CA Journal of Cultural Analytics

Bowman, Mike. "Text-Mining Metadata: What Can Titles Tell Us of the History of Modern and Contemporary Art?" *Journal of Cultural Analytics*, vol. 8, no. 1, June 2023, https://doi.org/10.22148/001c.74602.

ARTICLE

Text-mining metadata: What can titles tell us of the history of modern and contemporary art?

Mike Bowman[1]

[1] Birkbeck, University of London

Keywords: text-mining metadata, contemporary art, art titles, distant reading

https://doi.org/10.22148/001c.74602

Journal of Cultural Analytics

Vol. 8, Issue 1, 2023

The use of statistical text-mining to investigate the linguistic structure of textual resources has received limited attention in digital art history. In this paper the question I address is that of what text-mining titles as given in metadata can tell us about the history of modern and contemporary art. To investigate this question I constructed a dataset from the metadata for over 170,000 artworks, drawing on the online collections of 133 art museums in 30 countries. The use of topic modelling, parts-of-speech tagging and word counting allows me to identify large-scale and long-run patterns in the language used in the titles of those artworks. I set out an art historical reading of those patterns in which artistic interests signalled by the language used in titles come and go and are re-inflected, epistemic perspectives on the kinds of knowledge art can or should engender change, and artists engage with the ways that the title functions. My 'distant' reading is consistent with the canonical history of modern and contemporary art and cuts across the particularities of artist, period or movement which often feature in such accounts, providing a fresh perspective on that history. It also complements and extends the scholarship on the history of titles in the visual arts. The analytical framework and the dataset I have developed are not limited to answering the question addressed in this paper, and I consider some of the possibilities for future work.

1. Introduction

In an influential 2013 paper that set out the conceptual parameters within which much of the subsequent debate has been conducted, the visual theorist Johanna Drucker surveyed digital humanities activity in art history (Drucker). Drucker drew a distinction between *digitized art history*, which focuses on building digital resources and tools, and *digital art history*, where 'digital methods change the way in which we understand the objects of our enquiry' (Drucker 7). Whilst substantial progress had been made on the former, Drucker maintained that '… to date no research breakthrough has been made …' in digital art history (Drucker 5).

Drucker proposed several areas in which digital art history could develop, including engagement with the textual resources of art history. Textual materials such as art criticism, art history, philosophical writings on aesthetics, artist writings, exhibition catalogues and inventories are essential to the work of the art historian. Drucker expressed the belief that as more of them became available online, digital engagement would escalate dramatically. Tracing changes in terminology could 'expose aspects of the field that could only be partially glimpsed through traditional reading and study' (Drucker 8). Drucker also remarked that more sophisticated text-mining techniques had the potential to be 'touchstones of new practice and thought' (Drucker 10).

Art historians have made extensive use of digitized resources such as catalogues and inventories to explore questions around artistic networks, investigating the inter-relationships of artists, dealers, patrons and others in the art world (Fletcher and Heimreich; Joyeux-Prunel; Lincoln; Quach McCabe). However, much less has been done to investigate the linguistic structure and content of those resources (one notable exception is Greenwald), and studies utilising sophisticated text-mining methods have often been led by scholars in other fields (Smeets et al.; Garcia-Zorita and Pacios).

In this article I draw on the metatdata available in online art museum collections. The question I address is that of what text-mining those resources can tell us about the history of modern and contemporary art. My focus is on the titles of the artworks. A small number of articles and monographs have been published on the history of titles in the visual arts, (Gombrich; Bann; Welchman; Hoek; Yeazell). Titles may also play a key role in the readings developed by critics or art historians. The work of all these scholars shows that titles are not merely ways of identifying works of art but are of considerable art historical interest in their own right. Titles have mattered to those involved in the production, distribution and reception of art. They can indicate the subject matter of the work, have been a site for artistic innovation, can have an ideological effect, and influence the general viewer's or the critic's understanding and appreciation of the work. Indeed, one of those authors, Stephen Bann, observes that looking at titles is one way of 'retracing the history of modern art as a whole' (Bann).

What is common to the art historical literature on titles is that the authors look at individual artists, periods, or movements and give close readings of the ways titles contributed to the meanings of the works they named. In contrast, the account I set out in this article adopts a distant viewpoint. As self-contained textual units, titles are well-suited to sophisticated statistical approaches such as those developed for natural language processing. Such tools have been used by literary historians to look at a wide range of questions such as influence between authors, the nature of fictional language, how conceptions of gender difference have changed, the emergence of literary style, and racial discourse in Japanese literature (Jockers; Allison et al.; Piper; Underwood; Long). Literary historians have also used simple word counting of the titles of literary works to consider questions such as how literature reflects ideologies of gender, and to investigate their lexical richness (Moretti; Jockers). Looking at titles in aggregate, I bring both kinds of statistical technique into art history to explore the ways titles have been used. They allow me to identify large-scale and long-run patterns in the language used in titles, to which I give art historical interpretations.

To develop the readings I present in this article I needed a conceptual framework cutting across the use of titles by individual artists. And so, I have drawn on the work of scholars who have investigated the question of what a title is and how it functions. The literary theorist Gerard Genette was one of the first to provide such an account (Genette and Crampé). More recently, the semiotician Josep Besa Camprubi has surveyed the relevant literature (Besa Camprubi). Although the scholars he reviews differ in their methods of analysis and terminology, Besa Camprubi finds there is a 'remarkable convergence' between them over the three functions a title can be used to perform (Besa Camprubi 8). The first function Besa Camprubi identifies is naming. Titles always have a nominative function. The second function of the title is the semantic one of saying something about the work it names and contributing to the meanings it is given. Titles can also function 'seductively', and may be used to attract the attention of the reader.

In the next section I provide a critical review of my data sources and my dataset construction. This is followed by my distant readings of titles to give a perspective on the history of modern and contemporary art, including a critical comparison of my work with other authors who have written on the titles of art works and on the history of modern and contemporary art. In the concluding section I summarize those readings. and consider the potential for future work using the analytical framework I develop or my dataset, and reflect on the methodological lessons that can be drawn for those working in the digital humanities. The Appendix contains the main data tables, and the full results of my modelling can be found at https://doi.org/10.7910/DVN/MDGEYO.

2. Data Sources and Dataset Construction

Art museums are increasingly making their collections available online, and, although some such as the National Galleries of Modern Art in Italy and India had yet to do so when I was compiling my dataset, those that did allowed me to put together a dataset with a broad geographical spread. I included in my dataset institutions that identify themselves as being a modern, contemporary, or modern and contemporary art museum. I also included collections of modern, contemporary or modern and contemporary art held by other museums. I restricted my dataset to works given as paintings, sculptures, installations or works in new media, or, if that was not available, excluded works given as photographs or as executed on paper such as drawings, prints and editions. The reasons for this choice are that the former have standardly had their own titles over the whole of the period I examine in this article, whereas that was not the case with the latter. The works in most of the collections I sourced for my dataset are predominantly from the twentieth and twenty-first century and so for my analysis I restricted my dataset to artworks created in the years from 1900 to 2019. The metadata I collected for each work included the name of the artist, their nationality or place of birth (if given), the title, and the date or period of creation.

I cleaned and processed this metadata in several ways for inclusion in my dataset. I removed duplicates as identified by their collection numbers, as some collections have multiple records for the same object, for instance presenting images of the parts of an installation. In a manual pass through the data I excluded curatorial additions involving the use of text in brackets where I could be confident that was the case, for instance to give the native-language translation of a foreign-language title. However, it is very likely I have not identified all instances of curatorial additions in brackets. With a small number of paintings, the collection entry records titles for works on the front ('recto') and back ('verso') of the canvas. In these cases I retained the recto title.

In addition to art museums in English-speaking countries, several of the museums in my dataset provide titles in English. To allow a unified analysis of language use all across the titles in my dataset, titles in the native language of the country in which the museum was based were translated into English using Microsoft Translator.[1] Instances of non-native and non-English titles were identified manually, consolidated and processed through Microsoft Translator.[2] Microsoft Translator can give multiple translations for the same word. For instance, French titles which begin 'd'après' and which indicate a work 'after' that of another artist were translated into English as beginning 'after', 'according to', and 'based on', and the Portuguese phrase 'natureza morta' was translated as 'still-life' and 'dead nature'. In a manual pass I identified common multiple translations and replaced them with the appropriate English translation. In the examples just given, 'd'après' was translated as 'after', and 'natureza morta' as 'still-life'. This was supplemented by manual examination and automated replacement of common mistranslations, such as rendering the French title 'Nu' as 'Naked' rather than 'Nude'. In a final step I transformed instances of British English, such as 'colour' and 'realise', to American English, respectively 'color' and 'realize'.

Microsoft Translator provides literal translations with at times an unnatural word order in English, and my dataset is likely to include a residual number of mistranslations, multiple translations, and spelling variations. Notwithstanding these shortcomings, the English translations in my dataset are adequate for the readings I give in this article where titles are looked at in aggregate and large-scale trends are given art historical interpretations. It is not my aim to provide close readings of the titles of individual artworks, where an appropriate translation may be a key part of the interpretation. In total, 34% of titles, or around 60,000 in total, were translated into English using Microsoft Translator.

[1] The Microsoft translator can be found at https://www.microsoft.com/en-us/translator/.

[2] I investigated the use of several automated language recognition applications to support this task but found that the accuracy rates were too low to make the process more efficient than one involving manual examination only.

I made further transformations of the metadata. In titles there are several ways of indicating a numbered work, including 'number', 'no.' and '#'. I replaced all of these with 'number', to allow the overall trends in the numbering of works to be identified. In a semi-automated pass I transformed all Roman numerals to Arabic numerals, to allow the levels of use of each particular number to be identified. Manual checking was required to identify uses such as that of the first-person singular pronoun 'I', of 'x' as used in the sizes of paintings or to signify something unknown, of 'v' as a shorthand for versus, and of Roman numerals in proper names. For each work I recorded a single year as that in which it was created. In the large majority of cases the online collections record a year, a season, or a date in a year on which the work was created, and that year was entered into the dataset. If the date was given as 'around', 'before' or 'after' a certain year I recorded that year. In some cases the metadata gave a range of years, often the decade in which the work was executed. I excluded items where the ranges were longer than 20 years, otherwise I took the middle of the range as the year of creation. This will have introduced a bias into my dataset, boosting the number of works recorded as being created in those middle years, but will not have affected the long-run trends.

Altogether, my dataset includes the metadata on 171,000 artworks by around 35,000 artists from the online collections of 133 art museums in 30 countries covering Asia, Australasia, Europe, North America, and South America. Figure 1 presents the number of entries in my dataset for works with a given or estimated creation date in each decade from the 1900s to the 2000s. In terms of the number of words, the documents vary in size from 13,986 words in the titles of works created in the 1900s to 70,629 words in the titles of works created in the 1980s. Table 1 in the Appendix gives a list of the countries housing the collections included in my dataset, together with the associated number of entries. A complete list of the museums included in my dataset can be found at https://doi.org/10.7910/DVN/MDGEYO.

The interpretations I offer in this article are of modern and contemporary art as it has been constructed and presented through the collecting policies and practices of those 133 art museums. Over and above the geographical constraints, the selection of period, and the choice of media I have already discussed, my reading will also reflect the inclusions or exclusions of certain types of artist or work in those collections. Several of the collections are heavily weighted towards artists in the canonical history of modernism, and some include a large number of works left to the institution by artists in their estates. The art historian Terry Smith has argued that 'no museum has succeeded fully in becoming a site for contemporary art', and 'in general … the leading centers in Euro-America celebrate those artists who … perpetuate traditional subjects' (Smith 45). Women are substantially under-represented in the collections included in my dataset, with that bias being most acute with artists active in the early-twentieth century. I have not made any adjustments to the data to reflect these factors as there is no principled way on which to make them.

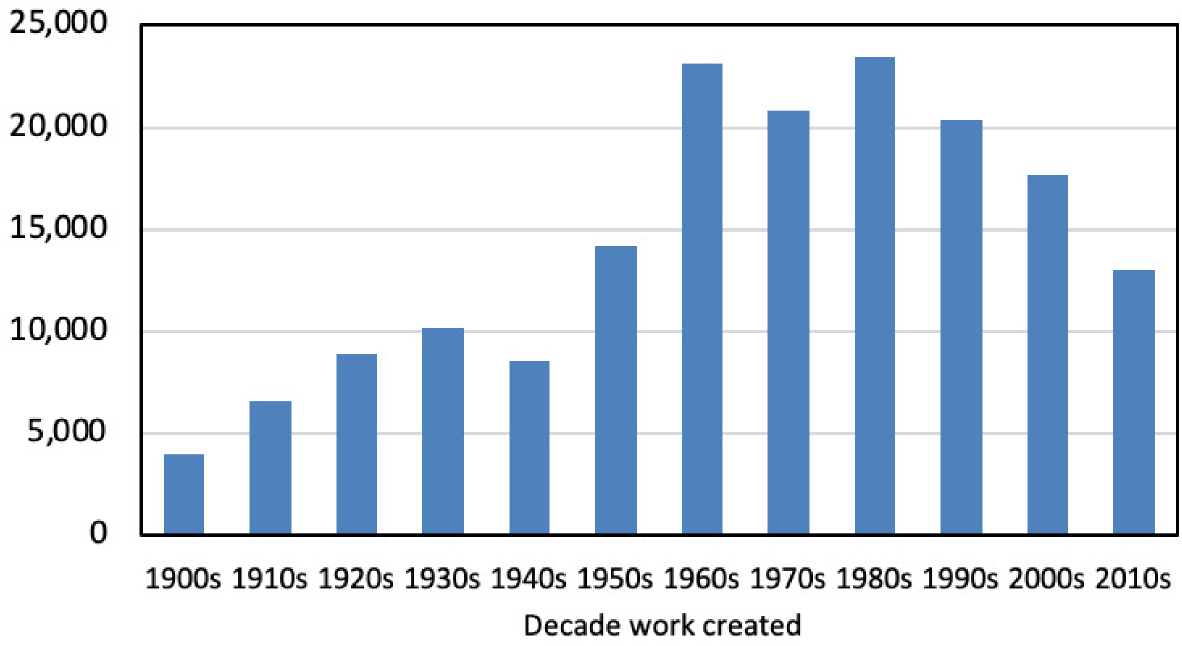

Figure 1. Number of dataset entries for works created in each decade, 1900s to 2010s.

My analysis of trends over time in the language used in titles is complicated by the fact that in some cases the titles as they are now presented would not have been part of the public face of titling at the time the works they name were created. The titles of works may have been changed after they were first exhibited, either by the artist, an owner, or a dealer. In other cases, works may not have been put on public display when created, for instance remaining with the artist until their death. With some works the original title may have been lost and the work recorded as 'no title' or 'untitled'.

Without a detailed study of the history of a sufficiently large selection of the titles in my dataset it is not possible to say with a high degree of confidence how the readings I present in this article have been influenced by these factors. However, I have reviewed the Catalogue Raisonnés of seven artists and this gives some comfort the impact is not substantial. In cases of re-titling or of titles given to works left in estates, the titles may be of the same kind as the titles of other works displayed by the artist at around the time they were created. For instance, Mark Rothko often re-titled by re-numbering paintings, and the titles used by the gallery managing his estate mainly consisted of color words, a practice Rothko also followed with some of his paintings. In addition, works where the original title has been lost are most likely to date from the early twentieth century. Works from that time recorded as 'no title' or 'untitled' make up a very small proportion of my dataset.

3. Titling in Modern and Contemporary Art

3.1. Topic Modelling - Titles as Signals of Artistic Interests

A topic model is a generative statistical model of a collection of documents (Blei et al.; Boyd-Graber et al.). Each 'topic' is a distribution over all the words occurring across those documents, and each document is modelled as distribution over all the topics. Prominent topics that feature with high weight across several documents represent recurring patterns of use of the words of highest weight in those topics. Topics prominent only in a single text or a small number of texts contain words whose use is distinctive of those documents. Topic modelling is therefore a useful tool for the identification of clusters and other patterns of language use across a body of texts.

The implementation of topic modelling I utilised is that developed by Andrew McCallum as part of his MALLET toolkit.[3] I followed the standard approach to the development of my topic model in including only 'content' words, which were also lemmatized. I excluded all punctuation marks and other non-alphanumeric symbols except for the hyphen. The hyphen is commonly used in titles such as 'self-portrait', and to have excluded it would have split such titles into two separate words each of which would have been treated independently in the modelling. It is common in topic modelling to exclude proper names. I investigated this option using the named entity recognizer developed by the Stanford University Natural Language Processing Group.[4] However, the level of accuracy was low, with many proper names missed and many phrases mis-identified as proper names, and so I retained proper names in the titles to be modelled. I then experimented with varying the ways in which titles are grouped and the number of topics to be identified. My aim was to see if a model could be developed that supported an art historical interpretation and where there were few or no redundant topics, in the sense of a topic of low weight in all of the document models. The topic model meeting those aims looked at the titles used for works created in each decade from the 1900s to the 2010s, twelve documents in total, and modelled them as being generated by 20 topics. Models that looked at each title as a separate text or grouped together date-ordered titles in 100 title parcels or by year were not as well-suited for the identification of long-term trends. Models with fewer than around 20 topics did not discriminate between the words used in the titles of works created in different decades as well as the 20-topic model. Those with more always included redundant topics, which typically had a weight of 3% or less in all of the documents modelled.

[3] Andrew Kachites McCallum, 'MALLET: A Machine Learning for LanguagE Toolkit.', which can be found at https://mimno.github.io/Mallet/index.

[4] The Named Entity Recognizer can be found at https://nlp.stanford.edu/software/CRF-NER.shtml#:~:text=Stanford%20NER%20is%20a%20Java,or%20gene%20and%20protein%20names

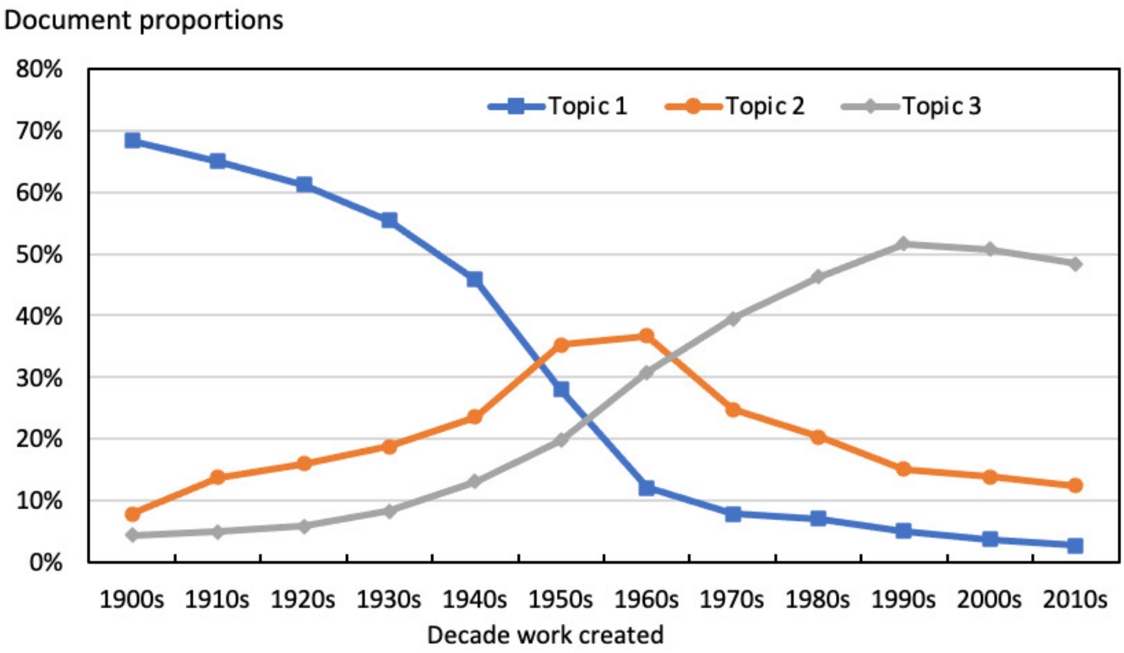

Figure 2. Topics 1, 2 and 3 proportions in the model for the titles of works created in each decade, 1900s to 2010s.

My model is dominated by three topics. For the titles of artworks created in every decade from the 1900s to the 2010s the most prominent topic in the model is one of those three topics, their combined weight is never less than 60%, and for all decades bar the 2000s and the 2010s they represent three of the four topics of highest weight. No other topic is significant in the model for more than two decades. Plotting the weights of the three dominant topics by time shows an inter-woven cyclic pattern of change, as presented in Figure 2. The first topic, 'topic 1', dominates in the models for the titles of artworks created in the early decades of the twentieth century. 'Topic 2' rises to be the most prominent in the models for the titles of artworks created in the 1950s and 1960s before declining. 'Topic 3' is the most prominent for the decades from the 1970s to the 2010s, although, as can be seen from Figure 2, its importance has also fallen back in recent decades.

To give an art historical reading of these trends I turned to the important words in each topic. As will be seen, these support the kind of rich interpretation of topic models developed by Andrew Goldstone and Ted Underwood, and Andrew Piper (Goldstone and Underwood; Piper 66–93). As these scholars stress, topics do not merely identify univocal themes. The ways words re-appear across topics, the importance of words within topics, and the prominence of topics in a text and across texts are all of interpretive significance.

As I develop my reading in this section and the next I will compare it with the work of John Welchman, whose monograph *Invisible Colors* provides an in-depth study of titles by over 200 modern and contemporary artists, and identifies some long-run themes in their use, and with the canonical history of modern and contemporary art (for recent surveys see Arnason and Mansfield; Bois et al.; Smith and, Hopkins). It is not possible to summarise and adequately reflect the complexity of that history in the space available in this article, but I will suggest some of the ways we might compare them. As will be seen, my readings tend to confirm the established narrative, giving some new ways of looking at that history. In my reading I will also return to the three basic functions of the title I discuss in the introduction.

A perspicuous way to examine the important words in a topic is through the word cloud, which presents the frequencies of the words in a text in a manner in which that information can be readily assimilated. In my case I generated the word clouds with frequencies given by the weights of the 50 most important words in each topic. The cut-off at 50 words did not represent a sharp break in the weights of the words in the topic. Rather, I chose it as using a smaller cut-off such as 20 words excluded some of art historical value, whereas including more words added little to my readings. In each word cloud, it is only the relative size of words that is of interpretive significance - the larger the word the more weight it has in the topic. The relative position and orientation in the cloud are not important – two words being close together or having the same orientation does not indicate any relationship between them. All the words bar the first person singular 'I' are in lower case. The basic meanings of the words in the clouds for topics 1, 2 and 3 will not have changed over the course of the period I have modelled, although, as we will see, some of their associations have.

The word cloud for topic 1 is presented in Figure 3. As can be seen, the most important words are predominantly those relating to the generic categories of landscape, portrait, and still-life, or to figurative subject matter including women, the seasons of the year, gardens, nudes, children, and flowers. Used in titles they would most likely have expressed some sort of artistic engagement with the figurative tradition or with figurative subject matter. As a shorthand, we can characterise topic 1 as 'figurational'. These figurational interests were the most prominent large-scale signals sent through the language used in titles during the first five decades of the twentieth century.

The generic hierarchy as institutionalised in Academic doctrine and teaching was fundamental to the nineteenth-century European art world, including the practice of artists and the judgements of critics, who worked both within and against the system (Mainardi; Teukolsky). Whilst the late nineteenth century saw the demise of history painting, topic 1 shows that it also remained important to many artists to present their work in relation to that conceptual and practical framework.

What displaced the figurational uses of titles were mainly those associated with the important words in topic 2. Its word cloud is presented in Figure 4. As that shows, the most important words include those describing the type of work – a painting or a sculpture for instance – or the formal elements of a painting such as shapes and colors, and the way in which they are arranged,

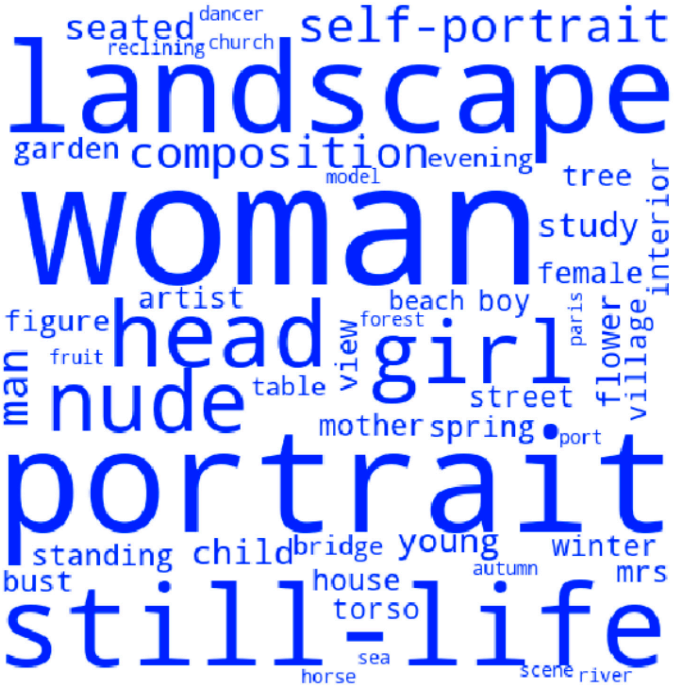

Figure 3. Word cloud for the 50 most important words in Topic 1.

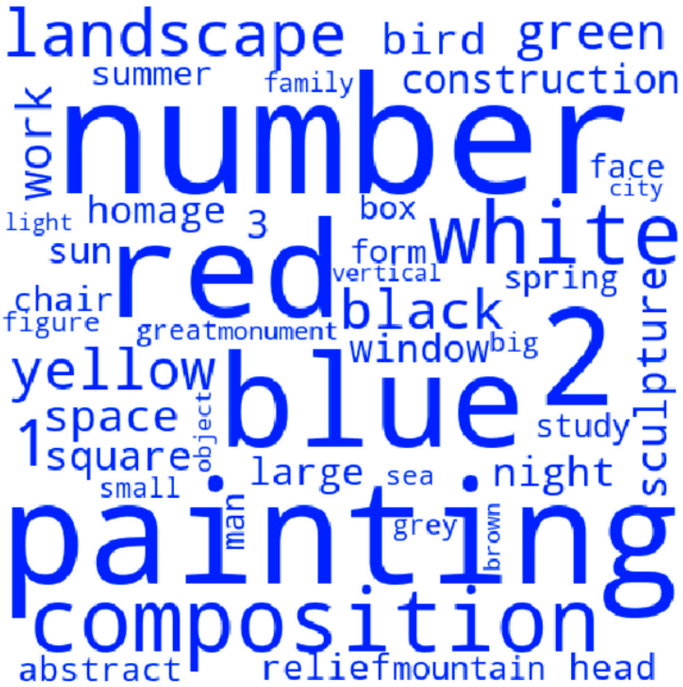

Figure 4. Word cloud for the 50 most important words in Topic 2.

with ‘composition’ in the top 10 words of the topic and ‘construction’ also featuring in the top 50. The words ‘abstract’ and ‘form’ are also important in this topic. I would characterise topic 2 as ‘abstract/formal’.

‘Composition’ is also an important word in topic 1, but its weight in topic 2 is much higher, which suggests its use was increasingly dissociated from figurative subject matter and so can be read as an expression of interest in composition itself. Of the generic labels in topic 1, only 'landscape' persist as an important word in topic 2, which indicates the continued engagement with that genre of painting as other kinds of figurative painting such as the still-life and the portrait went into a sharper decline.

'Number' is the word of highest weight in topic 2 and the numerals 1, 2 and 3 also feature in the top 50. Numerals can function to identify or to order. However, topic modelling does not allow these uses to be distinguished as it takes no account of the order of the words in the documents being modelled. The title 'Number 1' might signal a desire to minimise or eliminate the title's semantic role as something contributing to the work's meaning, restricting its function to the nominative one of identification. In contrast, the words 'Number 1' in the title 'Compositions in Red and Blue Number 1' might indicate a work to be seen as the first of a sequence and so have a significant semantic function. To understand how the word 'number' and numerals functioned we need to look at their presence in whole titles, not just as individual word tokens. I will come on to this later in this article.

The pattern of development of topics 1 and 2 in my model is consistent with the canonical narrative of modern art in the first half of the twentieth century. Whilst some artists in the early century continued to work within the figurative traditions of the nineteenth century, others in the main avant-garde movements of that time such as Fauvism, Cubism or Der Brücke challenged that tradition through giving different ways of seeing or representing figurative subject matter. Interests in regard to figuration continued to be a concern for many modern artists after the First World War, for instance with the revival of Realism in the 1920s through movements such as Neue Sachlichkeit in Germany, Traditionisme in France and Regionalism in the United States of America. The combination of natural forms to present an uncanny, alternate or subversive reality was associated with Magic Realism and was an important strand within Surrealism from the 1920s onwards. The emergence of abstract art as a significant trend within Modernism is often traced back to the 1910s and associated with artists and movements such as Wassily Kandinsky, Piet Mondrian, De Stijl and Suprematism. Abstraction grew in importance in the 1920s and 1930s, and from the 1940s to the 1970s it was a widely practiced form of Modernism, including various strands or schools of expressive and geometric abstraction such as Abstract Expressionism, Tachisme, Art Informel, Concrete Art, and movements such as Post-Painterly Abstraction, Minimalism and Op Art.

Topic 3 is the most prominent topic in my model for the titles of artworks created in recent decades. As its word cloud presented in Figure 5 shows, the strongest recent trends in titling include the emergence of the presentation of artworks as 'untitled', alongside the signalling of an engagement with seriality, and a continued use of numerals in titles. Indeed, 'untitled' dominates topic 3 much more than the words of highest weight do in topics 1 and 2. Presenting a

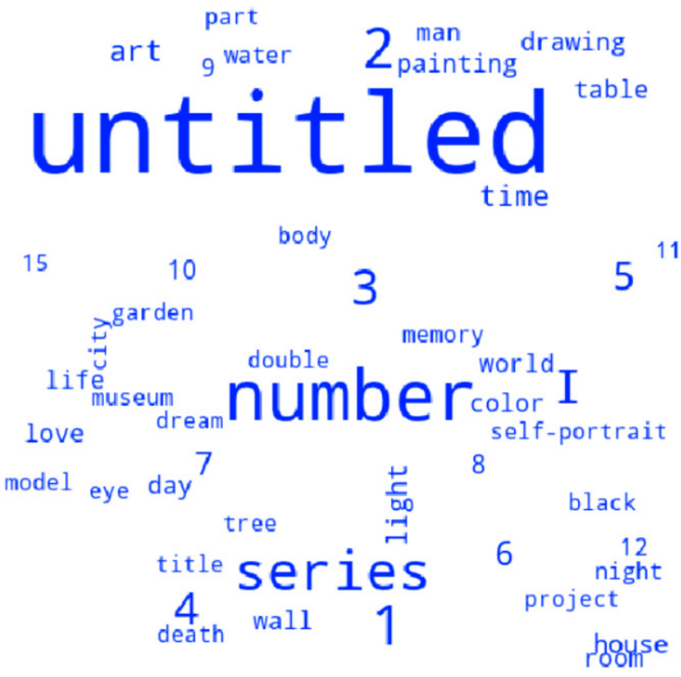

Figure 5. Word cloud for the 50 most important words in topic 3.

work as being untitled engages with the nominative aspect of the title, looking to efface that function. As with numerals, if used as part of a short title, 'untitled' is also a means of minimising its semantic role. However, that is not the case with longer titles where it appears in combination with other words. I will present more on my reading of the use of 'untitled' when I look at the way it features in whole titles.

The engagement with seriality is indicated through the presence in the top 50 words of topic 3 of 'series', 'part' and 'project', and the numerals from '1' thru '12'. Numerals were also important in topic 2, but are less prominent than in topic 3 and far fewer of them appear in the top 50. With topic 2 they would appear to indicate a stronger association with ordering or numbering individual works than with works to be seen as part of a larger whole. Indeed, 'series' and 'part' do not appear in the top 50 words for topic 2, whereas 'number' is more prominent in topic 2 than in topic 3.

The other important words of topic 3 can be read as signals of a range of possible artistic interests. Words such as 'art', 'museum', 'room', and 'wall' might indicate an engagement with the institutional context within which work of arts are displayed. In association with these interests, the word 'painting' that carries over as an important word from topic 2 to topic 3, might indicate a re-inflection of interest away from the work of art itself towards its institutional context. The occurrence of 'I', 'life', 'love', 'death', 'eye', and 'memory' suggest an engagement in some way with the somatic or the personal. 'Self-portrait' features as an important word in topic 3 and topic 1, but not topic 2, and when seen alongside the other important words in topics 1 and 3 suggests a shifting of interests in self-representation away from the figurative towards the personal. However the important words in topic 3 are read, and other readings are clearly possible, the interests expressed by them are more heterogeneous than those in the other two. I have characterised topic 1 as 'figurational' and topic 2 as 'abstract/formal'. Topic 3 does not have such a succinct summary.

The kinds of interests I have associated with topic 3 can also be seen in relation to recent developments in art. Since the 1960s art has been characterised by an artistic plurality. Working with a wide range of media artists have engaged with a diversity of styles and themes. Many artists have engaged with seriality as an element of their practice. Other important themes include those of race, personal identity and politics, and the making of art that is explicitly challenging and questioning of institutional structures and power relations. Artists have also looked to create a dialogue with styles or movements that came before them, and here we might see the continued significance of topics 1 and 2 in my model for the titles of artworks created in recent decades.

The move away from traditional art forms can also be seen in my dataset. If we look at those collections which classify work by art form, we find that painting dominated in the first half of the twentieth century, with paintings created in the 1900s or 1940s accounting for around 80% of the total, and for all other decades from the 1920s to the 1950s around 70% of all works. Since then the proportion of all works that are paintings has fallen consistently and substantially to 36% of works created in the 2010s.[5]

To complete my reading of the three most prominent topics in my model, we can note that there are no words that are important in all of them, which is an indication of how strongly the large-scale interests expressed through titles changed over the twentieth century.

We can also look at the art historical significance of the other topics in my model. Table 2 in the Appendix gives the top 20 words in each of these 17 topics, ordered by the decades in which they are prominent in the model. In all of the cases listed the topic had a weight of 4% or more in the model, and in 60% of cases the weight was 10% or higher. These topics are largely distinct, each sharing very few of their words of highest weight with the other nineteen topics. The most significant overlap is that the word 'composition' has by far the highest weight in the topic prominent only in the 1950s, indicating a peak in artistic engagement with the compositional in that decade over and above the persisting associations captured in topic 1 and topic 2.

[5] I have not given the comparable proportions for sculptures as in some of the collections in my dataset works are classified as both sculptures and installations, and similar works by the same artist are classified in some collections as sculptures and in others as installations.

One of the themes running through John Welchman's *Invisible Colors* is that of how artists engaged with the idea of composition and how that was expressed through their use of titles (Welchman, pp. 176 – 209 and 265 – 323). My account gives a measure of how significant that engagement was at different times through the twentieth century, and the peak in interest in the 1950s sheds further light on artistic motivations for what he terms the 'counter-compositional impulse' which emerged during that decade (Welchman 279).

Typically, the other topics can be read as reinforcing the interpretations I have given of topics 1 to 3. Often, they include words particularising the generic language featuring in those topics. From the 1900s to the 1940s, the topics prominent in those decades include important words giving specific locations or types of object as well as the names of individuals. Signals of artistic allegiance also feature, with 'synchromy', 'futurist, 'suprematist', and 'surrealist' words of high weight in prominent topics from that period. The practice of numbering works can be seen in the numerals of years featuring as important words in topics prominent during each of the decades from the 1950s to the 2010s. 'Modernism' is an important word in the topic prominent in the model for the 2000s, suggesting an artistic engagement with the modernist tradition from the twentieth century, and with the concept of modenism itself.

The other important words in these topics often relate to particular subjects of artists heavily represented in the collections included in my dataset. For instance, several collections have works from the project *Insertions into Ideological Circuits* by the artist Cildo Meireles, many of which were created in the 1970s. The river Gein was a popular subject with Piet Mondrian in his paintings from the 1900s as he moved towards abstraction.

The art historical value of the topic model is not limited to that provided by the most prominent topics in each decade and the most important words in those topics. The overall distributions of topics in the model for the titles of works created in each of the decades from the 1900s to the 2010s can give an indication of the diversity of titling practice in each period and how that has changed over time. The more diverse the titles used by artists the more topics there will be of significant weight in the model for that period to capture that variety. Conversely, the more titling is centred upon a limited set of words, the more the topic model will be concentrated in a few topics of high weight. A succinct measure of diversity or complexity in this sense of the degree of concentration of the weights in a topic distribution is its entropy, which has been used by a number of scholars to look at questions of cultural or linguistic complexity (Juola; Lincoln; Chen et al.; Rasterhoff et al.; Long 135–40). The entropy of a distribution can be thought of as the number of bits required to characterise it. The higher the entropy the more spread the distribution and the more bits required to characterise it.

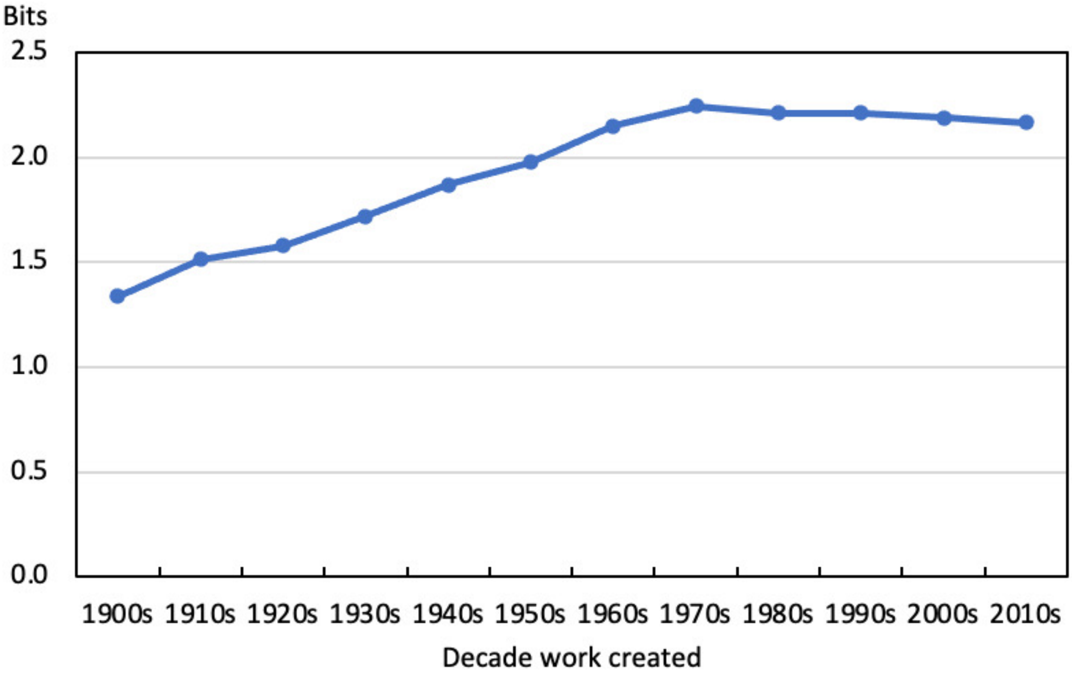

Figure 6. Entropy of topic model distributions for the titles of works created in each decade, 1900s to 2010s.

The entropy of the topic distribution in my model for each decade is shown in Figure 6. From an art historical perspective, the absolute levels of the entropy are less important than the change in its value over time. As can be seen, on this measure titling became persistently and significantly more diverse from the 1920s to the 1960s. As artists moved away from signalling a predominant interest in, or engagement with, figuration, in aggregate the words in the titles they used in each decade were sending out a wider range of signals. From the 1960s onwards the diversity of the interests signalled through the titles used for works created in each decade has changed little, with the entropy of the topic distribution roughly constant.

A more widely-used and more readily understood measure of linguistic variety in a given collection of texts is the type-token ratio, and I was interested to compare these two measures (Jockers 54–58; Long 139–46). Type-token ratios can only be compared meaningfully between documents of the same word length, or number of tokens. The documents with the higher type-token ratios contain more word types, and so draw on a richer vocabulary. I took as my base the decade with the fewest words used in titles, the 1900s with 13,986 words, and for each subsequent decade took the first 13,986 words in a randomized list of all the titles from that period. The results of my analysis are presented in Figure 6. As can be seen, it mirrors the trends in the diversity of titling practice as measured by the entropy of the topic model distribution for each decade. With the titles of works created in the 1900s, each word type was used 3.3 times on average. Using a comparable number of word tokens, each word type was used 2.5 times on average with the titles of works created in all decades from the 1960s to the 2010s.

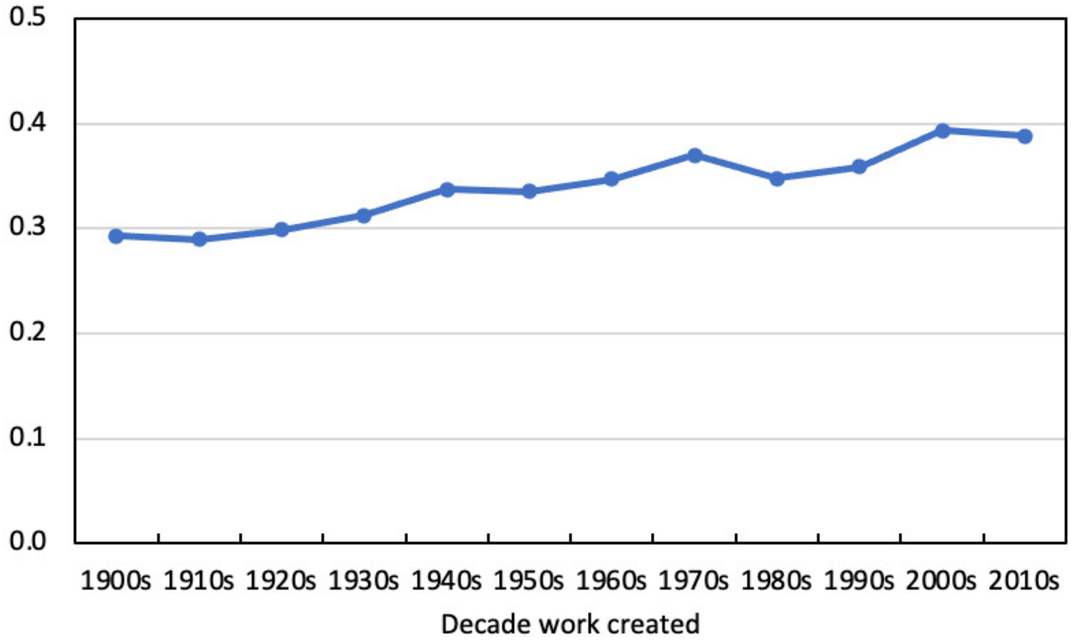

Figure 7. Type-token ratio of the titles of works created in each decade, 1900s to 2010s.

3.2. Parts of Speech - signals of epistemic perspectives

My topic modelling has allowed me to set out a reading of the large-scale trends in the use of the 'content' words of titles during the twentieth century. The use of other components of titles may also be of art historical significance. Parts-of-speech tagging is one way of investigating that question, as well as other issues more directly related to my interpretation of the topic model. It measures the use of all classes of 'syntactic token' including different types of word such as nouns, adjectives or verbs, and other types of token such as numerals, symbols, and punctuation marks.

One way of reading the parts of speech distribution for the titles of works created in each decade is as an indication of epistemic perspective, or the kind of knowledge art can or should produce. In the first half of the twentieth century the predominant parts of speech were nouns, prepositions, determiners, and adjectives. The sorts of titles they were used in were typically declarative or descriptive such as *Interior with Violin*, *The Seine near Paris* or *Composition in Red* and so expressive of a static, objective, or third-person epistemic perspective. Since the 1950s that stance has been changing to be one that is more dynamic, subjective, questioning, and opinionising. There has been a modest but steady move away from declarative and descriptive titles to the subjective, interrogative, exclamatory or imperative title, and which seems to suggest a re-orientation of the relationship between the artist, artwork and viewer. This has happened through an increase in the use of personal pronouns, of end-sentence punctuation such as the exclamation mark, the question mark, and the ellipsis, of modals such as 'can' or 'should' and 'wh-words' such as 'who', 'what', 'where' and 'how'. Examples of this kind of title

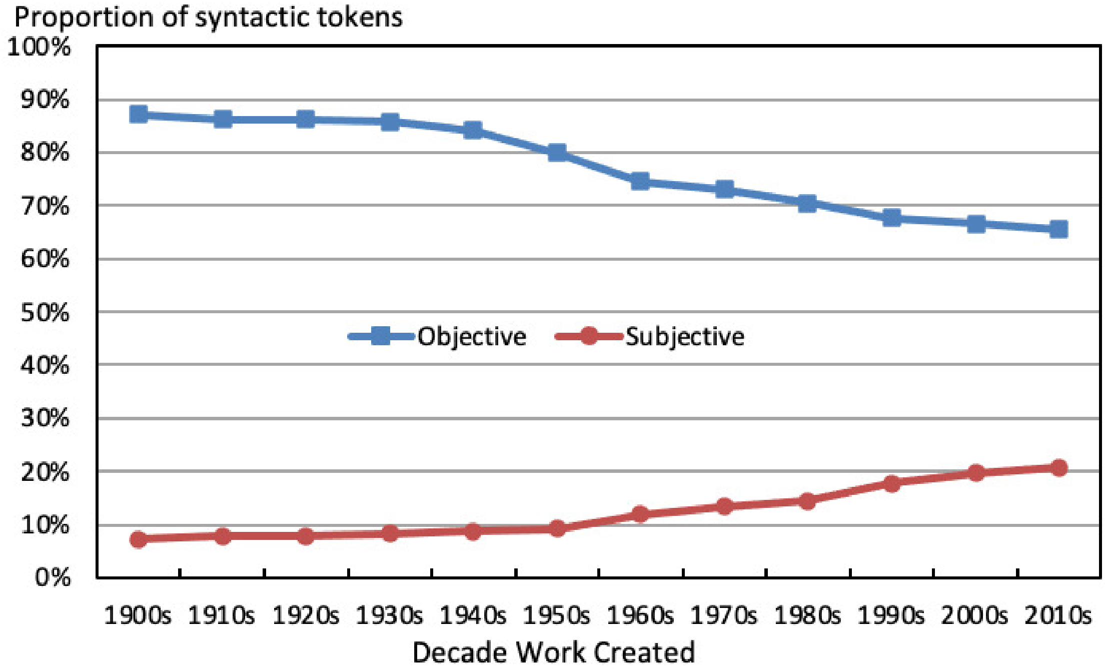

Figure 8. 'Objective' and 'subjective' parts of speech, proportion of all syntactic tokens in the titles of works created in each decade, 1900s to 2010s.

include *Listen and be Quiet!* and *May I Help You?*. The growing use of text in brackets as an aside from the artist is another way this has happened. The use of verbs and adverbs in titles has also grown. Through being used in titles such as *Succeeding the Past,* I would read this trend as showing an increased interest in process, agency, and change. Often verbs and other parts of speech have gone together in titles such as *How to Make a Refugee* and *We Could be Looking for the Same Things.*

The overall trends in the use of these different parts of speech is shown in Figure 8. For ease of presentation, I have labelled nouns, prepositions, determiners, and adjectives taken together, but excluding all instances of 'untitled', as 'objective'. I have labelled brackets, personal pronouns, end-sentence punctuation, modals, verbs, adverbs, and all instances of 'untitled' followed by additional text as 'subjective'. The remaining syntactic tokens are predominantly 'untitled' when used on its own and numerals, which I have excluded as I would not read them as strongly suggesting one epistemic stance or another. As can be seen 'objective' parts of speech have fallen from over 80% of all syntactic tokens in the titles of works created in each decade from the 1900s to the 1940s, to 65% in the 2010s. 'Subjective' parts of speech have risen from around 10% of all syntactic tokens in the titles of works created in the 1940s to 21% in the 2010s.

In my analysis I used the tagger developed by the Stanford University Natural Language Processing Group to determine the proportion of all syntactic tokens with each part of speech for the titles of works created in each decade.[6] The Stanford tagger has a reported accuracy of 97% in assigning parts of speech for English. A lower level of performance with titles is to be expected as titles are often short and so provide less context to the tagger than other texts. However, in my reading I grouped together all kinds of adjectives and nouns as 'objective' parts of speech, and so this measure is unaffected by mis-taggings such as singular nouns as plurals, nouns as proper nouns, and adjectives as nouns. Similarly, the 'subjective' parts of speech include verbs of all kinds, and so the measure is unaffected by mis-taggings such as past tense verbs as past participles. Some mis-taggings such as nouns as verbs and adverbs as adjectives will have affected the subjective and objective measures. To get an indication of how well the tagger was identifying the subjective and objective parts of speech I compared the results of using the tagger with my subjective assignments for 500 randomly selected titles. Overall it was 98% correct.

3.3. Word Counting – engagement with the functions of the title

John Welchman has examined the various ways artists used what he terms 'non-referential' titling, in which works were presented as 'untitled' or with numerical titles (Welchman 8). This is also a theme I read from the results of my topic modelling, where I suggested that the use of numerals and of 'untitled' in short titles were ways of minimising the role of the title in contributing to the meanings given to the work it names. I also commented my interpretation was incomplete as the context of use was not being taken into consideration. If we now look at how the word length of titles has changed and at the words used in them, then I would read these uses of titles as part of a much broader trend of 'minimal' titling extending through the twentieth century and into the twenty-first. The semantic role of titles is also lessened through reducing the number of words in them, for instance to simple descriptions of the subject matter or to generic labels. Titles such as 'landscape' and 'mother and child' were in common use in the early decades of the twentieth century, with short titles such as 'painting' and 'abstract composition' frequent in the 1940s and 1950s.

A consolidated measure of these trends is that of the prevalence of short titles of three words or fewer in length, which I have also split into short titles including 'untitled', short titles including a numeral and not including 'untitled', and other type of short title.[7] These trends are shown in Figure 9 and, as can be seen, minimal titling has always been a major component of titling practice. The proportion of all titles of length three words or fewer rose

[6] The Stanford University Natural Language Processing Group Part of Speech Tagger can be found at https://nlp.stanford.edu/software/tagger.shtml.

[7] The overlap between short titles including 'untitled' and those including a numeral is small. Of all short titles including 'untitled' or a numeral, only at most 2% in any decade include both.

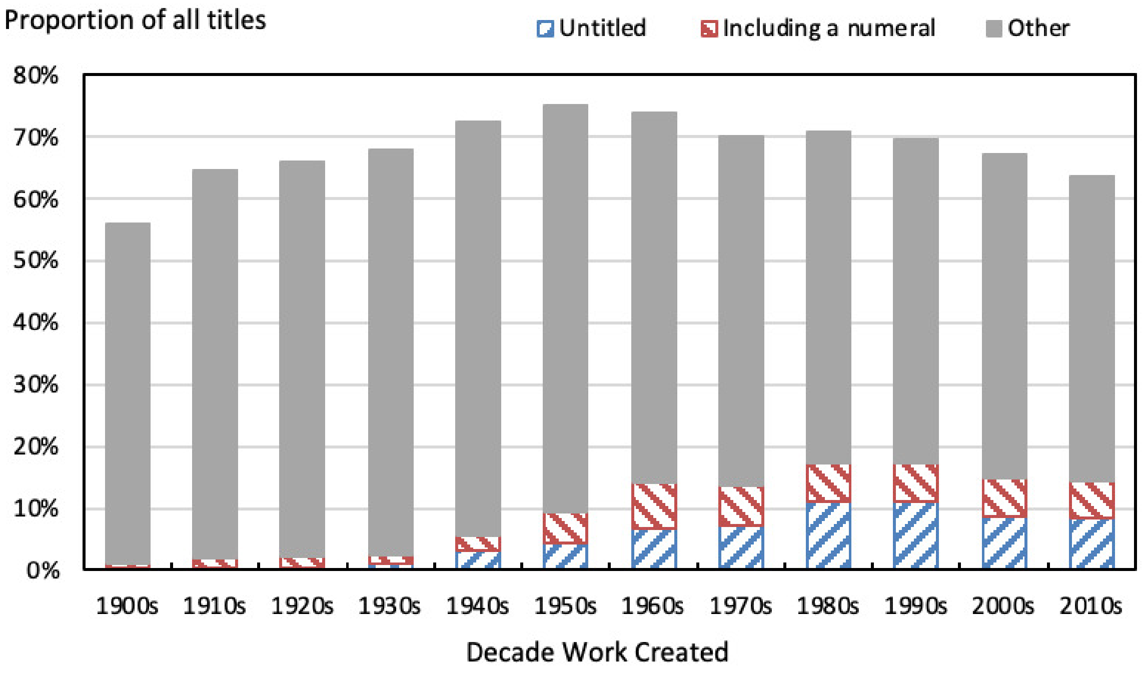

Figure 9. Minimal titling – proportion of titles of word length three words or fewer, works created in each decade, 1900s to 2010s.

persistently and markedly from 56% with works created in the 1900s to 75% in the 1950s before falling back to 64% in the 2010s. Within that overall trend, short titles including 'untitled' peaked at 12% of all titles of works created in the 1980s, and has declined since then to 9% of all titles in the 2010s. Short titles including a numeral but not 'untitled' peaked in the 1960s at 7% of all titles of works created in that decade. Other types of short title rose to 67% of all titles in the 1940s, before declining to represent 49% of the titles of works created in the 2010s.

In my reading of topic 3 I suggested that presenting a work as 'untitled' engages with the nominative function of the title, a dimension to that practice not considered by Welchman. The trend in this use of titles is shown in Figure 10. As can be seen, the presentation of works as untitled first emerged in the 1930s, and was most prominent in the 1990s, with 13% of the titles of works created in that decade including 'untitled'. Comparing Figures 8 and 9 shows that presenting works as untitled was predominantly associated with their use in short titles, and so simultaneously looking to minimise the semantic role of the title. Of the titles of works created in all decades from the 1960s to the 2010s only 1% to 2% included 'untitled' and were of word length four or more, and so had a significant semantic role, for instance with the titles often used by Dan Flavin in which text in brackets was utilised to give a personal dedication from the artist as with *untitled (to Don Judd, colorist)*.

Whilst minimal titling has declined in recent decades, the use of long titles and the syntactic complexity of titles have both been growing. After having declined in the first half of the twentieth century, use of long titles of ten words

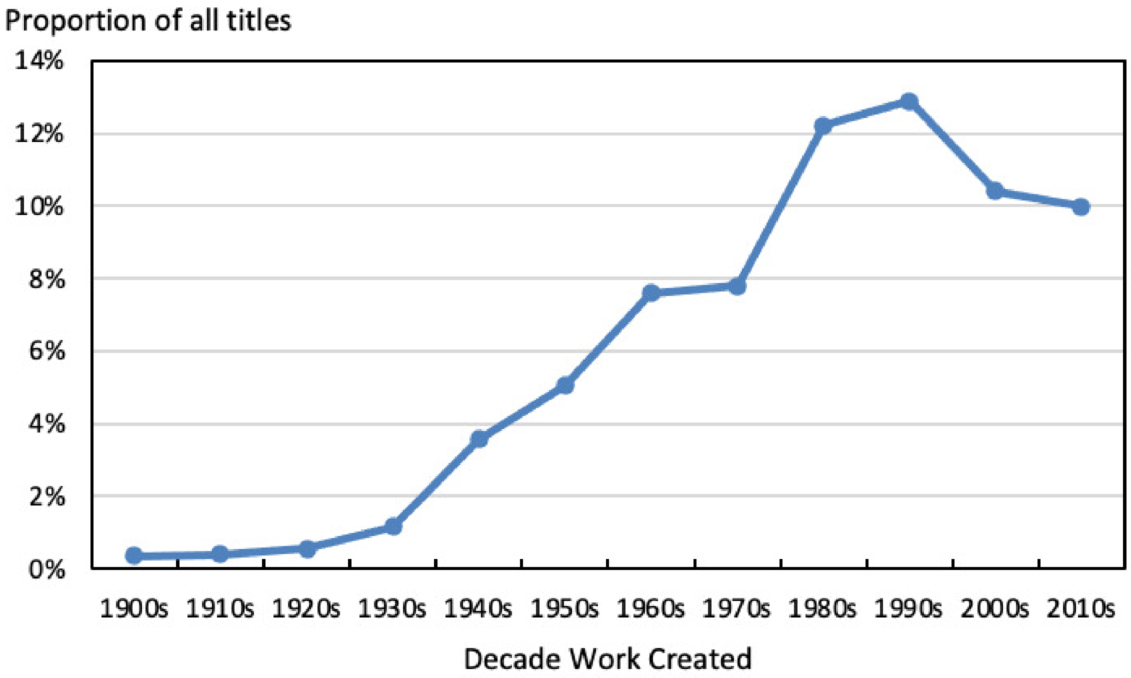

Figure 10. Proportion of titles including 'untitled', works created in each decade, 1900s to 2010s.

or more then picked up, growing from 1.0% of all titles of works created in the 1950s to 5.0% in the 2010s, as shown in Figure 11. The parts of speech model shows that in recent decades there has been more variety in the parts of speech used in titles, including increased use of all kinds of punctuation mark. A succinct measure of the syntactic complexity of titles in this sense is the entropy of the parts of speech distribution, and, as can be seen from Figure 11, this has been rising steadily from the 1940s. Long or syntactically complex titles can be used to do many things. One is that they would have stood out from the often-minimal titles around them and so functioned 'seductively', attracting the attention of the reader.

4. Summary and Discussion

The metadata provided in exhibition catalogues is an under-used resource in digital art history. In this article I have brought a range of statistical techniques to bear on those resources, using them individually and in combination to investigate how text-mining titles as given in the metadata with online art museum collections can be used to look at the history of modern and contemporary art. These techniques include topic modelling, parts-of-speech tagging and word counting. The dataset I compiled was metadata including the work's title and the year of creation for over 170,000 entries downloaded from the online collections of 133 art museums in 30 countries.

The perspective provided by my application of statistical text-mining techniques to this metadata is very different in scale and in content to that provided in the art historical literature. It cannot replace the kind of detailed account of titles given in that scholarship. Rather, text-mining, as I have used it, provides a distant viewpoint delivering rich numerical and visual descriptions.

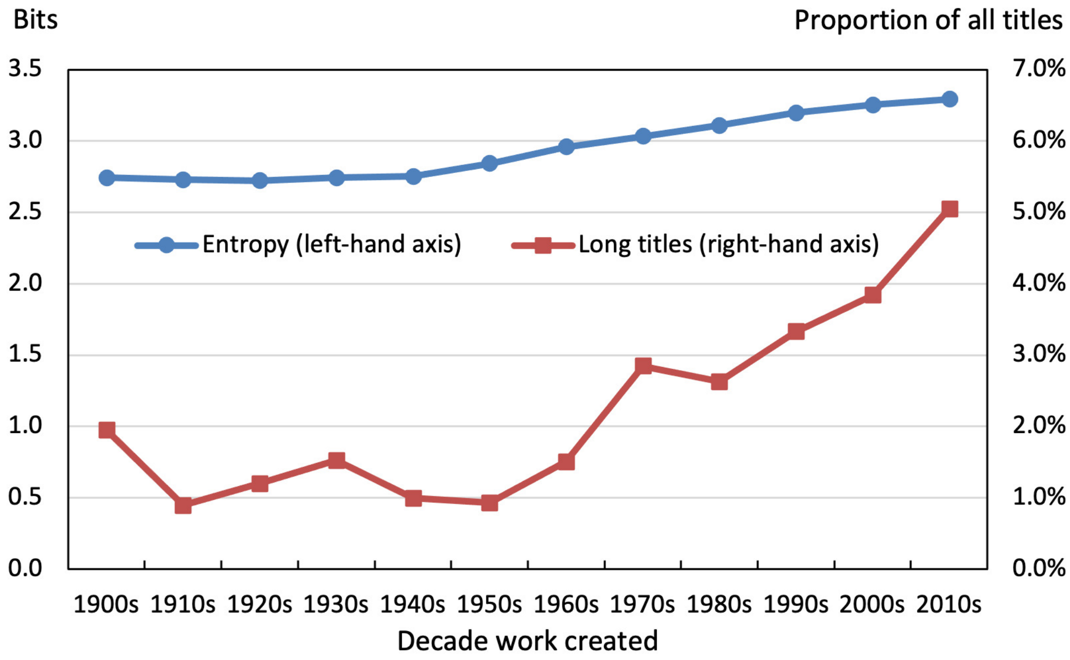

Figure 11. Long title of length 10 words or more, entropy of the parts of speech distributions, titles of works created in each decade, 1900s to 2010s.

These descriptions, whether numerical or graphical, give new ways of seeing the ways titles have been used and some of the developments within modern and contemporary art. They also support a reading in more recognisably art historical terms, which is consistent with that given in earlier accounts of titling and with the canonical history of modern and contemporary art. The long-run trends which I identify and give art historical interpretation to suggest some new ways of thinking about that history, cutting across the particularities of earlier accounts.

In my reading of the results of the topic model we can see an inter-woven cyclic pattern of change in which the large-scale use of titles to express certain interests increases in importance, becomes dominant and is then displaced. Titular signals in relation to figuration are displaced primarily by those relating to abstraction and the formal dimension to painting, and to numbering, which in turn give way to the more heterogeneous set of interests expressed through the words in topic 3. What cut across the expression of the two types of interest dominant in the first half of the twentieth century was an epistemic stance in which titles expressed an objective and stable perspective. In the second half of the twentieth century and into the twenty-first, the kinds of interest I read from topic 3 were associated with a more subjective and questioning epistemic position. Minimal titling, in which artists looked to minimise the semantic role of the title, was a major trend all through the twentieth century and into the twenty first. In recent decades the strongest trend in titling has been that of presenting works as untitled, which is a way of engaging with the nominative function of the title.

Statistical modelling of titles in aggregate also allows me to answer other questions that cannot be addressed through conventional art historical methods. In particular, through an appeal to the information theoretic concept of entropy and to type-token ratios I have interpreted my model results in terms of the levels of diversity and complexity in the use of titles and in their syntax.

The readings set out in this article represent a first step in the digital art historical analysis of titles and the metadata contained in online museum collections. The interpretive framework and the dataset are not limited to looking at titling in modern and contemporary art as a whole. To end this article I will suggest some of the ways in which they could be utilised to address a number of art historical questions, and in ways which both extend the digital methods of this article and use them in combination with conventional art historical methods.

The value of digital art history in transforming established narratives and in opening up new areas of enquiry often comes when it is used at scale or to investigate the complex inter-relationships between features represented in a dataset. Digital methods have been one way art historians have looked at questions around networks and geographies, often challenging the dominant histories of modern and contemporary art (Joyeux-Prunel; Nijboer). I am in the process of assigning a nationality to each of the artists named in my dataset. This will allow me to investigate the circulation of the ideas, interests and perspectives I have read from the results of my modelling between artists of different nationalities and to address questions such as whether those interest emerged earlier or were more prominent with artists of one nationality than with others. Assigning a gender to each artist would allow a similar analysis to be carried out for male and female artists.

The question of the canon is one which has engaged art historians since the 1980s. Scholars have looked to identify canonical formations, and to historicize and critique the production of canons and canonical value (Brszyski; Iskin). The art historian Robert Jensen has used statistical methods to address this question, counting the canonical through the levels of appearance of artists in 38 art survey textbooks published in America, and arguing that canonical status derives from innovation (Jensen). As he puts it 'quantitative analysis is uniquely suited to understanding the collective perceptions of a discipline...' (Jensen 28). My dataset could allow a comprehensive quantitative analysis of artistic canons in modern and contemporary art, as produced through the prominence of an artist in museum collections. It could look both in aggregate at my dataset and also at cross-country comparisons. The analytical framework I have developed provides one way in which canonical and non-canonical artists could be compared. I have not collected metadata on accessioning dates, but adding that to my dataset could also shed light on canon formation and canon change. It may be of museological interest to compare the collections of modern and contemporary art included in my dataset, and to look at the representation of artistic careers.

It would be of art historical value to have a detailed understanding of the emergence and adoption of the presentation of works as untitled and the conventional use of 'untitled' to describe such a work. My dataset provides a starting point, identifying artists who began presenting works that way. Supplementing this with traditional art historical methods including examination of contemporary exhibition catalogues, artist writing, journals and critical reviews, could allow that history to be written.

Looking further out, we can speculate on whether the cyclic pattern of change that can be seen in my topic model continues. If the topic modelling were repeated in, say, ten years, would topic 3 remain dominant or would a topic 4 have emerged, expressing new interests and concerns or subsuming some of the interests previously expressed through topic 3?

My work can also be considered from the methodological perspective. The interpretations I have developed show the value of considering a topic model as a whole rather than as a thematic analysis in which topics are treated independently and univocally. Parts-of-speech analysis is typically used in the digital humanities as a first step to further analysis. For instance, to identify the adjectives used in a text as part of an analysis of the sentiments being expressed or to isolate the proper nouns, which can then be extracted as part of the text preparation in topic modelling. It may also be used for classificatory purposes with the pattern of use of different parts of speech as a marker utilised to identify the author of a text. However, in the digital humanities it has been used only rarely to provide representations that are of further interpretive significance (Allison et al.). In my own reading I have looked to give art historical interpretations to all parts of the topic model and of the parts of speech model.

Researchers in the digital humanities often look at one quantitative technique, or at several sequentially, where one is seen as improving upon the previous in answering the research question. My work shows that there can be benefit in using statistical techniques in combination to develop a consolidated reading. The various perspectives provided by topic modelling, parts-of-speech analysis and count-based statistics complement each other and, in the cases of the use of the numerals and 'untitled', are required to develop a more complete interpretation.

Data repository: https://doi.org/10.7910/DVN/MDGEYO

Submitted: October 10, 2022 EST, Accepted: November 22, 2022 EST

REFERENCES

Allison, Sarah, et al. "Style at the Scale of the Sentence." *Stanford Literary Lab Pamphlets*, vol. 5, 2013, https://litlab.stanford.edu/pamphlets/.

Arnason, Hjorvadur H., and Elizabeth C. Mansfield. *History of Modern Art*. Pearson, 2013.

Bann, Stephen. "The Mythical Conception Is the Name: Titles and Names in Modern and Post-Modern Painting." *Word & Image*, vol. 1, no. 2, Apr. 1985, pp. 176–89, https://doi.org/10.1080/02666286.1985.10435674.

Besa Camprubi, Jose. *Nouveaux actes sémiotiques No 82: Les fonctions du titre*. Presses Universitaires de Limoges, 2002.

Blei, David M., et al. "Latent Dirichlet Allocation." *Journal of Machine Learning Research*, vol. 3, no. 1, 2003, pp. 993–1022, https://doi.org/10.5555/944919.944937.

Bois, Yves-Alain, et al. *Art Since 1900: Modernism, AntiModernism, Postmodernism*. Thames & Hudson, 2004.

Boyd-Graber, Jordan, et al. "Applications of Topic Models." *Foundations and Trends® in Information Retrieval*, vol. 11, no. 2–3, 2017, pp. 143–296, https://doi.org/10.1561/1500000030.

Brszyski, Anna, editor. *Partisan Canons*. Duke University Press, 2007.

Chen, Ruina, et al. "Entropy in Different Text Types." *Digital Scholarship in the Humanities*, vol. 32, no. 3, 2017, pp. 528–42, https://doi.org/10.1093/llc/fqw008.

Drucker, Johanna. "Is There a 'Digital' Art History?" *Visual Resources*, vol. 29, no. 1–2, June 2013, pp. 5–13, https://doi.org/10.1080/01973762.2013.761106.

Fletcher, Pamela, and Anne Heimreich. "Local/Global: Mapping Nineteenth-Century London's Art Market." *Nineteenth Century Art Worldwide*, vol. 11, no. 3, 2012, https://www.19thc-artworldwide.org/autumn12/fletcher-helmreich-mapping-the-london-art-market.

Garcia-Zorita, Carlos, and Ana R. Pacios. "Topic Modelling Characterization of Mudejar Art Based on Document Titles." *Digital Scholarship in the Humanities*, vol. 33, no. 3, Oct. 2017, pp. 529–39, https://doi.org/10.1093/llc/fqx055.

Genette, Gérard, and Bernard Crampé. "Structure and Functions of the Title in Literature." *Critical Inquiry*, vol. 14, no. 4, July 1988, pp. 692–720, https://doi.org/10.1086/448462.

Goldstone, Andrew, and Ted Underwood. "The Quiet Transformation of Literary Studies: What Can Topic Models of PMLA Teach Us About the History of Literary Scholarship?" *New Literary History*, vol. 45, no. 3, 2014, pp. 359–84, https://doi.org/10.1353/nlh.2014.0025.

Gombrich, Ernst. "Image and Word in Twentieth-Century Art." *Word & Image*, vol. 1, no. 3, July 1985, pp. 213–41, https://doi.org/10.1080/02666286.1985.10435861.

Greenwald, Diana S. *Painting by Numbers: Data-Driven Histories of Nineteenth-Century Art*. Princeton, Princeton University Press, 2021.

Hoek, Leo H. *Titres, toiles et critique d'art, Déterminants institutionnels du discourse sur l'art aux dix-neuvième siècle en France*. Brill, 2001.

Hopkins, David. *After Modern Art: 1945 – 2017*. Oxford, Oxford University Press, 2018.

Iskin, Ruth, editor. *Re-Envisioning the Contemporary Art Canon: Perspectives in a Global World*. Routledge, 2016.

Jensen, Robert. "Measuring Canons: Reflection on Innovation and the Nineteenth-Century Canon of European Art"." *Partisan Canons*, edited by Anna Brszyski, Durham and London, Duke University Press, 2007, pp. 27–54.

Jockers, Matthew L. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, 2013, https://doi.org/10.5406/illinois/9780252037528.001.0001.

Joyeux-Prunel, Béatrice. “Provincializing Paris, the Center-Periphery Narrative of Modern Art in the Light of Quantitative and Transnational Approaches.” *Artl@s Bulletin*, vol. 4, no. 1, 2015, https://docs.lib.purdue.edu/artlas/vol4/iss1/4/.

Juola, P. “Using the Google N-Gram Corpus to Measure Cultural Complexity.” *Literary and Linguistic Computing*, vol. 28, no. 4, June 2013, pp. 668–75, https://doi.org/10.1093/llc/fqt017.

Lincoln, Matthew. “Social Network Centralization on Print Production in the Low Countries, 1550 – 1750.” *International Journal for Digital Art History*, no. 2, 2016, pp. 1550–750, https://dahj.org/article/social-network-centralization.

Long, Hoyt. *The Value in Numbers: Reading Japanese Literature in a Global Information Age*. Columbia University Press, 2021, https://doi.org/10.7312/long19350.

Mainardi, Patricia. *The End of the Salon*. Cambridge University Press, 1993.

Moretti, Franco. “Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850).” *Critical Inquiry*, vol. 36, no. 1, Jan. 2009, pp. 134–58, https://doi.org/10.1086/606125.

Nijboer, Harm et. “Unthinking Rubens and Rembrandt: Counterfactual Analysis and Digital Art History.” *Digital Humanities*, 2019.

Piper, Andrew. *Enumerations: Data and Literary Study*. University of Chicago Press, 2018, https://doi.org/10.7208/chicago/9780226568898.001.0001.

Quach McCabe, Sophia. “Intermediaries and the Market: Hans Rottenhammer’s Use of Networks in the Copper Painting Market.” *Arts*, vol. 8, no. 2, June 2019, pp. 75–96, https://doi.org/10.3390/arts8020075.

Rasterhoff, Claartje, et al. “Measuring Innovation in the Art and Book Market during the Dutch Golden Age.” *Digital Humanities Benelux*, https://web.archive.org/web/20210119053958/https://2018.dhbenelux.org/programme/detailed-programme/.

Smeets, J., et al. “SMTP: Stedelijk Museum Text Mining Project.” *Digital Humanities*, 2016, https://hdl.handle.net/11245.1/69bd334a-49bf-40bb-927f-ac537024fc61.

Smith, Terry. *Contemporary Art: World Currents*. Laurence King, 2011.

Teukolsky, Rachel. *The Literate Eye, Victorian Art Writing and Modernist Aesthetics*. Oxford, Oxford University Press, 2009.

Underwood, Ted. *Digital Horizons, Digital Evidence and Literary Change*. Chicago, Chicago University Press, 2019.

Welchman, John C. *Invisible Colors: A Visual History of Titles*. New Haven and London, Yale University Press, 1997.

Yeazell, Ruth Bernard. *Picture Titles: How and Why Western Paintings Acquired Their Names*. Princeton University Press, 2015, https://doi.org/10.2307/j.ctvc779dd.

Appendix. Data Tables

Table 1. Countries with museum collections included in the dataset, and associated number of entries.

Country	Entries	Country	Entries
Argentina	1,025	Mexico	1,471
Australia	7,295	The Netherlands	6,689
Austria	2,756	New Zealand	3,158
Belgium	1,863	Peru	337
Brazil	2,371	Poland	1,555
Chile	474	Portugal	2,124
Cuba	404	Romania	1,362
Czech Republic	972	Slovakia	6,628
Finland	6,552	Slovenia	737
France	33,653	South Korea	6,881
Germany	9,573	Spain	9,185
Greece	746	Sweden	5,528
Ireland	1,321	Switzerland	5,179
Italy	2,196	United Kingdom	6,210
Japan	6,510	United States of America	35,102

Table 2. Top 20 words in the other seventeen topics in the topic model.

Top 20 words	Decades prominent
estaque rocky synchromy improvisation monhegan fernande gein sunlight kamenicky wertheimer futurist birch copy fence ogunquit murnau elbe saint-cloud moulin vanderbilt	1900s and 1910s
guitar clothespin cagnes chess lady drapery ploughing alsace teenage elephant suprematism boufarik meyer compotier tropical toulon instrument trumpeteer trapeze tobacco	1920s
santa presentation butterfly ana hall drapery shell harmony apollo independence disk 1937 1939 optical final surrealist rotorelief tatras guernica tuileries	1930s
project sculpture moving-static-moving blossom negro clown tivoli 1948 1949 gable sin 1945 sedona bacchante partisan store deadly story flower mary	1940s
composition 56 concept miniature 2nd 59 55 1950 cycladic spatial cathedral 1959 1956 prisoner diego linear 1957 1955 political upright	1950s
solomon electronic delta report position knot achrome cord finger melville force enclosure bleu nero carriage demonstration chung synthesizer flush announcement	1960s
piece relief poem american 1964 structure reclining 63 1967 1962 1969 paper 1965 concept prototype palgwae 1963 page 1968 bouncing	1960s
acrylic 71 collaboration blooming nirvana tube split signed attention seorak boomerang rite tape conical diane 79 lichtenstein trick poison verse	1970s
work canvas piece selected reel surface tape poem horizon 1970 original open circuit fischer artwork shut ideological portfolio insertion send	1970s
86 god 300 200 se horizon two-finger position directly photocopy exercise right-handedness 1988 site vertical apocalypse surface extensor 1981 1983	1980s
paradise angel pain 2000 passage suite wrong story case video mike dancing young survival ego jerusalem ghost name echo aids	1980s and 1990s
praescriptura balcony directional 1994 1996 1997 nascens cantus 1992 streets causa piraq web 95 a-b 1993 1991 1990 terrestris paradisus	1990s
ideal citizen lose cliff evidence delft modernism multimedia motel 180 strawberry poland caravaggio pe unwritten laid dolmen guided doppelganger temple	2000s
france 2001 2000 2006 scene 2004 2003 2005 2008 2009 2002 earthling make width 1998 2007 penelope map collage fedex	2000s
05 01 02 07 remix fly true mix black octopus conversation fake projected 06 stolen crying medusa finland auto exposed	2000s
weak acid cabaret founder mouth ultra blocked duplex bare 73 connection rude ruin model arm 1985 knot cuddly error tail	2010s
2010 walt store chapter ws copy 2012 paul 2011 panel ghost 2013 2017 soundtrack ramada joan 2014 self-construction 2016 berlin	2010s