

A Quantitative Study of Fictional Things

Andrew Piper, Sunyam Bagga

McGill University, 680 Sherbrooke St., Montreal, QC H3A 2M7, Canada

Abstract

In this paper, we apply machine learning based predictive models on two large data sets of historical and contemporary fiction to better understand the role that things play in fictional writing. A large body of scholarship known as “thing theory” has attempted to understand the function of fictional things within literature mostly by focusing on small case studies. We provide the first-ever estimates of the distribution of different types of things in English-language fiction over the past two centuries along with experiments to model their semantic identity. Our findings suggest that the most common fictional things are structural in nature, functioning akin to narrative props. We conclude by showing how these findings pose problems for inherited theories of fictional things and propose an alternative theoretical framework, embodied cognition, as a way of understanding the predominance of structural things.

Keywords

thing theory, embodied cognition, fiction, narratology, machine learning, natural language processing

1. Introduction

Over the past two decades, a large body of research has emerged in the field of literary studies focusing on the question of “things.” “Thing theory,” as this area has come to be known [8], has its origins in different research traditions, including the rise of material cultural studies [1], new historicism [18], and media theory [21, 33, 27]. At the heart of this work has been an attempt to move attention away from the symbolic aspects of often natural objects in literature – an interpretive tradition grounded in European Romanticism – towards the materiality, physicality, and the madeness of fictional things. In doing so, thing theory reorients critical attention around different kinds of things and different kinds of roles that things may play in the history of imaginative literature.

In this paper, we apply machine learning based predictive models on two large data sets of historical and contemporary fiction to better understand the role that things play in fictional writing. We define things for our purposes as any non-human object and thus distinguish them from people, places, or human body parts like faces and eyes. Our aim is to better understand the function that imaginary objects play in the practice of creative storytelling and how this has potentially changed over time. Despite a wealth of recent case studies that focus on particular types of things in individual books [17, 31, 20, 22, 32], only one work to date has used computational methods to study the broader distribution of fictional objects [28]. If we want to understand what Brown has called “a genuine sense of the things that comprise the stage

CHR 2022: Computational Humanities Research Conference, December 12 – 14, 2022, Antwerp, Belgium

✉ andrew.piper@mcgill.ca (A. Piper); sunyam.bagga@mail.mcgill.ca (S. Bagga)

ORCID 0000-0001-9663-5999 (A. Piper)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

on which human action, including the action of thought, unfolds” (5) [11], then it is imperative that we develop methods that can more sufficiently account for the broader population of things in creative writing.

Accordingly, we have two primary aims in this paper. First, we provide the first-ever estimates of the distribution of things in fictional writing over the past two centuries. For the first time, we can observe the kinds of things that populate fictional writing at least with respect to a single major language. Second, we conduct a series of experiments to better describe the semantic identities of things as a way of inferring their narrative function. Doing so can help us engage with longstanding debates about the nature and role of fictional objects within the field of thing theory.

Based on the data we provide here, we find that the most predominant type of fictional things since the nineteenth century are “infrastructural” in nature. By infrastructural we mean three key qualities: the first, ontological in nature, is that the most common types of things in fiction are more likely to be *structural* rather than, say, instrumental or vehicular or organic. The most prevalent things in fiction hold other things up. They provide narrative support. Thus for us infrastructure does not more narrowly refer to the contemporary usage of “airports,” “roads,” etc., but rather to human-made structures more generally, such as rooms, doors, tables, walls, and homes that help support other objects.

The second way we mean infrastructural is that these things are also distinctively semantically inert: as we will show through different experiments, structural things are less descriptively rich than other kinds of entities, less affectively charged, and more associated with physical rather than cognitive or emotional behavior. Such things are more semantically inert than other kinds of entities. They appear to function more as vehicles of embodiment than intellection. Third and finally, we mean infrastructural in the sense that such structural things appear regularly throughout narrative time. They are consistently woven throughout the whole narrative rather than functioning as singularly important entities. Their continued prevalence is a key part of their meaning.

Overall, then, the most predominant things in fiction appear to function as “props” in multiple senses of the word: they physically hold other things up; are descriptively shallow; and predominate throughout narrative space and time. As we discuss at the close of our paper, we think these insights have important implications in modifying the inherited views of thing theory and the role that fictional things play in literature.

2. Data

The primary datasets we use for this paper are, first, the Hathi1M dataset which consists of 1,671,370 randomly drawn pages from over 300,000 volumes of English prose in the Hathi Trust digital library [3]. These volumes span the years 1800-2000 and consist of an approximate balance of pages labeled as fiction and non-fiction. The second, the CONLIT dataset [25], is a collection of 1,934 works of English-language fiction drawn from eight different genres published between 2001 and 2021 largely North American in origin. Books in the contemporary collection were manually curated to represent popular writing aimed at reaching different readerships (i.e. “genres”). While the term “genre” has been understood in multiple ways within

Table 1

The five most common super-sense tags and most-frequent associated words

Noun Type	Examples	Verb Type	Examples
person	man,mother,father	stative	was,is,be
artifact	door,room,house	communication	said,say,told
communication	name,words,word	motion	go,came,come
location	place,side,city	contact	put,stood,sat
body	eyes,head,face	cognition	know,think,knew

Table 2

Primary super-sense tags used to represent things in this paper

Thing	Examples	Category
artifact	door,room,house	human-made
object	world,ground,sky	organic
animal	horse,dog,animals	organic
food	food,coffee,dinner	organic
plant	trees,tree,grass	organic

the research community [16, 30], we define genre for our purposes as a form of institutionally framed classification [19], where we use three broad categories of framing: cultural capital (bestsellers, prizewinners, elite book reviews), stylistic affinity (mysteries, science fiction, romance, etc.), and age-level (middle-grade and YA).

In order to detect entity-types in our data, we process all texts using bookNLP [4], which implements BERT-based models and has been shown to outperform other state-of-the-art systems for a variety of NLP tasks when applied to literary texts [5]. In addition to performing entity recognition, part-of-speech-tagging, and dependency parsing, bookNLP also provides 41 “super-sense” tags trained on SemCor’s implementation of the Wordnet taxonomy, examples of which are shown in Table 1. In Table 2, we provide a list of noun-types that we use to capture two categories of things as either human-made or organic. According to bookNLP’s documentation, the supersense tagging has an accuracy of 76% when applied to works of fiction.

Prior work on the computational detection of objects in fiction has applied a dictionary-matching approach using the Wordnet taxonomy to a small set of sixty nineteenth-century novels [28]. Our work updates this work by a) examining a considerably larger collection of fiction and b) using predictive modeling for word-type tagging. Problems of word disambiguation surrounding dictionary-matching methods are well known [24], which may prove especially problematic in fiction when, for example, highly prevalent character names (such as “Iris,” “Ivy” etc.) can be confused for objects when taken out of context. Additionally, in contrast to Tenen [28] who only counts entities that appear as grammatical objects, we condition on all appearances of nouns given that objects can play an important role as the subject of actions. Our results thus depend on the predictions made by bookNLP’s super-sense tags for all nouns in a given book.¹

¹All data and code can be found in the following repository: <https://doi.org/10.6084/m9.figshare.21382020.v1>

3. Results

3.1. The prevalence of things

Fig. 1 provides an overview of the distribution of noun-types across three historical periods. As we can see, humans dominate fictional narrative regardless of historical time-frame. Nevertheless, human-made things (“artifacts”) are second, a surprising result given no prior theory has indicated this kind of predominance. Indeed, we find that the rate of artifacts across all three collections is 2.4x higher than all other natural objects combined and 3.3x higher if we condition on the contemporary data. As Fig. 1 indicates, the prevalence of artifacts increases from the nineteenth-century to the present (note the reversed order of persons and artifacts by period).

We can also observe that this increase of artifacts over time is specific to fiction (Fig. 2). Despite the simultaneous growth of artifacts in fiction and non-fiction over the course of the nineteenth century, by the early twentieth century we see fiction investing more heavily in the utilization of human-made things. By the end of the twentieth century, artifacts in fiction are almost 50% more frequent than in non-fiction, averaging roughly 3 more occurrences per page than an equivalent work of non-fiction. Artifacts also represent the semantic category with the single strongest growth over this time period. Human-made things come to dominate fiction beyond what we expect from their overall prevalence in English-language writing more generally.

3.2. The nature of things

To better understand the nature of human-made things and their distribution in fiction, we break down the artifact category by various sub-types. To do so, we use Wordnet’s hypernym taxonomy. For each word predicted to be an artifact in our dataset, we crawl the hypernym tree for the word’s most common sense and store all types that occur below the level of “artifact.” For example, the word “house” would produce the following sub-types in the following order of ascending generality: *dwelling*, *housing*, *structure*. We condition on 2,500 / 1,800 of the most common artifacts in the contemporary / Hathi collection, which account for just under 80% / 90% of all occurrences of artifacts in each collection. We then manually review and clean the labels, producing the following sub-types and their counts in Table 3.

As we can see, “structures” are the most common kind of object in fiction by a factor of 1.5 over the next most frequent type, human-made instruments. Indeed, the frequency of structures is higher than all of the organic objects combined. This holds for the Hathi1M collection with the exception that there are considerably more natural objects utilized in the nineteenth- and twentieth-century data. Beyond this divergence, the ranking and even the top words of the various categories are remarkably similar (full data provided in the supplementary material).

3.3. Semantic frameworks of things

With a clearer picture of the distribution of the types of fictional things, we take the next step of trying to better understand the semantic identity of these things, and structural things in particular. Knowing that human-made things dominate fiction in terms of object types and

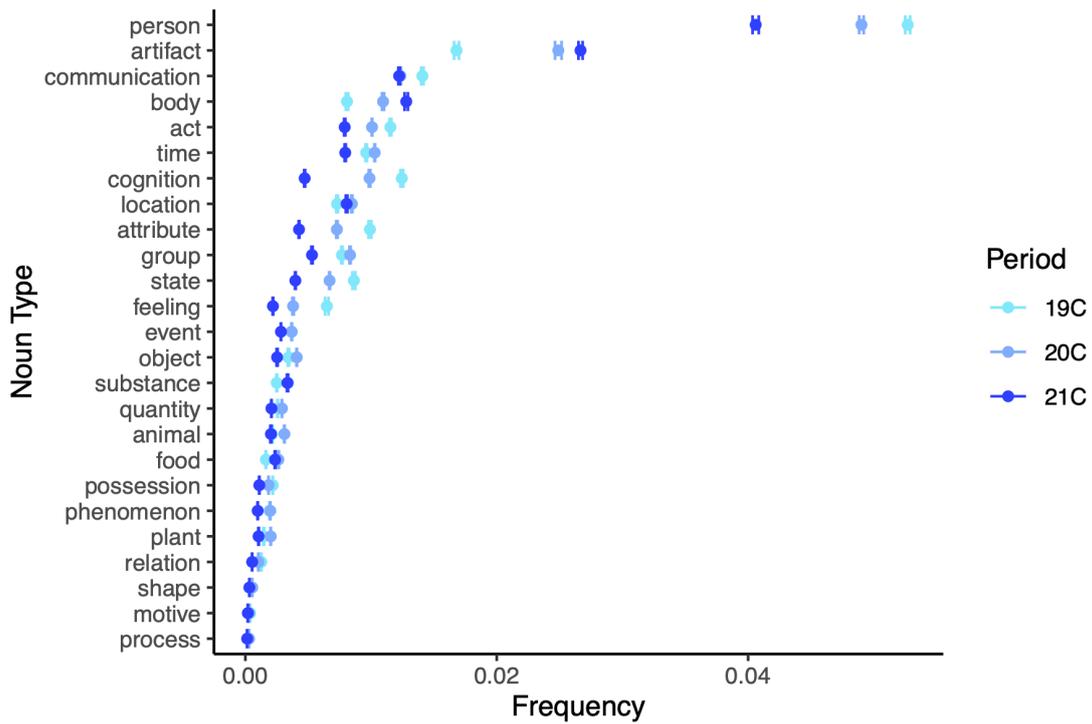


Figure 1: Distribution of entity-types by historical period

knowing that within that group structures are the most common sub-type can give us important information about the potential narrative function of fictional things. Here we wish to deepen our understanding by undertaking a series of tests to gain further insights about the semantic frameworks that surround structural things. We note that when we refer to “structures” in this section we combine words labeled as “structures” and “furnishings” because we consider tables, chairs, etc to also function as structural forms for our purposes.

3.3.1. Modification

Wall [31] has argued that by the nineteenth century fictional things assume a more ornamental or decorative function, that is, they move from being inert, non-descript objects like “the pot” famously identified by Virginia Woolf in her essay on Robinson Crusoe [34] to serving as elaborate descriptive props for writers like Balzac, Flaubert, and Dickens. In order to test this theory, we measure the ratio of the rate of the modification of artifacts divided by the rate of modification of other kinds of entities. Doing so allows us to understand how much more likely artifacts are to be modified given the overall rate of modification per book / year, which we assume varies from book to book and from year to year. To detect whether an entity has been modified, we condition on the following three forms of grammatical modification: adjectival modifiers, participial modifiers, and clausal modifiers, which are represented by the Stanford

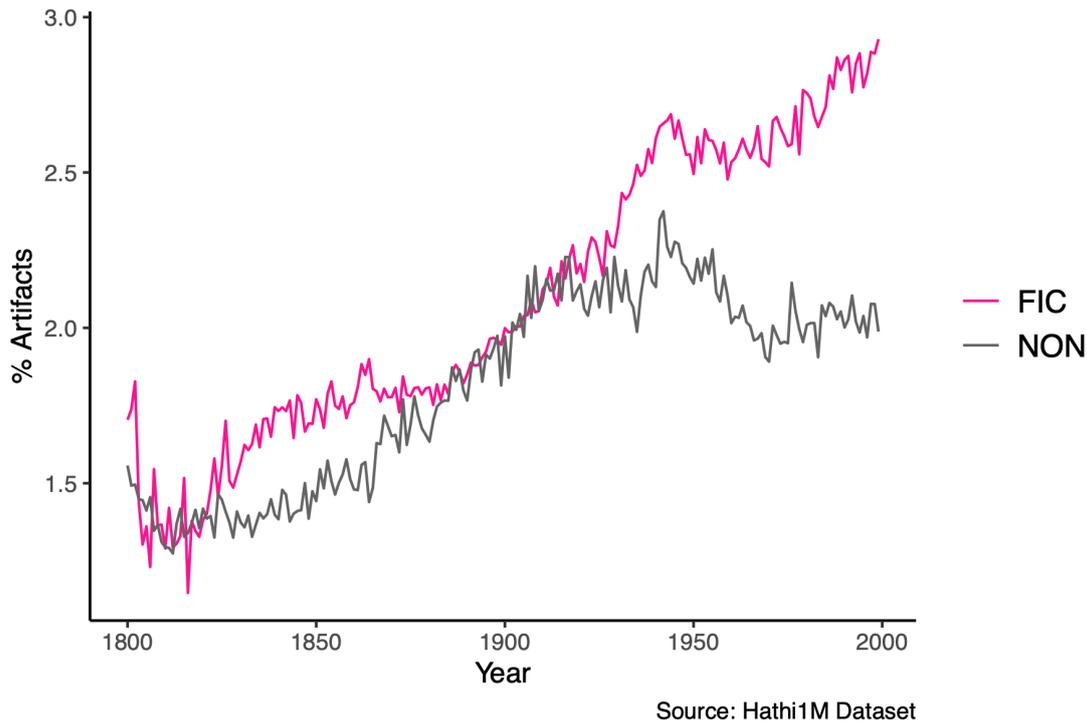


Figure 2: Distribution of artifacts over historical time in fiction and non-fiction

Dependency Parser as “amod,” “partmod,” and “rcmod” respectively. We show examples of each form in Table 4 and the results of our comparisons in Table 5.

3.3.2. Affect

In addition to description, we also measure the rate of “affect” associated with different kinds of fictional things. Research has emphasized the ways in which objects in fiction are often embedded in psychically or emotionally charged contexts [12]. In order to measure relative levels of affect surrounding artifacts, structures, and other kinds of entities, we use the NRC valence-arousal-dominance lexicon [23] and calculate the average valence and arousal for a random sample of sentences drawn from each book / year. We then subset our samples by whether they contain artifacts and structures and compare the distributions of those values for each historical period. We use Cohen’s d as a measure of effect size when comparing categories.

As we can see in Table 5, when it comes to modification we see a small increased rate of modification of artifacts when compared to other kinds of entities, which shifts in the opposite direction when we condition only on structures and furnishings. Structures are unique in their lower levels of description. For affect, we see a small negative relationship between artifacts and their valence scores for the historical data and a large negative relationship for the contemporary collection, which grows stronger as we move from all artifacts to our subset of

Table 3

List of sub-types and counts in the contemporary collection with the most frequent words for each type.

Category	Type	Count (per 100K)	Examples
artifact	structure	874	door,room,house
artifact	instrumentality	557	phone,gun,bag
artifact	clothing	277	clothes,shirt,pocket
organic	object	254	world,ground,sky
organic	food	225	food,coffee,dinner
organic	animal	197	horse,dog,animals
artifact	furnishing	149	bed,table,chair
artifact	vehicle	135	car,ship,truck
organic	plant	102	trees,tree,grass
artifact	creation	63	picture,pictures,photo
artifact	line	9	rope,string,napkin
artifact	plaything	5	cards,doll,toy

Table 4

Examples of our three types of modification. Entities are underlined and modifiers appear in italics

amod	partmod	rcmod
<i>broad</i> <u>streets</u>	the <u>net</u> became <i>cracked</i>	the <u>tube</u> that <i>surrounded</i> it
<i>short</i> <u>sleeves</u>	a <u>hook</u> <i>measuring</i> about two inches	the <u>coffin</u> which <i>sat</i> on the table

Table 5

Rates of modification and valence scores associated with different kinds of entities for each period. Cohen's d represents the effect size when comparing the distributions of rates associated with a given type and all other entity types (e.g. artifacts v. non-artifacts).

Period	Type	Measure	d
19C	Artifacts	Modification	0.001 (negligible)
20C	Artifacts	Modification	0.36 (small)
21C	Artifacts	Modification	0.33 (small)
19C	Artifacts	Affect	-0.29 (small)
20C	Artifacts	Affect	-0.36 (small)
21C	Artifacts	Affect	-1.57 (large)
21C	Structures	Modification	-0.13 (small)
21C	Structures	Affect	-1.79 (large)

structures. This suggests that structures are functioning less ornamentally and less affectively than other kinds of fictional entities.

3.3.3. Narrative Distribution

Another way to understand the meaning of objects in fiction is to observe their distribution across narrative time [10]. Fictional things may play a strategic role by occurring at key turning points or narrative entry or exit points or conversely they may play a more infrastructural role by occurring regularly throughout narrative time. As Boyd [10] has shown, different linguistic features exhibit meaningful changes over narrative time. In Figure 3, we plot the fraction of structures and other artifacts over narrative time for our contemporary collection. As we can see, we observe a preference for human-made things to appear towards the opening sections of a narrative, suggesting that things function partially as “narrative establishments.” However, in real terms the change is very slight, with a decline of roughly 15 total mentions of structures between the first and final sections of an average narrative. In other words, evidence also suggests that in addition to establishing narrative frames, structures and other artifacts also play an infrastructural role in their continued presence over narrative time.

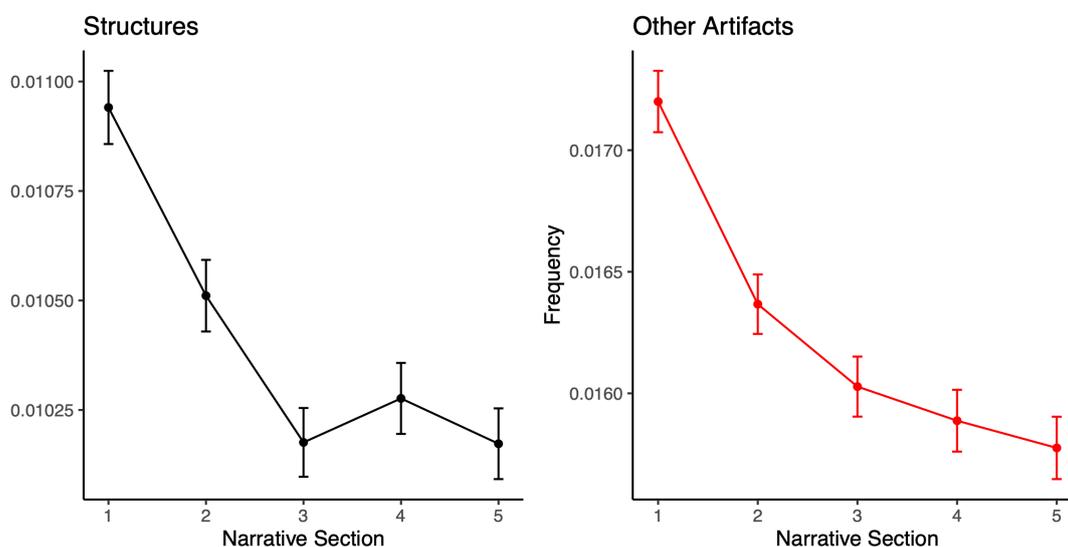


Figure 3: Frequency of structures and other artifacts over narrative time

3.3.4. Association

Finally, we explore collocate word types associated with structures to better understand how they are semantically contextualized. As the theory of distributional semantics suggests, the company that words keep has a strong influence on their meaning [29]. We thus calculate the likelihood of a word-type appearing in a sentence with a structure compared to sentences that only contain other kinds of non-human-made entities (Table 6). Once again we use bookNLP’s super-sense tags rather than condition on individual lexemes.

As we can see, structures are most strongly associated with actions associated with contact and motion and entities associated with substances and locations, while they are negatively cor-

Table 6

Word types positively and negatively associated with structures across all genres in the contemporary collection. G represents the likelihood ratio of each word type collocated with structure words.

Positive Association	G	Negative Association	G
verb.motion	313144	verb.communication	77159
verb.contact	200322	noun.communication	57418
noun.location	51015	noun.person	35933
noun.substance	39907	verb.cognition	35847
verb.perception	20811	noun.cognition	27290

related with actions of communication, cognition, and emotion. In other words, when readers encounter structural things in fiction, they are considerably less likely to encounter moments of affect or cognitive reflection. Structural things are more likely to cause characters to move and touch, not think and emote.

4. Discussion

Summarizing our findings to this point, we can say that according to our data the most common kinds of things in fiction are structural in nature, i.e. human-made and supportive. These may include rooms, houses, doors, windows, tables, roads, kitchens, walls, stairs, apartments, among many other kinds of things. Such things are less likely to be modified and less likely to be associated with affective feelings. They are consistently deployed throughout narrative time and they are more likely to facilitate corporeal rather than cognitive or communicative behavior in characters.

These findings raise challenges for some of the more prominent scholarly arguments that have fallen under the heading of thing theory. Brown’s claim that fictional things are important because of their “labor of infusing manufactured objects with a metaphysical dimension” [11] or Calvino’s claim that “in a narrative any object is always magic” (33) [13] do not fit well with the semantic nature of the structural things that we have shown here. An individual door or room may be magical or metaphysical (e.g. the door in Tieck’s *Bluebeard’s Castle* or Gregor’s room in Kafka’s *Metamorphosis*), but broadly speaking such structural things are less likely to be associated with ornamentation, affect, or cognition. Their semantic inertia appears to matter more to their identity rather than their individual affective or descriptive depth.

Similarly, Freedgood’s [17] assertion concerning the denotative value of things – the way their value lies in their ability to point outwards to the world – also does not capture the broad semantic behavior of fictional things that we are seeing. While readers may connect these objects to their historical life contexts, in terms of their narrative function within the texts themselves it is far more likely that the rooms, houses, tables, and doors that predominate serve as generic props rather than point deictically to rich historical or inner mental worlds. The same holds for the longstanding critical interest in technological things [21, 27, 33]. While the impulse to write about technologies like telephones, radios, and typewriters captures the predominance of human-made objects within fiction since the nineteenth century, it fails to

account for the most common kind of artifactuality within fiction, that of structural things.

Barthes [6] has argued that such objects are in fact “narratively useless,” i.e. their function is to resist interpretation and instead signify the idea of “referentiality” (143). They are there merely to produce what he calls a “reality effect.” Lamb [22] makes a similar point in arguing that the value of fictional things is their “irrelevance to any human system of value” (11). While these theories address the semantic inertia of structural things (i.e. their supporting or background nature), they do not provide a framework for understanding why so many seemingly useless things are so common in fictional storytelling. Rather than see these highly frequent objects as “useless” or “irrelevant” – akin perhaps to the idea of junk DNA – we would argue that another theoretical framework outside of thing theory may provide a productive means of understanding these objects’ narrative function.

The framework we would suggest falls under the heading “embodied cognition,” an increasingly studied (and still debated) framework within cognitive science [26] that has also found increasing resonance in literary studies [2, 9, 14]. The key argument that embodied cognition makes is that thinking is not localized in the brain but transpires through body-environment interactions. Thought is distributed throughout one’s object world.

One way to understand the predominance of structural things in fiction, then, would be as a means of activating this idea of “embodied cognition” through fictional narration. Structural things shift the focus, as Randall Beer, one of the early proponents of embodied cognition, argued, “from accurately representing an environment to continuously engaging that environment with a body so as to stabilize appropriate co-ordinated patterns of behavior” (97) [7]. The strong corporeal and low decorative aspects of such objects may potentially produce, in Andy Clark’s words, “a constantly available channel that productively couples agent and environment” (15) [15]. The persistent recurrence of such objects are thus neither narratively useless nor to be understood principally as vehicles of introspection. Rather, they may be means of activating an experience of “the extended mind,” enabling the experience of embodied cognition in the minds of readers through these objects’ recurrent physical and semantically inert presence. Seen in this light, one of fiction’s modern social functions could be that it helps readers activate the particular mode of thought known as embodied cognition.

By empirically accounting for the types of things in fiction and their semantic identity across large number of documents, our findings pose challenges to inherited critical theories about the role of things in fictional narratives. While those theories are undoubtedly valid for the individual objects and works they address, they fail to account for the most predominant kinds of fictional things and their semantic behavior across large amounts of literature. Rather than argue that such things are unimportant, we contend that it is highly important to account for the most prevalent kinds of things in fiction if we are to understand the function of fictional things.

While we suggest embodied cognition as one possible avenue for understanding the function of such structural things, future work can test this theory further either through more text samples drawn from different cultural domains and languages or through empirical reader studies. Work in embodied cognition has a long history of measuring human attention and problem solving and many of these approaches could be productively applied towards understanding reader behavior. The coupling of large-scale observational data with empirical reader studies offers an ideal synthesis to continue to better understand the social and psychological functions

of fictional things.

Acknowledgments

This research was generously supported by the Social Sciences and Humanities Research Council of Canada (895-2013-1011).

References

- [1] A. Appadurai. *The Social Life of Things: Commodities in Cultural Perspective*. Cambridge University Press, 1988.
- [2] P. B. Armstrong. *Stories and the Brain: The Neuroscience of Narrative*. JHU Press, 2020.
- [3] S. Bagga and A. Piper. “HATHI 1M: Introducing a Million Page Historical Prose Dataset in English from the Hathi Trust”. In: *Journal of Open Humanities Data* 8 (2022).
- [4] D. Bamman. *BookNLP, A Natural Language Processing Pipeline for Books*. 2022. URL: <https://github.com/booknlp/booknlp>.
- [5] D. Bamman. “Litbank: Born-literary natural language processing”. In: *Computational Humanities, Debates in Digital Humanities* (2020).
- [6] R. Barthes and R. Howard. “The Reality Effect”. In: *The Rustle of Language*. 1981, pp. 141–149.
- [7] R. D. Beer. “Dynamical Approaches to Cognitive Science”. In: *Trends in Cognitive Sciences* 4.3 (2000), pp. 91–99.
- [8] B. Bill. “Thing Theory”. In: *Critical Inquiry* 28.1 (2001), pp. 1–22.
- [9] G. Bolens. *Kinesic Humor: Literature, Embodied Cognition, and the Dynamics of Gesture*. Oxford University Press, 2021.
- [10] R. L. Boyd, K. G. Blackburn, and J. W. Pennebaker. “The Narrative Arc: Revealing Core Narrative Structures Through Text Analysis”. In: *Science Advances* 6.32 (2020), eaba2196.
- [11] B. Brown. *A Sense of Things: The Object Matter of American Literature*. University of Chicago Press, 2010.
- [12] B. Brown. *Other Things*. University of Chicago Press, 2016.
- [13] I. Calvino. *Six Memos for the Next Millennium*. Harvard University Press, 1988.
- [14] M. Caracciolo and K. Kukkonen. *With Bodies: Narrative Theory and Embodied Cognition*. Ohio State University Press, 2021.
- [15] A. Clark. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press, 2008.
- [16] R. Cohen. “History and Genre”. In: *New Literary History* 17.2 (1986), pp. 203–218.
- [17] E. Freedgood. *The Ideas in Things*. University of Chicago Press, 2009.

- [18] C. Gallagher and S. Greenblatt. *Practicing New Historicism*. University of Chicago Press, 2000.
- [19] C. Garrido Castellano. “The Institution of Institutionalism: Difference, Universalism and the Legacies of Institutional Critique”. In: *Culture, Theory and Critique* 59.1 (2018), pp. 59–73.
- [20] M. Jacobus. *Romantic Things*. University of Chicago Press, 2012.
- [21] F. A. Kittler and J. Johnston. “Gramophone, Film, Typewriter”. In: *Literature Media*. Routledge, 2013, pp. 31–49.
- [22] J. Lamb. *The Things Things Say*. Princeton University Press, 2022.
- [23] S. Mohammad. “Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (volume 1: Long papers)*. 2018, pp. 174–184.
- [24] R. Navigli. “Word Sense Disambiguation: A Survey”. In: *ACM Computing Surveys (CSUR)* 41.2 (2009), pp. 1–69.
- [25] A. Piper. “The CONLIT Dataset of Contemporary Literature”. In: *Journal of Open Humanities Data* 8 (2022).
- [26] L. Shapiro. *Embodied Cognition*. Routledge, 2010.
- [27] B. Siegert. *Cultural Techniques*. Fordham University Press, 2015.
- [28] D. Y. Tenen. “Toward a Computational Archaeology of Fictional Space”. In: *New Literary History* 49.1 (2018), pp. 119–147.
- [29] P. D. Turney and P. Pantel. “From Frequency to Meaning: Vector Space Models of Semantics”. In: *Journal of Artificial Intelligence Research* 37 (2010), pp. 141–188.
- [30] T. Underwood. “Genre Theory and Historicism”. In: *Journal of Cultural Analytics* 2.2 (2016), p. 11063.
- [31] C. S. Wall. *The Prose of Things*. University of Chicago Press, 2014.
- [32] S. Wasserman. *The Death of Things: Ephemera and the American Novel*. U of Minnesota Press, 2020.
- [33] D. S. Wershler-Henry. *The Iron Whim: A Fragmented History of Typewriting*. Cornell University Press, 2007.
- [34] V. Woolf. *The Second Common Reader*. Houghton Mifflin Harcourt, 1986.