TRANSCRIBED* . Wheeler, clasping his	hands,	says, *TRANSCRIBED* . A watchman stands behind him holding his
the design. The Prince of Wales stands, looking down dejectedly, his	hands	tied behind him by a rope held by Pitt, who is
ose to Irving capers a lank-haired fellow in old-fashioned dress, his	hands	together, his eyes raised sanctimoniously. In his coat-pocket is
4. John Bull, a stout, plainly dressed citizen, stands full-face, his	hands	in his coat-pockets, between Townshend *BRACKETED* and Hood *BRA
In the carriage a young man hides his face in his	hands,	while four ladies try to console him, one saying, *TRANSCRIBED*
to a sheet printed in two columns. John Bull, on his	hands	and knees, blows into the sails of a fleet of men-of-war
and ungainly, seated on a close-stool, his chin supported on his	hands.	His face is contorted and he clutches *TRANSCRIBED* , one of th
a paper: *TRANSCRIBED* ; he holds up two fingers. Lansdowne puts his	hands	over his ears, saying, *TRANSCRIBED* *BRACKETED* *BRACKETED* Bed
say: *TRANSCRIBED* ; *TRANSCRIBED* and, *TRANSCRIBED* The Quaker, his	hands	folded, answers: *TRANSCRIBED* . A Quaker stands by an open grav
dapper little man who crouches on the ground, holding up his	hands	behind his back. He sings: *TRANSCRIBED* Behind Blücher the Tsar
fat Dutchman in baggy breeches smoking a pipe, stands with his	hands	as well as legs, and he has a patch over one
sleeves are pinned to his coat, showing that he has lost	hands	on hips. Between them, a hand on the head of each,
the right and left of this sit jauntily two corpulent parsons,	hands.	His left hand supports his sabre. He wears the Waterloo medal
left, slightly smiling, two fingers extended, as if about to shake	hands	outstretched. The upper part of the Temple has undergone a chang
supplication. Spain and France are in flight towards the left, their	hands	and legs; a man full-face, holding his hat, stands stiffly on
. Two of the figures are puppets, with strings attached to their	hands	and legs; a man full-face, holding his hat, stands stiffly on

Figure 1b: Every fiftieth line from the concordance of 'hands', with lines sorted alphabetically according to words to the left of 'hands'.

can be gleaned by generating a sorted concordance and then displaying every nth word. For example, Figure 1b shows every fiftieth line from the concordance for *hands*, sorted according to words to the left of *hands*. The fact that each line here shows *hands* being used as a noun seems a reasonable basis to conclude that this is its typical usage in the BMSatire Descriptions corpus.[8]

In an attempt to give an overall view of the informational content of the corpus, Table 3 groups the 300 most frequent words according to the kind of information they provide. The main distinction is between 'content descriptors' i.e. words used to refer to and describe the entities and actions depicted in the prints, and 'meta/special words' i.e. words from the special language of curation. We separated a set of function words although there is a fuzzy boundary with the prepositions that we included in the 'meta/special' set, and we identified a small set of polysemous words that do not have a clear predominate usage in corpus. The decisions about where to place each word were, to a great extent, based on intuition; but sorted concordances were skimmed to check cases where there was thought to be potential multiple meanings. The sum of frequencies of the top 300 words is 727,157 out of a total of 1,129,475, so they alone account for approximately 64 per cent of the corpus. Thus, we might tentatively assume that Table 3 gives a reasonable impression of the kinds of information typically provided by George's descriptions.

Overall, the content descriptors give the impression that George concentrated on the physical *who* and *what* of the prints, with concrete rather than abstract nouns and verbs, corresponding with a rather literal and generic kind of description, although there is some naming of specific individuals. With regards to the differences in frequency between the content descriptors, e.g. between *man* and *woman*, we expect these are a consequence both of what is depicted in the prints and of George's choices about what to describe and what to omit. However, any such claims could only be substantiated with a thorough examination of the prints themselves and so text analysis can only provide a starting point for such an investigation.

Within the meta/special words there are many prepositions which could be taken to reflect George's preoccupation with conjuring the prints in the reader's mind, by specifying the spatial relationships between the principal entities and, at the same time, guiding the reader's eye around the print. Here, we note that George's catalogue was intended for use outside of the British Museum print room where readers would not have had access to the prints she described. This suggests that curatorial descriptions made before and after the availability of collections in microfilm or digitized form may need to be thought about differently, particularly where descriptions from both periods are used together in collection search and discovery.

We also see what we suspect are an unusually high number of capitalized function words and prepositions, e.g. *He*, *She*, *His*, *Two*, *Behind* and *Below*. We suspect that this is a result of George tending to write short atomized sentences and/or wanting to make certain information (such as position and clothing) more salient by giving it its own sentence, e.g. 'The guard is Lady Conyngham: she stands up, blowing her horn. She wears a guard's greatcoat and satchel over her dress and holds a blunderbuss.' rather than 'The guard is Lady Conyngham who is wearing a guard's greatcoat and satchel over her dress and holds a blunderbuss: she stands up, blowing her horn.' This is another example of how a simple corpus linguistic analysis may suggest a new line for further investigation. On a separate point, it should be noted that the high frequency of words such as *I* and *if* probably indicate that a certain amount of transcribed text (speech, labels and signs) remains in the corpus: our process for inserting '*TRANSCRIBED*' relied on pairs of quotation marks which are not always present in the text we selected. In the original volumes, transcribed text is formatted in italics and pairs of quotation marks were added inconsistently to de-italicised text as descriptions were entered into the British Museum internal database, later to be used as the basis for the British Museum's Collections Online service (Griffiths 2010).

Looking at the most frequent words tells us what is most commonly written in a corpus, but curatorial voice is also characterised by what words are not used. Thus we used a different corpus linguistic technique – keyness – to identify a set of words that are used unusually infrequently in the corpus, in comparison with a reference corpus. Keyness is computed as a measure of how much more or less frequently a word occurs in one corpus

Content descriptors	Nouns (mostly people, body parts, objects, clothes)	**man, hand, head, hat, table, woman, paper, arm, hands, men, ground, wall, arms,** heads, wig, dress, crown, sword, cap, coat, shoulder, door, feet, legs, chair, hair, book, officer, gown, uniform, pocket, shoulders, window, papers, mouth, breeches, neck, words, lady, pair, foot, flag, leg, *dog, horse, women, clouds, room, floor, bottle, group, crowd, ribbon, waist, tail, boy, feathers, soldiers, body, expression, boots, eyes, pipe, trousers, hats*
	Verbs (mostly physical actions)	**stands, says, holding, holds, wearing, saying, wears, has, sits, dressed, looks, stand, seated,** looking, standing, lies, answers, hangs, points, leans, raised, walks, *sit, say, wear, rests, pointing, held, supported, turns, kneels, covered, attached, lie, puts*
	Adjectives (physical properties, appearance)	**large, small, long,** open, fat, young, little, cocked, high, stout, *huge, short, elderly, old, ragged, tall, military, broken, decorated, tiny, thin*
	Names (mostly people)	**Fox, Lord, John,** Pitt, Napoleon, Duke, King, Wellington, Bull, George, Prince, Mrs, *North, Sheridan, House, Queen, Lady, Sir, Burke, III, Grey*
Meta/special	Art terms	**inscribed, design, title,** figures, figure, Plate, scene, *portrait, engraved, view, print, etched*
	Prepositions for spatial organisation	**left, right, behind, profile, Behind, towards, background, extreme, under, above, beside, Below,** side, front, foreground, between, Above, Beneath, below, Beside, centre, *facing, upper, corner, length*
	Misc.	**No., BMSat, See, &c.,** see, &, *BM, Satires*
Function words		**the, a, of, and, is, in, his, with, on, to, The, are, A, by, from, which, an, at, He, her, who, On, he, two, him, one, as, In, other, up, it, for, out, over, three, or, their, that, but, each, round, them, His, its, Two, have, She, I,** through, this, being, into, An, be, off, One, both, down, against, she, whose, four, all, They, while, From, very, they, another, others, across, not, also, before, you, much, *my, forward, about, like, Next, At, if, been, was, together, Three, more, next, some, these, third, This, only, second, 1, probably, where, All, There*
Words that are highly polysemous in the corpus		**back,** face, French, *part, full, drawn, fire, end, glass, showing, British, headed, smoke*

Table 3: The 300 most frequent words in the BM Satire Descriptions corpus, tentatively grouped by informational content: top 100 shown in bold; 101-200 shown in normal type; 201-300 shown in italics.

relative to another, and as a measure of the statistical significance of that difference. Table 4 shows a list of negative keywords, i.e. words occurring unusually infrequently in the BMSatire Descriptions corpus compared with, in this case, the British National Corpus (BNC).[9] Note, the BNC was intended to be representative of British English across a wide variety of written and spoken sources from the late twentieth century, thus the keyness results will partially be due to linguistic change since the time that George was writing and general differences between written and spoken language. That caveat aside, it is notable that George tends not to use past or future tenses (*was*, *had*, *were*, *said*, *will*), personal pronouns (*I*, *you*), modals (*would*, *could*), causality (*because*) and informal language (*it's* and *don't*). In part this suggests a systematic difference between George's curatorial voice and that of, for example, the contemporaneous curatorial work of Arthur Popham in the 1950 catalogue *Italian drawings in the Department of Prints and Drawings in the British Museum: the fourteenth and fifteenth centuries* (Popham 1950). This is a work which focuses on attribution and provenance and, in its description of objects, it more often switches between different tenses, makes causal connections and uses modals to hedge statements.

---

was, that, I, be, for, to, you, had, it, were, not, would, have, we, said, can, will, there, been, when, they, this, could, what, time, It, do, so, But, know, then, more, any, as, no, all, because, people, er, should, now, years, it's, got, work, about, if, such, get, did, or, don't, think, she, way, but, may, your, than, new, me, even, If, well, year, go, We, And, You, some, only, our, how, need, per, might, made, going, used, I'm, use, good, want, just, really, thought, It's, local, What, must, government, something, went, course, after, too, system, like, came, So

---

Table 4: A set of 100 negative keywords extracted from the BMSatire Descriptions corpus, i.e. words occurring unusually infrequently compared with the British National Corpus.

3.2 Curatorial Interpretation/Evaluation

The frequency-led approach in Section 3.1 gave the impression that the descriptions mostly refer to observable physical entities and their appearances and actions, and do so in relatively generic and objective terms. In this subsection we report five complementary corpus analyses that elucidate ways in which George also interprets and evaluates in places. Broadly speaking, for these analyses we work in a top-down manner, such that we start with an assumption about how an aspect of interpretation/evaluation is realized in words and then count those words. This approach allows us to observe phenomena that were missed by the frequency-led approach.

We recognize that knowledge is situated and that the notions of description and interpretation/evaluation are used by different people in different ways, so we do not attempt to make any hard and fast definitions here. However, for the analyses reported below, we found it helpful to conceive of statements as being more or less 'descriptive' and 'interpretive/evaluative'. At the description end of the scale are statements that could be considered objective and are less likely to be controversial: the writing and comprehension of such statements have minimal reliance on contextual information and specialist knowledge. Moving towards the interpretation/evaluation end of the scale, the writing and comprehension of statements about the contents of an image rely increasingly on contextual information and specialist knowledge, and hence tend to become more a matter of opinion. For example, at the descriptive end of the scale we would find a statement about observable entities and actions in generic terms, such as 'a woman runs down a street'. Naming the woman and the street would require some contextual information and/or specialist knowledge, so a statement such as 'Mary Smith runs down Oxford Street' would be placed along the scale, towards interpretation/evaluation. Further along still would be a statement like 'a woman dashes down a street' which suggests more urgency to the woman's action, and hints at a story in which her action is motivated by, say, the wish to catch something or the need to escape from something. At this end of the scale actions could be interpreted differently depending on the viewer's understanding of the story and the character. Also at this end of the scale would be statements that make

value judgments that are influenced by cultural norms, and perhaps personal prejudices, for example 'the extremely overweight man sits down'.

This characterization of description and interpretation/evaluation aligns with the definition of 'Content - description' in Spectrum 5.0 and with Panofsky's separation of description and interpretation in the iconographic method. In Spectrum 5.0 'Content - description' is defined as: 'A general description of a depiction in an object, or description of an object without making interpretation' (Collections Trust 2017) which suggests that a description comprises statements that refer to an object's self-evident content or visual appearance. In Spectrum, interpretation arises from the use of collections and can be used to improve catalogue records such that interpretation could be taken to mean both assertions on the meaning of an object and value judgements made about things depicted in an object. Previously, and contemporaneously to the production of the *Catalogue of Political and Personal Satires*, Erwin Panofsky's iconographic method found favour among scholars and curators who studied art and other objects with symbolic significance. Like the Spectrum standard, the iconographic method sought to separate description and interpretation.[10] Subsequently, drawing on Panofsky, Shatford proposed three distinct levels for analyzing the subject of images in order to index visual content: pre-iconographic – a generic description of entities and actions; iconographic – where specific people, places, etc. are named; and, iconological – the meanings conveyed by the image (Shatford 1986).

So how can we go about identifying instances of interpretation/evaluation in a corpus? Consider again the examples given previously: 'a woman runs down a street' and 'a woman dashes down a street'. The shift towards interpretation arises from referring to the same action in a different way by considering the character's mental state (emotions, needs, desires) and how this fits within an unfolding narrative. This kind of shift was accounted for by the narratologist Alan Palmer in his conceptualization of a 'Thought-Action Continuum', exemplified by the difference between the statements 'a person stands behind a curtain' and 'a person hides behind a curtain' (Palmer 2004). Palmer and Salway (2015) analyzed audio description – a verbal account of visual information provided for partially-sighted and blind audiences – in terms of the thought-action continuum by looking at how action descriptions were modified with extra words around verbs, and at how troponyms of generic verbs were used. This approach was our starting point in order to look for instances of interpretation/evaluation in the BMSatire Descriptions corpus, as reported in Sections 3.2.1 and 3.2.2.

3.2.1 Modifying action descriptions

In order to look for ways in which action descriptions may be modified to add interpretation/evaluation, we inspected sorted concordances for three of the most frequent verbs in the corpus – *walks*, *looks* and *says*. The overall impression was that most instances of action descriptions, at least for the selected verbs, are not modified and remain purely descriptive. However, there were some signs of interpretation/evaluation. The most striking case was the use of adverbs; see some examples in Table 5a. The use of these adverbs suggests that George has interpreted something about the mental states of the people depicted in the prints. Less frequent but still noticeable was the use of the pattern 'with + ABSTRACT_NOUN' as in 'looks with much satisfaction'; see Table 5b for examples.

Having identified an apparent pattern through manual inspection of concordance lines, i.e. the use of -ly adverbs to modify common verbs, we then made a more systematic analysis by filtering the frequency word list to include only the 1,009 words ending -ly. The 100 most frequent -ly words are shown in Table 6, in which we highlight those that we think are likely being used to interpret something about somebody's actions or to make an evaluative judgement such as about somebody's appearance.[11] The highlighted words have a total frequency of 2,940 which, spread across 9,330 descriptions, seems to be quite significant regarding the overall tone of George's curatorial descriptions, especially because we may expect many more instances to be found among the other 909 -ly words that occur between 1-25 times each. We speculate that some adverbs were used by George to give some sense of the story playing out in a print, e.g. 'walks dejectedly' suggests a character's disappointment with a recent event, and others were used evaluatively, e.g. *fashionably* and *ruffianly* could be shorthand for expressing a view about social status.

Henry Grey in gown and bands **walks amorously** with a meretricious-looking Susanna
Wellington, wearing a top-hat, **walks dejectedly**
Princess Elizabeth **walks possessively** with her husband
She **looks alluringly** at the spectator
The prostrate Pope **looks angrily** over his shoulder at Napoleon
J.B. **looks distrustfully** at Peel
The Duke **looks down disconsolately**
Richard **looks furtively** out at the wolves
who **says apprehensively**: *TRANSCRIBED*

Table 5a: Examples of verbs being modified with -ly adverbs to add interpretation of characters' mental states to the action description.

John **looks with much satisfaction** through a peep-hole in the wall
He turns round to **look with anxiety** at Sancho Panza
both turn their heads to **look with satisfied nonchalance** towards Napoleon
the Prince, who **says, with contemptuous arrogance**, *TRANSCRIBED*
The old man **looks with impotent anger** at his young wife
 *The drops his knife and **says with shocked disapproval**: TRANSCRIBED*
a fashionably dressed man who **walks with jaunty complacency**

Table 5b: Examples of verbs being modified using 'with'.

3.2.2 Lexical choices: troponyms

As well as modifying a verb with an adverb, it is also possible to add an interpretation to an action description by using a troponym, e.g. writing 'she saunters' rather than 'she walks' makes an interpretation of a character's mood or mental state. We used WordNet to make a list of troponyms for *say*, *look* and *walk* and then collated their frequencies;[12] see Table 7. The results show much higher frequencies for the generic verbs than for their troponyms, especially since it is questionable whether *shout* and *scream* should be taken as troponyms of *say*. This analysis seems to confirm the overall impression that George preferred an informally controlled language using mostly generic nouns and verbs, with adverbs added when necessary.

---

elderly (586), only (434), probably (408), partly (365), **fashionably** (323), respectively (276), ugly (249), slightly (217), evidently (215), **grotesquely** (198), melancholy (181), fly (177), heavily (173), **plainly** (167), **angrily** (154), apparently (152), **violently** (130), **fiercely** (123), family (110), immediately (109), **stiffly** (106), closely (105), burly (101), **intently** (96), **eagerly** (88), **delightedly** (85), July (84), enormously (82), freely (70), sly (70), horizontally (67), neatly (67), similarly (67), **arrogantly** (64), comely (64), scarcely (62), **slyly** (62), completely (61), Immediately (59), elaborately (59), **menacingly** (59), **wildly** (58), **gloomily** (57), Only (54), scaly (54), **furtively** (53), **savagely** (53), **tipsily** (53), vertically (53), broadly (52), grossly (52), **furiously** (51), poorly (51), **complacently** (50), **obsequiously** (50), equally (46), diagonally (45), faintly (43), **frantically** (41), **aggressively** (40), **derisively** (40), **despairingly** (40), **sternly** (40), **contemptuously** (39), nearly (39), presumably (39), realistically (39), **dejectedly** (37), oddly (37), roughly (37), **curiously** (36), **gleefully** (36), Probably (35), tightly (34), **triumphantly** (34), **fixedly** (33), **deprecatingly** (32), **impassively** (31), seemingly (31), **apprehensively** (30), **awkwardly** (30), **smartly** (30), **elegantly** (29), Folly (28), **vigorously** (28), butterfly (27), especially (27), **ruffianly** (27), **sourly** (27), **amorously** (26), **anxiously** (26), directly (26), **disconsolately** (26), immensely (26), **painfully** (26), **precariously** (26), symmetrically (26), **ferociously** (25), fully (25), **meditatively** (25)

Table 6: The 100 most frequent words ending -ly, with interpretive/evaluative words highlighted and with word frequency in brackets.

Verb	Troponyms
say (8228)	shout (637), exclaim (431), scream (112), cry (107), yell (25), plead (8), proclaim (6), enjoin (2), allege (0), aver (0)
look (2822)	gaze (265), stare (244), gape (80), peep (60), regard (53), admire (27), ogle (26), gawk (0), gawp (0), peek (0)
walk (1090)	stagger (122), stride (108), shuffle (6), strut (4), prowl (2), amble (1), lollop (0), saunter (0), skulk (0), stroll (0), stumble (1), traipse (0)

Table 7: Frequencies of some common verbs and their troponyms: the frequencies combine all forms of each verb.

There are some issues that should be noted with this kind of analysis. The chosen verbs – *walk*, *say* and *look* – were, in our judgement, the most likely of the frequent verbs to have troponyms used by George, compared with, e.g. *hold*, *wear* and *sit*. The selection of interpretive troponyms from WordNet also involved our judgment, i.e. we did not select every semantically related verb. Such judgements can be prejudiced by prior assumptions and ideally more time would be spent identifying potential troponyms for more verbs. Furthermore, the frequency counts for troponyms combine the counts for all the observed forms of the verb, e.g. *gaze*, *gazes*, *gazed* and *gazing*, which increases the chance of including the counts of words being used in other ways. Concordances were used to make a quick check that each word was being used as a verb, e.g. *cries* as a verb rather than as a noun. However, for the more frequent words it was not feasible to make an accurate count, so some of the troponym counts will be higher than they should be.

3.2.3 Lexical choices: synonymous pairs

It is not just with verbs that a curator has choices to make about using words with related meanings. Within the set of adjectives found in the 300 most frequent words (Table 3) there are two pairs that have roughly similar meanings but that suggest a different regard for their subjects – *fat* and *stout*, and *old* and *elderly*. So we might ask how and why George used these alternative forms and whether it has something to do with her making evaluative judgments.

A quick skim of concordances for the four words suggested some variation according to the gender of the person being described. To look into this further we needed a baseline of how much more often men and women are mentioned in the corpus. The frequency of *man* is 4,500 and the frequency of *woman* is 1,937, so *man* occurs about 2.3 times more frequently than *woman*.[13] Thus, if George were using *fat* and *stout* interchangeably to describe men and women then we should expect to see '*fat man*' and '*stout man*' occurring about 2.3 times more often than '*fat woman*' and '*stout woman*' respectively. In fact we observe that '*fat man*' (82) only occurs about 1.1 times more often than '*fat woman*' (73): in other words, George is using *fat* with *man* about half as often proportionately as she uses it with *woman*.

These numbers are laid out in Table 8, along with the equivalent numbers for *stout*, *old* and *elderly*. This gives the impression that George has a systematic preference for describing men as stout and elderly, and women as fat and old.[14] A little further investigation suggested that there may be a class basis for George's preferences, supported by the frequencies for these words being used with *lady*, i.e. as with man, there is a preference for *stout* and *elderly*: '*fat lady*' (16), '*stout lady*' (25), '*old lady*' (5), '*elderly lady*' (14). Further, although the numbers are small, when describing body size and age together, George seems to prefer using *fat* with *old*, and *stout* with *elderly*: '*fat old*' (14), '*stout old*' (2), '*fat elderly*' (13), '*stout elderly*' (20). Altogether this analysis suggests an evaluative aspect to George's descriptions whereby she accorded more respect to middle/upper class characters, perhaps reproducing her own views, the prevailing views of the 1930s-1950s British professional class, and/or the prevailing views in late-Georgian London.

3.2.4 Descriptive and interpretive cues

On a different tack, we note previous work that analyzed the language in a corpus of art gallery captions and found that the verbs *depict* and *convey* were frequently and consistently used to signal respectively: the description of pre-iconographic and iconographic image content, and the interpretation of iconological image content (Salway and Frehen 2002). For example: 'this work depicts two women eating seafood at the famous Parisian restaurant Prenier' and 'this composition conveys the claustrophobia of the interior of an omnibus'.

	man	woman	Ratio man:woman	Ratio compared with BASELINE
BASELINE	**4500**	**1937**	**2.3:1**	**x1**
fat + _____	82	73	1.1:1	x0.5
stout + _____	302	41	3.2:1	x1.4
old + _____	71	140	0.5:1	x0.2
elderly + _____	180	50	3.6:1	x1.6

Table 8: Showing how George tended to use 'stout' and 'elderly' with 'man', and 'fat' and 'old' with 'woman'.

Using a thesaurus and researcher judgment, two sets of verbs were identified corresponding to *depict* and *convey* and then counts of their frequencies in the BMSatire Descriptions corpus were made, and concordances were skimmed in order to see how George used them. The frequencies are shown in Table 9: here each frequency is for all observed forms, including nouns, e.g. *depict*, *depicts*, *depicting*, *depicted* and *depiction*. Examining the concordances for these words confirmed that they are mostly, if not entirely, cues for description and interpretation. However, it is not always the case that a particular verb was used consistently by George for either description or interpretation, e.g. 'Parted bed-curtains show Syntax asleep' (description) and 'The luxury of the room is shown by an arcaded wall' (interpretation).

	Frequencies of synonyms
depict	show (870), illustrate (321), depict (181), portray (2)
convey	represent (718), indicate (713), express (460), suggest (206), symbolize (87), imply (49), denote (33), convey (8), connote (6), evoke (6), impart (0), transmit (0)

Table 9: Frequencies of candidate cues for description and interpretation in the BMSatire Descriptions corpus.

The total number of instances of the candidate cues listed in Table 9 is 3,660 which, in a corpus of 9,330 descriptions, suggests a substantial phenomenon that warrants further investigation as an important facet of curatorial voice in this and perhaps other corpora. However, the potential to use computational techniques beyond counting frequencies and retrieving concordance lines is limited by the variety of ways in which these words are used, and further investigation must rely on close reading of concordances lines, possibly with regard to entire descriptions and the corresponding prints. To give a sense of the richness and complexity here, Table 10 gives some examples to reflect the typical, but not necessarily exclusive, usage of just the most frequent form for each of the most frequent candidate cues. In some cases a word switches from being a cue for description to a cue for interpretation, or vice versa, when used as part of a phrase, e.g. *depicted* and '*depicted as*'.

3.2.5 Hedging

The degree of confidence that a writer has in an assertion can be indicated using a hedge, e.g. 'it may be the case that…'. Hedging need not only be associated with interpretation and evaluation in curatorial descriptions, but we might hypothesise that it will be more likely used when a curator is going beyond what they consider to be self-evident in the object being

Cues	Examples
depicted (165)	In the upper part is <u>depicted</u> a meeting of the General Court Fores's shop is <u>depicted</u> with approximate correctness The room is <u>depicted</u> in detail The upper part of the print <u>depicts</u> topographically the English Channel
depicted as (15)	Mrs Billington is <u>depicted as</u> Clara in Sheridan's opera *TRANSCRIBED* Fox is <u>depicted as</u> a fox
illustration (63)	An <u>illustration of</u> the Reform agitation and its association with the demand for reduction of taxation and abolition of sinecures
illustration to (47)	<u>illustration to</u> Lavater's *TRANSCRIBED*
showing(430)	The interior of an inn bedroom, <u>showing</u> a large four-post bed with check curtains the lower part of an open window, <u>showing</u> a heap of money-bags
showing that (119)	clouds of smoke <u>showing that</u> a naval battle is in progress With the bales is a porter's knot, <u>showing that</u> Temple had tried to carry them
representing (330)	which he is showing to figures <u>representing</u> Britannia, Hibernia, Scotia, and America. a fleet of men-of-war <u>representing</u> British sea-power a skeleton, <u>representing</u> Death
indicated (375)	A mountainous landscape is <u>indicated</u> A window and a door are <u>indicated</u> The heads of a cheering crowd are <u>indicated</u>
indicated by (131)	The haste of Tom's exit is <u>indicated by</u> one stockinged foot its speed is <u>indicated by</u> the petticoats of Mrs Fitzherbert
suggesting (68)	On the right are books and an overturned stool <u>suggesting</u> a struggle with closed eyes <u>suggesting</u> death rather than sleep The second attendant wears a hat, <u>suggesting</u> that she is a milliner
expression(428)	His <u>expression</u> is one of perplexed anxiety
a|an * expression (206)	leaning back with <u>an</u> inscrutable <u>expression</u>
an expression of (128)	she looks up at him with <u>an expression of</u> dignified surprise

Table 10: Examples of the most frequently observed cues for description and interpretation.

described. In its own right hedging is interesting as an aspect of curatorial voice because it reveals the curator's voice quite explicitly as they flag their own uncertainty. It may also be the case that, when a catalogue is being revised or repurposed, hedges draw the attention of subsequent curators to more work that is needed to move towards certainty.

As a first step towards investigating hedging in the BMSatire Descriptions corpus we compiled a set of potential hedges and collated their frequencies in the corpus, albeit in a rather ad hoc manner: (i) we used a thesaurus to get synonyms of *perhaps*; (ii) from Table 6 we noted some frequent -ly words that look like hedges; (iii) based on prior knowledge of English grammar we considered some modal verbs; and, (iv) we skimmed a list of frequent n-grams to identify hedging phrases.[15] These candidate hedging words and phrases are shown in Table 11 with their frequencies.

In order to give a flavour of how George used hedges, Table 12 shows a selection of hedging examples, organized tentatively from George indicating more to less certainty. The

	Hedging words and phrases
Synonyms of perhaps	perhaps (206), Perhaps (29), possibly (15), Possibly (3), maybe (0), Maybe (0)
Frequent -ly words	probably (408), evidently (215), apparently (152), presumably (39), Probably (35), seemingly (31)
Modals	may be (98), seems (64), seem (48), seeming (8), could be (7), might be (3)
Frequent n-grams	appears to be (103)

Table 11: Frequencies of candidate hedging words and phrases.

examples show that hedging is actually used for qualifying statements about what is depicted and about attribution, as well as, if not more than, interpretive statements about symbolism and allusion. As with further investigation of the descriptive and interpretive cues mentioned in 3.2.4, further investigation of these hedges, e.g. to analyze variation over the course of George's work or variation between different kinds of prints, would rely mostly on close reading.

4. Discussion and concluding comments

The man, <u>evidently</u> Bute, is dressed partly in tartan
they are <u>evidently</u> constables and their hangers-on
one is Burdett, one wears a cocked hat, and is <u>probably</u> Cochrane.
The supporters are two stags, <u>probably</u> an allusion to the old gibe that citizens were cuckolds.
The panorama is <u>probably</u> by Heath who went to Glasgow to paint panoramas
The presiding alderman <u>appears to be</u> Harley
A small poodle also <u>appears to be</u> in distress.
Little Lansdowne capers behind him, <u>apparently</u> dancing a Highland fling
Seago wears a caplike wig, <u>apparently</u> of worsted.
The pyre <u>seems</u> to be made of money-bags
He wears what <u>may be</u> intended for a fool's cap.
The near leader <u>may be</u> Windham.
The ladies are <u>perhaps</u> dressed as shepherdesses.
Tierney squeezes a lemon into a glass, <u>perhaps</u> indicating his parsimony
Part of the interior of a large church, <u>perhaps</u> intended for Westminster Abbey.

Table 12: Examples of hedges, tentatively ordered from marking more to less certainty.

We have shown how simple corpus linguistic techniques can be used to characterize the language of curatorial descriptions in a variety of ways. In closing, we consider the potential value of such analyses for generating new knowledge about curatorial voice and for opening up new research directions, and what relevance such analyses may have for curatorial practice and for repurposing curatorial descriptions in digital search and access systems.

Section 3.1 showed how a frequency-led approach can generate a broad characterization of the informational content of descriptions. In the case of the descriptions in the *Catalogue of Political and Personal Satires* the emphasis is clearly on the physical *who* and *what* of the prints, along with specifying the spatial relations between entities and the use of some technical concepts related to print production. Further, there are signs that the descriptions tend to be written in the present tense with a formal style and a preference for short clear sentences. These are not surprising findings for anyone who is familiar with George's work, or the conventions of print scholarship; however, they serve to make the point that aspects of curatorial voice can be elucidated in this way, such that the approach could be usefully applied to less familiar collections. Furthermore, future work could use the BMSatire Descriptions corpus as a reference corpus for a keyness-based analysis of another catalogue, i.e. to identify words that are more or less frequent in that catalogue and hence elucidate differences in curatorial voice.

It was necessary to complement the frequency-led approach with some top-down analyses in order to identify the ways in which George included interpretation/evaluation in the *Catalogue of Political and Personal Satires*. The general point here is that important phenomena will be missed by a frequency-led analysis if they manifest in many different words, none of which is very frequent on its own. Section 3.2 presented analyses which were directed by preconceived ideas about how interpretation/evaluation might manifest. The results suggested that, in particular, George used adverbs when interpreting the mental states of the characters depicted in prints, and a small set of cues sometimes flag where she addressed the iconological content of prints. Such findings provide a basis for further investigations by making it possible to automatically retrieve examples of interpretation/evaluation for close reading, and perhaps by enabling quantitative analyses, e.g. to compare the amount of interpretation/evaluation for different kinds of prints. However, top-down analyses are subject to the problem of 'seek and ye shall find': that is, results are in part determined by preconceived notions of how the phenomenon of interest will manifest in words. For example, in Section 3.4 we used a thesaurus to generate a set of candidate descriptive and interpretive cues to look for, but later, when reading some descriptions, we noticed that George often used the phrases 'intended as' and 'intended for' as interpretive cues: these were not in the thesaurus and hence were missing from our analysis.

Broadly speaking, in our case study each of the analyses gave some insight into curatorial voice and/or suggested ideas for further investigation, even though none was conclusive on its own. We therefore believe that we have demonstrated that important aspects of curatorial voice do manifest in linguistic features that can be detected with corpus linguistic analysis. This means it is possible to leverage large catalogues for systematic empirical research into curatorial voice in ways that would not be feasible with manual methods alone. That said, it is apparent that addressing questions about curatorial voice also requires the close reading of whole descriptions, probably alongside collection objects and consideration of historical and cultural contexts. For example, in the 'Curatorial Voice' project we are evaluating the use of corpus analysis to address questions about how George's curatorial voice was shaped by historical and cultural factors, how it changed over time and how it subsequently shapes and constrains interpretation of the prints. Whilst we are encouraged by the outcomes reported in this paper, we recognize some issues that must be considered as we move forwards.

Foremost is the matter of what kinds of claims we can make based on the observed linguistic results from the BMSatire Descriptions corpus. First, the impression that George's voice is – at a surface level – clear, neutral, and confident, must be balanced against the expectations of institutional labour and of writing for an academic publisher (the *Catalogue of Political and Personal Satires* was published by Oxford University Press). Second, the BMSatire Descriptions corpus does not include all the text in the *Catalogue of Political and Personal Satires*. Although the descriptions are the most substantial part written by George, an

investigation of her voice should also consider the descriptions in relation to the introductory texts, indices and transcriptions of print titles. For example, George may not have speculated on the meaning of a given print in its description because she did so in her introductory essay to the volume containing it. Third, George's use of hedging could be due to a particular set of prints being unclear or lacking supporting documentation in the historical record. So, for example, it would not be appropriate to make claims about her use of hedging changing over time without examining corresponding examples of prints. Fourth and finally, frequency effects in the BMSatire Descriptions corpus may have less to do with George's voice and more to do with frequency effects in the prints she was describing, e.g. the relative frequency of the words *woman* and *man* will be determined in part by the relative frequency that women and men are depicted in prints, so again examination of prints is required before strong claims are made. Nevertheless, we maintain that some claims based on the observed linguistic results alone remain robust irrespective of these caveats, e.g. the statistical analysis of George's preferred usage of *fat*, *stout*, *old* and *elderly* (Section 3.3).

In a different direction, and more speculatively, the 'Curatorial Voice' project is looking to the broader applicability of using corpus linguistic techniques to elucidate curatorial voice(s) in a given catalogue. It seems to us that the kinds of corpus analysis presented in this paper could be usefully applied to current and future museological practice with regards to both the production of curatorial descriptions and to the re-purposing of legacy descriptions for accessing and analyzing collections. Given a set of guidelines for producing curatorial descriptions, corpus techniques could be used to check the extent to which guidelines are being followed at a macro-level, e.g. by identifying what aspects of objects tend to be referred to or not, and by gauging the overall extent of description versus interpretation/evaluation. Further, such analysis could form a basis for plans to edit and enhance a catalogue by providing areas to focus on and estimates of the person time required. It could also be that a corpus-based characterization of the language used in an exemplary catalogue could be used to develop or refine guidelines by identifying that catalogue's distinctive linguistic features.

When planning to use a legacy catalogue as the basis for accessing a collection through text-based searches, it would be helpful to have an overview of the common vocabulary in order to understand what search terms are likely to be effective. Thinking beyond free-text searches, the use of corpus techniques to identify linguistic structures such as descriptive and interpretive cues that flag iconographic and iconological content, and spatial relations between the entities depicted in an image, might enable the automatic generation of structured representations of image content for enhanced search and discovery.

Acknowledgements

Notes

[1]   https://curatorialvoice.github.io/

[2]   We use 'curatorial voice' to unify a dispersed literature on the production, authority, and legacy of descriptive acts in museums and the cultural sector more broadly. Zachary Kingdon's work on ethnographic collections (2019) is framed by the quiet archival trace of early curatorial processes. Katy Hill (2016) tells us that late nineteenth and early twentieth century object description was considered to be feminised labour. Bowker and Star (2000) describe how the motivations behind organizing logics rarely survive the deployment

of those organizing logics. And Agostinho *et al*. (2019) and Safiya Umoja Noble (2018) caution against combining historical data and algorithmic systems to make predications on the basis that historical data always represent the racial and gendered oppressions of their own time. For us, this literature is united by a commitment to push back against notions that object description was or is a 'common sense' activity, and therefore provides a foundation from which to investigate curatorial voice.

3   In brief: word lists identify a set of the most frequent words in a corpus; keyness analysis identifies sets of words that are unusually frequent or infrequent in a corpus compared with another corpus; and, concordances provide a convenient overview of how a particular word is used in a corpus (McEnery and Hardie 2012).

4   For a thoughtful introduction to what scholars mean when they invoke 'close reading' as method, and how those traditions might be usefully combined with computational approaches to text, see Eve (2019), especially 3-11.

5   https://public.researchspace.org/sparql

6   The text strings in quotation marks and the bracketed text strings that were replaced are available in Baker and Salway (2019a).

7   For details of AntConc setting and the complete frequency list see frequencyList.txt in Baker and Salway (2019b).

8   For details of how to reproduce the concordances discussed here, and in Sections 3.2.4 and 3.2.5, see concordances.txt in Baker and Salway (2019b).

9   See keyness.txt in Baker and Salway (2019b) for technical details about how keyness was measured and the complete results file.

10 As the iconographer Ernst Gombrich was a known correspondent of M. Dorothy George and as the objects she described were feted for their iconographic potential, the iconographic method provides a further framework for thinking about the distinction between description and interpretation. But we also note that iconography was discredited for privileging the authority of knowledge found principally in the Global North (Cassidy 1993). And so drawing on the history of knowledge organization (Turner 2017), we acknowledge that interpretation is entangled with content description.

11 See ly_words.txt in Baker and Salway (2019b) for the complete list of -ly words.

12 WordNet is an online thesaurus-like resource, https://wordnet.princeton.edu/

13 A further line of inquiry might look into why this difference comes about, i.e. whether as a reflection of the content of the prints and/or George's selection of what to describe. Of course there are other ways in which women and men are referred to in the descriptions, e.g. pronouns, proper names, *lady* and *gentleman,* and gendered professional roles: whilst these would be important to account for in extended work into the representation of gender, it is not relevant for the point we are making in the current analysis.

14 This observation could and should be tested in further work with a measure of statistical significance.

15 This was a list of all word sequences between two and five words long that occurred 100 or more times in the BMSatire Descriptions corpus.

References

Anthony, L. (2018) 'AntConc (Version 3.5.6) [Computer Software]', Tokyo, Japan: Waseda University, http://www.laurenceanthony.net/software

Agostinho, D., D'Ignazio, C., Ring, A., Thylstrup, N.B. and Veel, K. (2019) 'Uncertain Archives: Approaching the Unknowns, Errors, and Vulnerabilities of Big Data through Cultural Theories of the Archive', *Surveillance & Society* 17 (3/4) 422–41, https://doi.org/10.24908/ss.v17i3/4.12330

Baker, J. and Salway, A. (2019a) 'Creation of the BMSatire Descriptions corpus', Zenodo, http://doi.org/10.5281/zenodo.3245037

(2019b) 'Corpus Linguistic Analysis of the BMSatire Descriptions corpus', Zenodo, http://doi.org/10.5281/zenodo.3245017

Bowker, G.C. and Star, S.L. (2000) *Sorting Things Out: Classification and Its Consequences*, Cambridge, Mass: The MIT Press.

Cassidy, B. (1993) 'Introduction: Iconography, Texts, and Audiences', in Brendan Cassidy (ed) *Iconography at the Crossroads: Papers from the Colloquium Sponsored by the Index of Christian Art*, Princeton University, 23-24 March 1990, 3-15, Princeton, NJ: Index of Christian Art, Department of Art and Archaeology, Princeton University.

Çelik, Z. (1996) 'Colonialism, Orientalism, and the Canon', *Art Bulletin*, 78 202-5.

Collections Trust (2017) Spectrum 5.0. 'Introduction to Spectrum 5.0', accessed 12 June 2019, https://collectionstrust.org.uk/spectrum/spectrum-5/

Duncan, C. (1995) *Civilizing Rituals: Inside Public Art Museums*, London; New York: Routledge.

Eve, M. (2019) *Close Reading with Computers: Textual Scholarship, Computational Formalism, and David Mitchell's Cloud Atlas*, Stanford: Stanford University Press.

George, M.D. (1935-1954) *Catalogue of Political and Personal Satires Preserved in the Department of Prints and Drawings in the British Museum*, London: British Museum Publications Limited.

Greene, C.S. (2016) 'Material Connections: "The Smithsonian Effect" in Anthropological Cataloguing', *Museum Anthropology*, 39 (2) 147-62, https://doi.org/10.1111/muan.12121

Griffiths, A. (2010) 'Collections Online: The Experience of the British Museum', *Master Drawings*, 48 (3) 356-67.

Hill, K. (2016) *Women and Museums 1850-1914: Modernity and the Gendering of Knowledge*, Manchester: Manchester University Press.

Kingdon, Z. (2019) *Ethnographic Collecting and African Agency in Early Colonial West Africa: A Study of Trans-Imperial Cultural Flows*, New York; London: Bloomsbury Visual Arts.

McEnery, T. and Hardie, A. (2012) *Corpus Linguistics*, Cambridge: Cambridge University Press.

Noble, S.U. (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York: New York University Press.

Palmer, A. (2004) *Fictional Minds*, Lincoln; London: University of Nebraska Press.

Palmer, A. and Salway, A. (2015) 'Audio Description on the Thought-Action Continuum', *Style*, 49 (2) 126-48.

Popham, A.E. and Pouncey, P. (1950) *Italian Drawings in the Department of Prints and Drawings in the British Museum: The Fourteenth and Fifteenth Centuries*, London: Trustees of the British Museum.

Putnam, L. (2016) 'The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast', *The American Historical Review*, 121 (2) 377-402, https://doi.org/10.1093/ahr/121.2.377

Salway, A. and Frehen, C. (2002) 'Words for Pictures: Analysing a Corpus of Art Texts', in *Proceedings of TKE 2002: 6th International Conference on Terminology and Knowledge Engineering*, Nancy, France, 28-30 August 2002, Le Chesnay Cedex, France: INRIA.

Shatford, S. (1986) 'Analyzing the Subject of a Picture: A Theoretical Approach', *Cataloging & Classification Quarterly*, 6 (3) 39-62.

Turner, H. (2017) 'Organizing Knowledge in Museums: A Review of Concepts and Concerns', *Knowledge Organization*, 44 (7) 472-84, https://doi.org/10.5771/0943-7444-2017-7-472

Authors

*****Andrew Salway**
Sussex Humanities Lab
Silverstone SB211
Arts Road
Falmer
East Sussex
BN1 9RG

andrewsalway@gmail.com
07956 923939

Andrew Salway is a research fellow in digital humanities working at the Sussex Humanities Lab, University of Sussex. He has 20 years' experience of developing and applying computational text analysis techniques in interdisciplinary research projects, to improve access to digital media archives and to develop new modes of humanistic and social science inquiry.

******James Baker**
Sussex Humanities Lab
Silverstone SB211
Arts Road
Falmer
East Sussex
BN1 9RG

drjameswbaker@gmail.com
07875 021299

James Baker is a Senior Lecturer in Digital History and Archives at the University of Sussex and at the Sussex Humanities Lab. He is a Software Sustainability Institute Fellow, a Fellow of the Royal Historical Society, and holds degrees from the University of Southampton and latterly the University of Kent, where in 2010 he completed his doctoral research on the late-Georgian artist-engraver Isaac Cruikshank.