







Small Worlds

Measuring the Mobility of Characters in English-Language Fiction

Matthew Wilkens¹ 
Elizabeth F. Evans²
Sandeep Soni³
David Bamman⁴ 
Andrew Piper⁵ 

1. Information Science, Cornell University , Ithaca, USA.
2. English, Wayne State University , Detroit, USA.
3. Quantitative Theory and Methods, Emory University , Atlanta, USA.
4. School of Information, University of California , Berkeley, USA.
5. Languages, Literatures, and Cultures, McGill University , Montréal, Canada.

Citation

Matthew Wilkens, Elizabeth F. Evans, Sandeep Soni, David Bamman, and Andrew Piper (2024). "Small Worlds. Measuring the Mobility of Characters in English-Language Fiction". In: *CCLS2024 Conference Preprints* (3). [10.26083/tuprints-00027393](https://doi.org/10.26083/tuprints-00027393)

Date published 2024-05-28

Date accepted 2024-04-04

Date received 2024-01-19

Keywords

fiction, mobility, geospatial analysis, narratology

License

CC BY 4.0 

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 3rd Annual Conference of Computational Literary Studies at Vienna, Austria, in June 2024.

Abstract. The representation of mobility in literary narratives has important implications for the cultural understanding of human movement and migration. In this paper, we introduce novel methods for measuring the physical mobility of literary characters through narrative space and time. We capture mobility through geographically defined space, as well as through generic locations such as homes, driveways, and forests. Using a dataset of over 13,000 books published in English since 1789, we observe significant "small world" effects in fictional narratives. Specifically, we find that fictional characters cover far less distance than their non-fictional counterparts; the pathways covered by fictional characters are highly formulaic and limited from a global perspective; and fiction exhibits a distinctive semantic investment in domestic and private places. Surprisingly, we do not find that characters' ascribed gender has a statistically significant effect on distance traveled, but it does influence the semantics of domesticity.

1. Introduction

What does it mean for a novel's characters to be mobile? And what effects does spatial mobility have on the novel, the story world it imagines, and the novel's greater cultural significance?

Narrative, especially long narratives, almost always involve a change of location or setting. This is an essential component of what narrative theorists identify as the world-building/world-changing function of narration (Bruner 1991; Herman 2009). Whereas setting was once regarded as the unimportant "background" of fictional narrative, it is now broadly recognized as a vital interface with the material and social world (Evans 2025; Evans and Wilkens 2024; Hones 2022; Ryan et al. 2016; Tally Jr 2012). As Friedman 1998 summarizes, "Setting works as symbolic geography, signaling or marking the specific cultural locations of a character within the larger society."

For some genres – the travelogue, the quest narrative, the adventure story, even the Bildungsroman – movement through space is an essential component of the genre’s meaning and identity. The inter-relatedness of space and time in narrative – that the movement through space involves a movement through time – has been influentially theorized by Bakhtin 2010 as the concept of the *chronotope*. For Bakhtin, the space-time nexus has a generative function when it comes to narrative.

In this paper, we introduce novel methods for measuring the physical mobility of characters through narrative space and time. We capture mobility in two fundamental ways. First, we define mobility as the movement through geographically-defined space and measure the distance that characters travel between countries, cities, regions, and other mappable places. Second, we examine mobility as movement through the non-geographic semantic spaces of rooms, streets, and others “generic” locations.

The geographic plotting of novels has long been theorized as an important component in constructing narrative meaning (Moretti 1999; Piatti et al. 2009; Ryan et al. 2016; Wilkens 2013). To take one classic example, the characters of Jack Kerouac’s *On the Road* (1957) travel not only because they want to get from point A to point B (at the novel’s start, New York City to Denver), but also because the road represents to them freedom, discovery, adventure, sex, and, for the narrator, Sal Paradise, writing material. When Sal reflects on his younger self, “I was a young writer and I wanted to take off,” he makes use of the double-meaning of “take off” – he wants his writing career to blossom, and he wants to be in motion. The two, and all that being on the road represents to Sal, are necessarily connected: “Somewhere along the line I knew there’d be girls, visions, everything; somewhere along the line the pearl would be handed to me” (Kerouac 2002, 8). For the “girls” Sal and his friends meet along the way, travel is a less-viable choice. While many of them also long for new horizons, women are generally represented by Sal and by the novel as a feature of the landscape, rooted in place, and as lacking in intellectual range as they are in geographic reach. Movement through geographically-defined space captures the variety of ideological meanings embedded in mobility, as well as the range of cultural restrictions imposed upon it.

In addition to this focus on geographic space, we also measure movement through what we term “generic space.” For many narratives, mobility may be characterized as a movement between generic spatial entities such as rooms, streets, parks, forests and homes. In Marilyn Haushofer’s feminist novel *The Wall* (*Die Wand*) from 1963, an invisible wall rises up one day to cut off the unnamed protagonist from the rest of the world. The remainder of the novel involves her moving back and forth between rural hunting lodges and the wall in the Austrian alps. In this case, movement through generic rather than geographically specified space grounds the novel’s reflections on the constraints of female identity, rooting the novel in a more allegorical mode.

Our work is thus tied to prior research in the broader area known as the spatial humanities (Bodenhamer et al. 2010; Roberts et al. 2014). Whether qualitative or computational in nature, this work is grounded in the significance of spatial structures for understanding cultural and narrative meaning. Where prior work often captured space as a static construct (the atlas or map as the principle theoretical frame), the concept of mobility can be a useful addition to this work by adding a dimension of narrative time.

Mobility, then, is a way of understanding the world-building function of fictional narratives. How and where characters move through space is integral to the construction of narrative meaning as much as the specific qualities of the individual places themselves. Modeling mobility at large scale can thus begin to provide insights into the more general chronotopes that shape storytelling across different cultures, genres, and historical time periods.

In the body of our paper, we first describe and validate the model we use to predict narrative mobility derived from prior work (Soni et al. 2023). We then describe a variety of measurements of mobility based on this model as applied to two primary datasets. The first is the CONLIT corpus of contemporary prose, which includes 2,754 works of English prose published since 2001 drawn from twelve different genres. The second is a collection of 10,629 novels by American authors published between 1789 and 2000.

As a way of understanding the function of the different kinds of mobility we are interested in, we examine the relationship between our mobility measurements and particular social categories. These include the effects of instrumentality (fictional versus non-fictional narratives), prestige (award-winning novels versus bestsellers), audience age-level, and pronoun-signaled character gender on character mobility.

Questions of narrative mobility – of what mobility is and how we recognize it – matter when we consider the significance of mobility for human cultures more generally. For Cresswell 2006, “mobility is central to what it is to be human.” Not only do people move from the moment of birth, but cultures blend, splinter, and evolve. And because mobility carries ideological meanings, it also shapes the stories we tell. As Cresswell emphasizes, the modern Western meaning of mobility is not stable: “[m]obility as progress, as freedom, as opportunity, and as modernity, sit side by side with mobility as shiftlessness, as deviance, and as resistance” (1-2). As *On the Road* suggests, the two understandings of mobility can even coexist within a single text. One of the consistent attributes of mobility is its ability to participate in a shifting process of meaning-making. This paper aims to introduce methods for understanding the dynamics of character mobility within literary narratives as part of a broader goal of understanding how mobility has been framed and understood over time.

2. Data and Methods

2.1 Data

We work with a corpus of 13,383 books published between 1789 and 2021. All books are in English; the large majority are works of fiction. The corpus was assembled from a range of sources as described below. The distribution of volumes across subcorpora is shown in table 1.

All subcorpora except CONLIT contain only fiction. As detailed in Piper 2022, CONLIT contains twelve different genres distributed across fiction and non-fiction writing, dating from 2001-2021. EAF and Wright comprise subsets of the novelistic fiction by US authors cataloged in Wright 1965 and digitized by a consortium of academic libraries (Digital Library Program 2012; Electronic Text Center 2000). Chicago I and II include novels by American authors published between 1880 and 2000, sourced from the Chicago Text

Collection	Label	Books	Begin	End
Early American Fiction	EAF	488	1789	1850
Wright Bibliography of American Fiction	Wright	1,052	1850	1875
Chicago Novel Corpus I	Chicago I	2,608	1880	1945
Chicago Novel Corpus II	Chicago II	6,481	1946	2000
CONLIT Contemporary Literature	CONLIT	2,754	2001	2021

Table 1: Subdivisions of the research corpus.

Lab (Long and So 2020).

Our corpus offers nearly uninterrupted coverage of American fiction over more than 230 years. It is especially rich in twenty-first-century writing, for which it contains extensive metadata concerning instrumentality, prestige, and audience type. The corpus does not include a meaningful amount of writing by non-North American authors, nor writing originally published in languages other than English. For this reason, our analysis and conclusions should be understood to apply primarily to the North American, English-language contexts that are well represented in our source collections.

2.2 Methods

2.2.1 Modeling Sequences of Places

From each volume in our corpus, we extract the ordered sequence of locations associated with each of its characters using the method developed in Soni et al. 2023. In brief, we use BookNLP (Bamman 2020, 2021) to identify characters and locations that cooccur within a rolling ten-token window in each source text. The same system performs coreference resolution, consolidates multiple forms of address to single characters, and records pronominally signaled character genders. We then train a BERT-based model to identify possible relationships (including NO RELATION) between each cooccurring character-location pair. From the full set of cooccurrences, we select those that describe a character as occupying the identified location (having relation IN). This method differs significantly from earlier work, in that it allows us both to place characters in specific locations and to trace character movements over narrative sequences.

The locations identified may be geopolitical entities (GPEs), such as nations or cities, facilities (FACs), such as homes or offices, or other locations (LOCs; typically natural settings). In principle, any of these locations might correspond to real, mappable places (England, Mt. Everest) or to imaginary or generic entities (the house, a street corner, Hogwarts). In practice, most GPEs are real and mappable; most FACs and LOCs are not. We separate our character sequences into GPEs and others. For GPEs, we retrieve detailed geographic information from open and commercial sources as described in Evans and Wilkens 2018. For non-GPEs, we remove stopwords ([the house | a house | her house] → house), but do not perform geolocation.

After processing, we have two lists of locations (GPEs and others, respectively) that are occupied sequentially by each character in each book. In some of our experiments, we are interested in transitions between locations. We call each case in which a character occupies a location different from the one immediately preceding it a *hop*. For example, a character having the GPE sequence [London, Boston, California] undergoes two hops,

London → Boston and Boston → California. If a character occupies the same location multiple consecutive times, we treat that sequence of unchanging locations as single instance. For GPE sequences, we exclude hops for which the distance between locations is conceptually ill-defined, such as London → England or California → USA.

2.2.2 Measurements

Here we present the primary measures used in our analysis, along with a list of dependent variables analyzed in table 5. In most cases, we restrict our calculations to the single most commonly occurring character in each book, which we call the *protagonist*. We condition on protagonists because we observe that the majority of overall mobility in the average book is associated with the most frequently occurring character.

Distance: The total geodesic distance (in miles) between sequences of geographic places (GPEs) that are inhabited by the book’s protagonist. This represents the sum of the distances traversed over all valid hops for the character. We exclude a subset of common hop types that are conceptually ill-defined, including hops between cities and the first-level administrative regions (states, provinces, etc.) or nations that contain them, and between first-level regions and the nations to which they belong. We allow hops between any locations at the same administrative level (city to city, state to state) and between different administrative levels when the lower-level location is not contained by the higher-level one (for example, neither Los Angeles → California nor Los Angeles → United States is allowed, but Los Angeles → Iowa is). We make an exception for hops involving continents, which we allow.

GPEs: Number of distinct geographic places inhabited by the main character (India, Toronto, New York, California).

Generics: Number of distinct generic places inhabited by the main character (room, kitchen, street, yard). These are annotated as LOC and FAC by BookNLP.

Semantic Distance: The average semantic distance between all sequentially inhabited generic places. Semantic distance is calculated as one minus the cosine similarity between word vectors for each generic place using the Glove 6B Wikipedia pretrained model with 100 dimensions (Pennington et al. 2014). Multi-word phrases average each word’s vector in the phrase. Stop words and punctuation are removed. Semantic distance aims to capture the semantic similarity of places given a general understanding of those terms.

Deictics: Frequency of “here” and “there” relative to all generic place names per book.

Generic / GPE Ratio: Total number of generic locations divided by the total number of GPEs per book.

Character Count: Number of references to a book’s protagonist.

Tokens: Total tokens per book.

Start Finish Miles: The direct geodesic distance between the first and last locations inhabited by the protagonist of each book.

2.2.3 Independent Variables used for CONLIT 173

The number of documents for each class are listed in parentheses. 174

Fictionality: The category designation between FIC (fiction) (1,934) and NON (non-fiction) (820). Also referred to as “instrumentality.” 175
176

Prestige: Sub-divided between genre labels PW (prizewinners) (258) for high prestige 177
and BS (bestsellers) (249) for low prestige. 178

Youth: Sub-divided between genre labels MID (middle-grade books) (166) and NYT 179
(*New York Times* reviewed), PW, and BS (926). 180

Female: Uses the inferred gender categories “she/her/hers” (744) and “he/him/his” 181
(1,180) for protagonists in fiction. The very small number of other pronominal designa- 182
tions are removed. 183

2.2.4 Distance Validation 184

The computational pipeline by which we produce our hop sequences and distance 185
measurements is complex and subject to multiple uncertainties. To validate our results, 186
we examined 10,000-word chunks extracted from the beginning of 30 novels sampled at 187
random from the CONLIT subcorpus. For each sample, we annotated by hand the set of 188
true geographic locations occupied by the main character; determined the geographic 189
coördinates of those locations; and calculated the distance traversed by that character. 190
We also labeled each sample’s holistic mobility from 1 (lowest mobility) to 5 (highest 191
mobility). We found that our algorithmic distance was linearly correlated with human 192
measurements at $R^2 = 0.525$ ($p \approx 0$). We also found that the mean distance traveled by 193
protagonists in high-mobility samples (those with ratings of 4 or 5) was much higher 194
than the mean distance traveled in low-mobility samples (ratings 1 or 2; $\bar{x}_{high}/\bar{x}_{low} = 3.6$; 195
 $p < 0.008$). We note as well that randomly distributed errors in our pipeline will tend 196
to reduce the observed significance of results derived from our data, hence that we 197
generally understate the statistical significance of our findings (see Spearman [1904] 198
1987). We are thus confident that our GPE-derived distance measures serve in aggregate 199
as an acceptable class of proxies for character mobility. 200

2.2.5 Regression Analysis 201

To evaluate the impact of each social category, which serve as our independent variables, 202
we conducted a linear regression analysis. For this analysis, we incorporated binary 203
dummy variables corresponding to each primary class, namely fiction, prestige, youth, 204
and female character. Additionally, we introduced control variables to account for 205
potential confounding factors, such as genre, point of view, book length (measured in 206
tokens), and character mention frequency (character count). 207

The outcomes of this analysis, including the directionality of the effect for each depen- 208
dent variable and the statistical significance represented by p -values, are summarized 209
in table 5. In our supplementary materials, we present comprehensive results, encom- 210
passing sample mean estimates, R^2 values, and the precise p -values obtained from the 211
analysis. 212

It is important to acknowledge the significance of our chosen control variables due to the variability they exhibit in our data. For instance, non-fiction texts exhibit a higher average length compared to fiction, whereas fiction registers a markedly higher average character count, with fictional protagonists being referenced significantly more frequently. Consequently, employing a uniform normalization technique would be inadequate to address the multifaceted disparities inherent in our dataset.

3. Results

Overall Distance. In table 2, we show the mean distance traveled, mean number of unique GPEs, and mean number of unique generic locations in each of our subcorpora.¹ Figure 1 visualizes the evolution in these quantities over time. As we can see, the average number of unique places, whether GPE or Generic, has more than doubled since the nineteenth century along with the total distance traveled by primary characters.

Collection	Distance	GPEs	Generics	Hops
EAF	13,139	5.9	37.5	5.8
Wright	10,477	5.3	43.8	4.9
Chicago I	21,026	8.4	72.9	9.3
Chicago II	37,023	13.8	113.0	16.3
CONLIT fiction	38,024	13.3	123.9	15.6
CONLIT nonfiction	131,263	35.8	120.8	60.8

Table 2: Means of distance, number of unique GPEs, number of unique generic locations, and number of hops by subcorpus.

Routes Traveled. Figure 2 presents a global map capturing the movement by protagonists between places in fictional narratives. This figure plots the aggregate hops taken by all fictional protagonists over the full corpus; the width of the line connecting each (undirected) origin and destination is proportional to the share of all hops represented by that location pair. While we visualize here only the aggregated results for the full corpus, the supplemental materials provide visualizations by subcorpus and by historical era. There is very little variation in the high-level appearance of this map over historical time. As table 3 further illustrates, the patterns of movement between places within (broadly American) fiction are highly stable and formulaic over historical time.

Gender and Mobility. Previous work has found that novels enriched in she/her characters contain fewer GPEs and that the GPEs in those narratives are less widely separated than are those in he/him-enriched novels (Evans and Wilkens 2024). As shown in table 4, we calculate the mean distance traveled and the count of unique GPEs / generics by pronominally indicated character gender. We find over the full corpus that the average male-gendered protagonist in fiction occupies more unique GPEs, fewer unique generic locations, and covers slightly more ground than does the average female-gendered protagonist. But, surprisingly, the difference in distance traveled is not statistically significant either in aggregate or within the individual subcorpora.

Social Effects on Mobility. Focusing specifically on the contemporary data, we measure

1. Median values of these quantities are lower, since their distributions include a long tail of large values, but the observed historical trends and relationships between subcorpora do not differ meaningfully under that metric. The same is true of the total (as opposed to unique) number of GPEs and generic location mentions. Full results are available in the supplementary material.

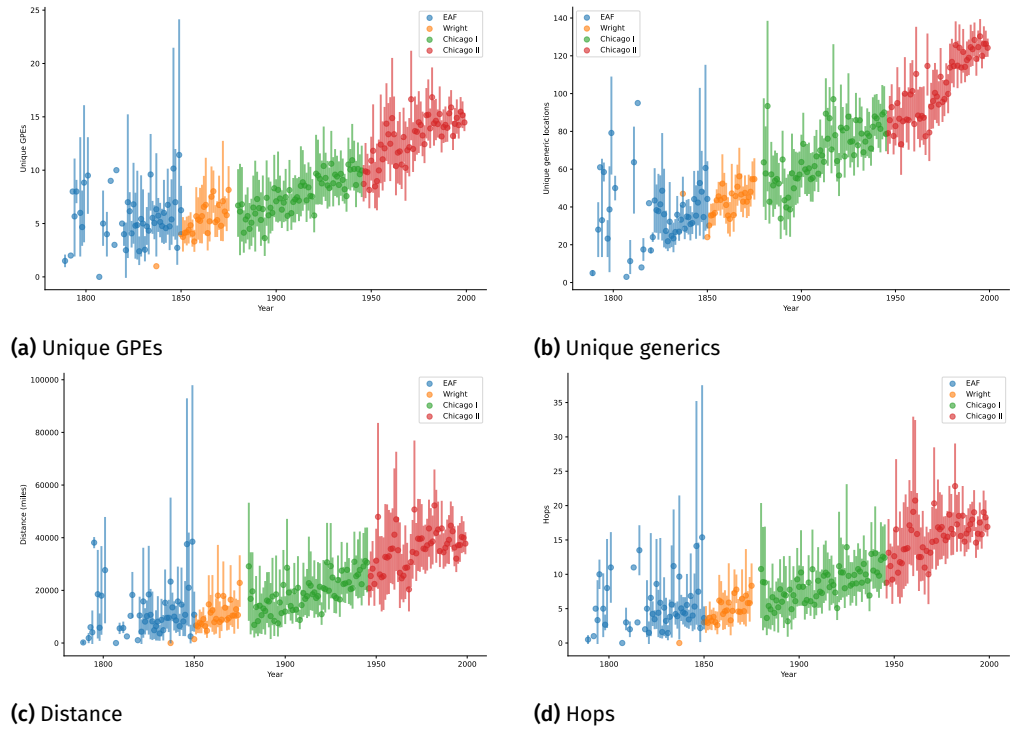


Figure 1: Unique GPEs, unique generic locations, protagonist distance, and hop count over time by subcorpus and year. Markers represent yearly means; bars are 95% confidence intervals.

GPE	Most frequent hops
New York	America, Paris, Manhattan, London, New York City
London	New York, England, Paris, America, France
America	New York, London, England, California, India
Paris	France, New York, London, Chicago, England
California	New York, Los Angeles, San Francisco, America, Chicago
Generics	Most frequent hops
room	house, home, kitchen, bedroom, place
house	room, home, kitchen, place, living room
home	house, room, kitchen, school, place
kitchen	house, room, home, living room, bedroom
place	room, house, home, apartment, kitchen

Table 3: Most frequent inhabited locations in the fiction facet of CONLIT followed by the most frequent subsequent locations (“hop”) in descending order of frequency.

Feature	she/her	he/him	<i>p</i>
Distance	29,943	31,134	0.1990
GPEs	11.08	11.85	0.0008 ***
Generics	102.0	95.8	0.0008 ***

Table 4: Key mobility metrics by narrativized character gender in fiction in the full corpus.

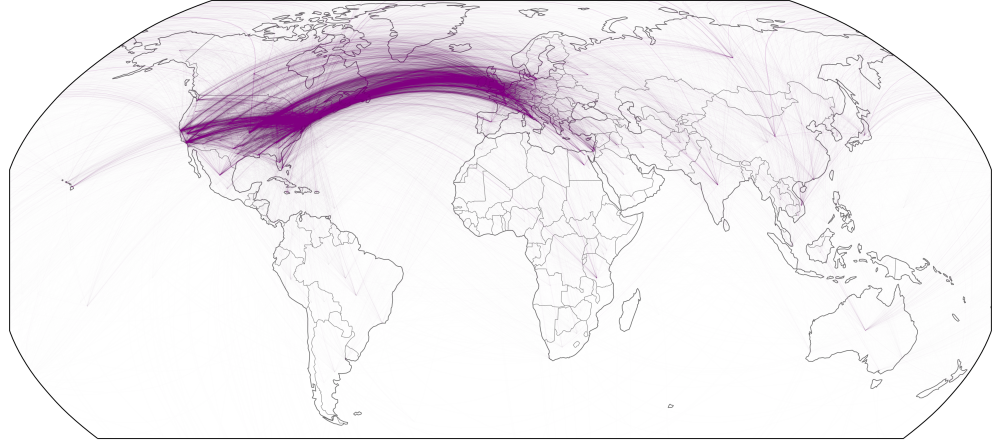


Figure 2: Aggregated character hops in the corpus. Line widths are proportional to the total number of hops between each pair of locations.

the effects of different social categories on character mobility using the regression models described above. As shown in table 5, we find that both fictionality and age-level have the strongest negative association with mobility, i.e., both categories significantly lower the distance traveled and the frequency of place names mentioned (both GPE and generic). We also observe a greater reliance on generic place names in both of these categories. Finally, as with the full corpus, we find that, after controlling for genre-related factors, there is no meaningful difference in the distance traveled between differently gendered characters.

Measure	Fictionality		Prestige		Youth		Female	
	valence	<i>p</i>	valence	<i>p</i>	valence	<i>p</i>	valence	<i>p</i>
Distance	-	***	+	.	-	***	+	.
GPEs	-	***	-	.	-	***	+	.
Generics	-	***	+	.	-	***	+	***
Semantic Dist	-	*	+	***	+	.	-	**
Deictics	+	***	-	***	+	.	-	.
Generic/GPE Ratio	+	***	+	.	+	***	+	.

Table 5: Results of regression analysis for each measure across our primary categories in the CONLIT corpus. Valence captures whether the estimate for the primary category (e.g. Fictionality) is lower than or higher than its opposite (e.g. Non-Fictionality). We provide standard significance codes (*** < 0.001, ** < 0.01, * < 0.05, . > 0.05). Full results, including the estimates and R^2 values, are supplied in the supplementary material.

In addition to our regression analysis, we also seek to identify ways in which mobility may differ *qualitatively* even when overall quantitative levels are similar. We employ the Fightin’ Words method of Monroe et al. 2017 to identify GPEs and generic places that are over- and underrepresented in facets of our corpus (figure 3).²

We observe that contemporary fictional narratives are often enriched in imaginary, extraterrestrial, historical, and otherwise “peripheral” GPEs (Maine, Taos, Sri Lanka) relative to nonfictional narratives, which are themselves enriched in sites of political power and armed conflict. Fiction is also enriched in generic locations that are private

2. Specifically, we use the method described in Monroe et al. 2017, section 3.5.1, equation 23, with an informative Dirichlet prior calculated over all volumes in the corpus.

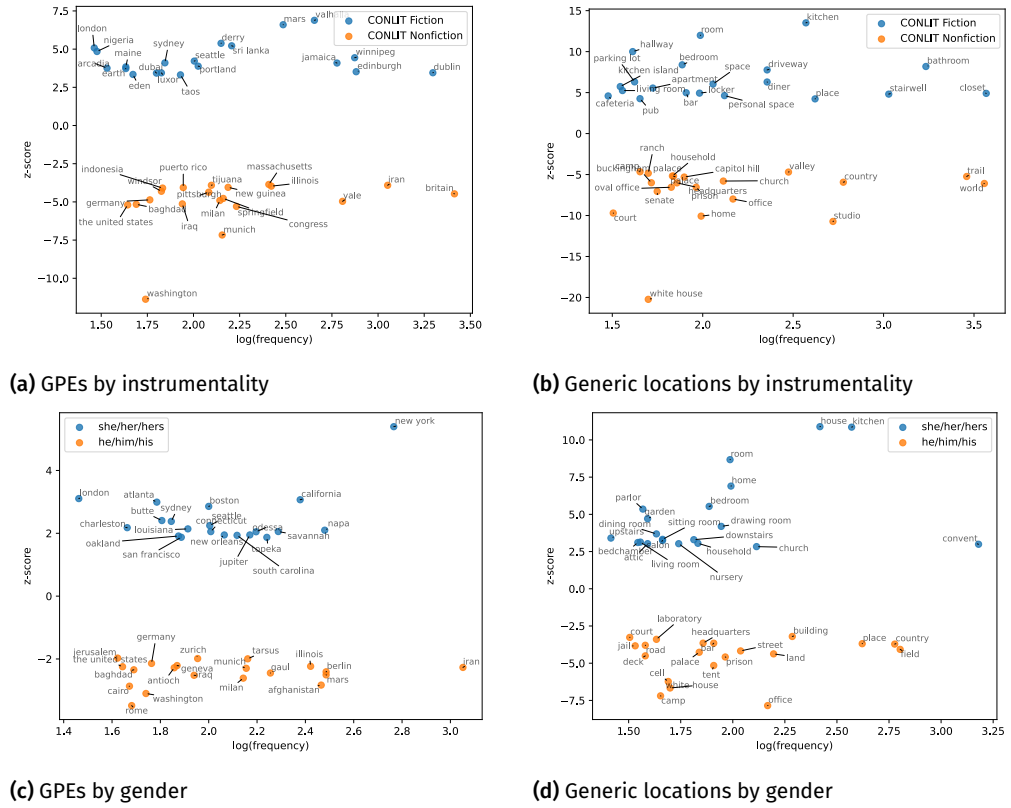


Figure 3: Distinctive location use across instrumentality and character gender facets in CONLIT. The x -axis represents the log of the frequency of each term in the indicated corpus; the y -axis represents the z -score of the term in the indicated facet relative to the other facet, informed by a weighted prior calculated over the full corpus.

and semi-public interior spaces, whereas nonfiction preferentially locates its characters in public sites of power and work.

Within fiction, we find that she/her characters are distinctively located in major and evocative urban localities; he/him characters are assigned preferentially to historical and contemporary sites of power and to those of American political and armed conflict. Generic locations are distributed by gender in ways that resemble their allocation between fiction and nonfiction, she/her characters occupying domestic interiors, he/him characters disproportionately found in public, power-infused sites.

4. Discussion

Our results paint a clear picture of the spatial constraints of fictional worlds. When compared with non-fictional narratives, characters in contemporary fiction travel less distance, visit fewer geographic and generic places, inhabit generic places that are semantically more similar to each other, and rely far more on generic places than on geographic ones. They also utilize deictic markers like “here” and “there” with far greater frequency. Fictional worlds are smaller worlds, both geographically and semantically.

Interestingly we see little effect on these measures if we look at social categories like prestige or gender. Prizewinning novels do not travel further or utilize more geographic places when compared to more market-driven fiction. They do tend to use fewer deictics

and employ more semantic diversity among non-geographic places, suggesting greater sophistication at the vocabulary level. Books aimed at middle-school audiences generally use far more limited narrative worlds, as would be expected.

The results surrounding character gender are surprising, given our assumptions that she/her characters would more likely be associated with social constraints that could effect their mobility. This turns out not to be the case. For both the historical and contemporary data, women were no more likely to be associated with diminished levels of mobility.

At the same time, when we examine the distinctive places associated with she/her characters, we do see more expected outcomes. She/her characters are more likely than he/him characters to be associated with domestic, private, and semi-public spaces. If we compare the results for fiction/nonfiction presented in figures 3a and 3b to those for character gender in figures 3c and 3d, we see how the locations distinctively occupied by she/her and he/him characters map closely to those of fiction and nonfiction protagonists, respectively. While we are not yet in a position to assert a blanket spatial homology between fictionality and gender, the resemblance is sufficiently persuasive to merit further investigation.

In addition to these small-world effects at the level of geodetic distance, we also find that the *connections* between geographic places in fictional worlds are remarkably predictable (figure 2). Fictional worlds are “small” not just in the sense of the overall distance characters travel, but also in the diversity of places they move between. There is a NATO- or grand-tour-driven center surrounded by a much less traveled periphery. Fictional characters spend their time moving between a very small portion of the world.

These results accord well with previous work that examined the distribution of named locations (without regard to character associations) in British and American fiction (Wilkens 2016), though there exists some evidence suggesting that British fiction underwent greater evolution of its geographic imagination over the twentieth century than did American (Wilkens 2021). Future work could begin to replicate these methods for more geographically diverse fiction produced around the world to model the spatial archetypes of mobility. Does every region have its spatial center of gravity and its exotic periphery? To what extent are centers and peripheries shared across nations, languages, and periods? Is every regional literature as constrained as the North American example or do other regions have very different network structures of mobility?

When it comes to changes to mobility over historical time, we see that the distance traveled by fictional characters has been increasing, as have the number of GPEs and generic places. One of the drivers of this is that fictional narratives have also been getting longer over time and the frequency of references to the main character has been increasing as well.³ If we normalize by book length, we still see meaningful increases over time; if we normalize by character count (that is, by the number of all character references that pertain to the protagonist), we see slower growth in distance traveled and essentially zero rise in the count of unique GPEs (figure 4). The same is true when we compare highly protagonist-centered first-person narratives to more widely

3. We note in passing that these measures of average book length and protagonist concentration over nearly 250 years of North American literature are novel in the critical and computational literature. They likely merit future investigation.

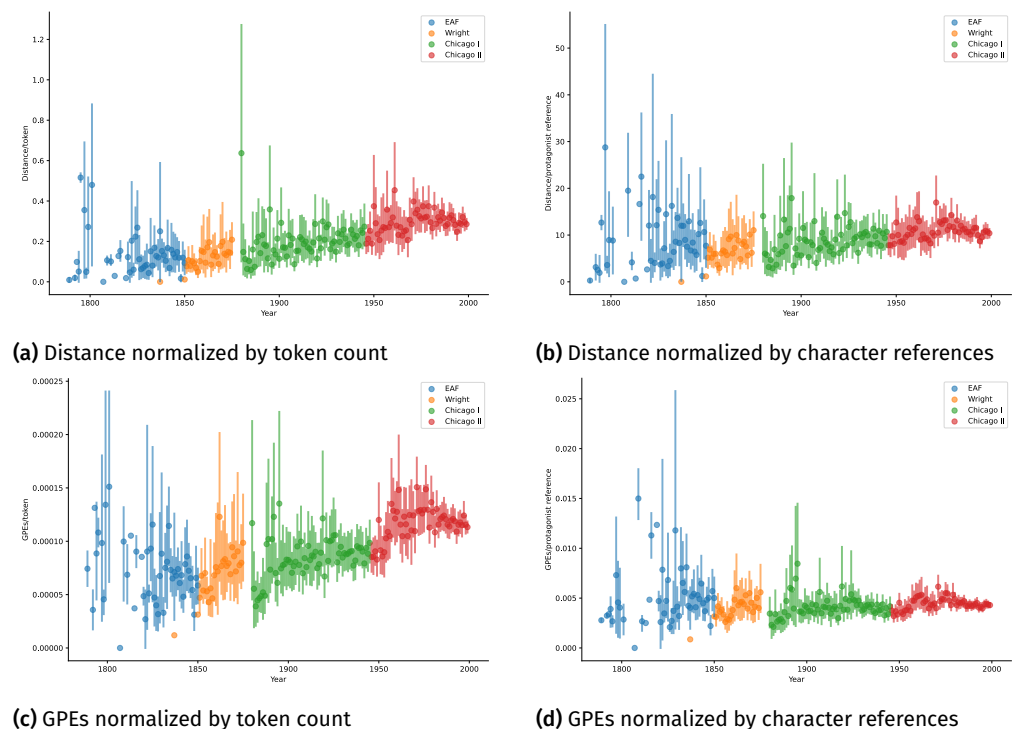


Figure 4: Average fictional protagonist distance and count of unique GPEs by year and subcorpus, normalized by volume length or by count of character references.

character-dispersed third-person alternatives. What this tells us is that, as books have 320
 become longer and more protagonist-centered, main characters are traveling relatively 321
 further and moving between geographic places more often, but much of this growth 322
 can be accounted for by the sheer increase in character references (allowing for more 323
 places to be counted and thus more distance to be traveled). There does not appear to 324
 be an obvious ceiling on the range or rate of protagonist mobility, even in long books 325
 with potentially saturated story worlds. 326

The final way in which we understand the small-world effect of fiction is through our 327
 examination of the lexical differences between spatial entities in fiction when compared 328
 with non-fiction (figure 3). When we do so, we quickly confirm several differences 329
 that we might have expected, but have not previously quantified. Compared to fiction, 330
 nonfictional narratives overrepresent sites of power, including official political locations 331
 like White House, Oval Office, Senate, Washington, Buckingham Palace (and “palace” 332
 generically), and Capitol Hill; sites of carceral power (court, prison); workplaces (studio, 333
 office, headquarters); and locations of present and historical conflict as experienced 334
 primarily from the United States (Baghdad, Iraq, Iran, Munich, Yijuana). Fiction, by 335
 contrast, overrepresents domestic and semi-public spaces (kitchen, hallway, bedroom, 336
 bathroom, apartment, cafeteria, pub, and many more), driveways, and parking lots. As 337
 has long been theorized, fiction is preëminently occupied with domestic and private 338
 space (Armstrong 1987; McKeon 2006). 339

On the other hand, the distinctive geographic spaces of fiction are often extremely distant 340
 or otherworldly (Valhalla, Mars, Arcadia, Eden). Fiction compensates for its small- 341
 world effects – either in the real-world or through generic private spaces – by investing 342

at least partially in telling narratives focused on the most distant places imaginable.⁴ 343
It is worth considering what a new genre of fiction might look like that inverted this 344
escapism-power dynamic and focused instead on immersing readers in the central 345
locales of power and punishment rather than the private chambers of imaginary locales. 346

The major limitation of our study, beyond the need for cultural expansion, is that our 347
models cannot account for distances between unreal places or extraterrestrial locations, 348
which are identified by our entity model, but are not easily localizable in terrestrial 349
space. As we mention above, one could argue that the role of genres like fantasy and 350
science fiction is precisely to undo the small-world effects of fiction (Dubourg and 351
Baumard 2022). In simulating vast travel, they reverse the constraints of fictionality. 352
At the same time, the fact that we see these genres still exhibiting lower diversity of 353
generic places and higher semantic constraints between them relative to nonfictional 354
narratives suggests a basic conflict between the expansiveness of space on the one hand 355
("to the moon and back") and the constraints of fictional places that are limited to 356
rooms, vehicles, and home-like structures. 357

5. Conclusion 358

Our project has attempted to add two important methodological dimensions to prior 359
research on literary spaces. First, relying on new models that locate characters in space 360
(Soni et al. 2023), we are able to give a *character-centred* account of fictional spaces. 361
Second, by looking at the sequencing of space we are able to observe the effect that 362
narrative time has on the construction of space, for which we employ the construct of 363
"character mobility." 364

Applying our models to a large collection of historical and contemporary Anglophone 365
fiction, we make the following key observations concerning the small-world effects of 366
fiction: 367

1. **Fictional worlds are small in the sense of the distance traveled by characters.** 368
When compared to the movements of non-fictional characters (subjects of memoirs, 369
biography, or historical narratives), fictional protagonists travel less than half the 370
distance of their non-fictional counterparts. Generic places are also much more 371
relied upon and far more semantically similar to each other than is the case in 372
non-fiction. 373
2. **Fictional worlds are small in the constrained routes that characters travel.** Fic- 374
tional characters stick to a very familiar set of pathways that leave much of the 375
world un- or under-explored. 376
3. **Fictional worlds are semantically small in the types of generic spaces they** 377
foreground. Fictional characters are much more likely to be located in domestic 378
or private spaces when compared to their non-fictional counterparts. 379
4. **Fictional worlds have been expanding over historical time.** The distance traveled 380
by fictional characters has doubled since the nineteenth century, but much of this 381

4. We say at least partially because these are not the most common locations in contemporary fiction (which are all-too-familiar places like New York, London, and America). Rather, these are the locations that are present in fiction and that are extremely under-represented in works of non-fiction.

increase can be accounted for by the increased centralization of main characters.	382
5. She/her characters do not move less but they do spend more time in the kitchen.	383
Insights into the gendered nature of mobility reject assumptions about the spatial	384
limitations of women characters, but support their over-representation within	385
domestic spaces.	386
We look forward to continuing this work to gain a deeper and more culturally diverse	387
understanding of the relationship between fictional narratives and character mobility.	388

6. Data Availability 389

Data and supplementary materials are available at <https://anonymous.4open.science/r/small-worlds-A0B2/> 390
391

7. Acknowledgements 392

The authors thank Yasmine Chim for her assistance compiling validation data. The 393
research reported in this article was supported by funding from the National Science 394
Foundation (IIS-1942591, to DB) and the National Endowment for the Humanities 395
(HAA-271654-20, to DB; HAA-290374-23, to MW). 396

8. Author Contributions 397

Matthew Wilkens: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, validation, visualization, writing - original draft 398
399

Elizabeth F. Evans: formal analysis, writing – review & editing 400

Sandeep Soni: methods, data analysis, software 401

David Bamman: funding acquisition, methods, resources 402

Andrew Piper: conceptualization, data curation, formal analysis, project administration, investigation, writing – original draft 403
404

References 405

- Armstrong, Nancy (1987). *Desire and domestic fiction: A political history of the novel*. Oxford 406
University Press. 407
- Bakhtin, Mikhail Mikhailovich (2010). *The dialogic imagination: Four essays*. University of 408
Texas Press. 409
- Bamman, David (2020). “LitBank: Born-Literary Natural Language Processing”. In: 410
Computational Humanities. Ed. by Jessica Marie Johnson, David Mimno, and Lauren 411
Tilton. Debates in the Digital Humanities. 412
- (2021). *BookNLP. A natural language processing pipeline for books*. [https://github.com](https://github.com/booknlp/booknlp) 413
[/booknlp/booknlp](https://github.com/booknlp/booknlp). Accessed: 2022-01-30. 414
- Bodenhamer, David J, John Corrigan, and Trevor M Harris (2010). *The spatial humanities: 415
GIS and the future of humanities scholarship*. Indiana University Press. 416

- Bruner, Jerome (1991). "The narrative construction of reality". In: *Critical inquiry* 18.1, 1–21. 417 418
- Cresswell, Tim (2006). *On the move: Mobility in the modern western world*. Taylor & Francis. 419
- Digital Library Program (2012). *Wright American Fiction*. Tech. rep. Indiana University Libraries. <https://webapp1.dlib.indiana.edu/TEIgeneral/welcome.do?brand=wright>. 420 421 422
- Dubourg, Edgar and Nicolas Baumard (2022). "Why imaginary worlds? The psychological foundations and cultural evolution of fictions with imaginary worlds". In: *Behavioral and Brain Sciences* 45, e276. 423 424 425
- Electronic Text Center (2000). *Early American Fiction Collection*. Tech. rep. University of Virginia Library. <https://jti.lib.virginia.edu/eaf/>. 426 427
- Evans, Elizabeth F., ed. (2025). *Cambridge Critical Concepts: Space and Literary Studies*. Cambridge University Press. 428 429
- Evans, Elizabeth F. and Matthew Wilkens (2018). "Nation, Ethnicity, and the Geography of British Fiction, 1880-1940". In: *Journal of Cultural Analytics* 3.2. 10.22148/16.024. 430 431
- (2024). *Gender and Literary Geography*. Cambridge University Press. 432
- Friedman, Susan Stanford (1998). *Mappings: Feminism and the cultural geographies of encounter*. Princeton University Press. 433 434
- Herman, David (2009). *Basic elements of narrative*. John Wiley & Sons. 435
- Hones, Sheila (2022). *Literary geography*. Taylor & Francis. 436
- Kerouac, Jack (2002). *On the Road*. Penguin Classics. 437
- Long, Hoyt and Richard Jean So (2020). *US Novel Corpus*. Tech. rep. University of Chicago Textual Optics Lab. https://textual-optics-lab.uchicago.edu/us_novel_corpus. 438 439 440
- McKeon, Michael (2006). *The secret history of domesticity: Public, private, and the division of knowledge*. JHU Press. 441 442
- Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn (2017). "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict". In: *Political Analysis* 16.4, 372–403. 10.1093/pan/mpn018. 443 444 445
- Moretti, Franco (1999). *Atlas of the European novel: 1800-1900*. Verso. 446
- Pennington, Jeffrey, Richard Socher, and Christopher D Manning (2014). "Glove: Global vectors for word representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. 447 448 449
- Piatti, Barbara, Hans Rudolf Bär, Anne-Kathrin Reuschel, Lorenz Hurni, and William Cartwright (2009). "Mapping literature: Towards a geography of fiction". In: *Cartography and art*. Springer, 1–16. 450 451 452
- Piper, Andrew (2022). "The CONLIT dataset of contemporary literature". In: *Journal of Open Humanities Data* 8. 453 454
- Roberts, Les, Thomas Thevenin, Julia Hallam, Andrew Beveridge, Ruth Mostern, Humphrey Southall, Niall A. Cunningham, Robert M Schwartz, and Elijah Meeks (2014). *Toward spatial humanities: Historical GIS and spatial history*. Indiana University Press. 455 456 457
- Ryan, Marie-Laure, Kenneth Foote, and Maoz Azaryahu (2016). *Narrating space/spatializing narrative: Where narrative theory and geography meet*. The Ohio State University Press. 458 459 460
- Soni, Sandeep, Amanpreet Sihra, Elizabeth Evans, Matthew Wilkens, and David Baman (2023). "Grounding Characters and Places in Narrative Text". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long* 461 462 463

- Papers*). Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, 11723–11736. [10.18653/v1/2023.acl-long.655](#).
- Spearman, Charles [1904] (1987). “The Proof and Measurement of Association between Two Things”. In: *The American Journal of Psychology* 100.3/4, 441–471.
- Tally Jr, Robert (2012). *Spatiality*. Routledge.
- Wilkens, Matthew (2013). “The geographic imagination of Civil War-era American fiction”. In: *American Literary History* 25.4, 803–840.
- (2016). “The Perpetual Fifties of American Fiction”. In: *Neoliberalism and Contemporary Literary Culture*. Ed. by Mitchum Huehls and Rachel Greenwald-Smith. Baltimore: Johns Hopkins UP, 181–202.
- (2021). “‘Too isolated, too insular’: American Literature and the World”. In: *Journal of Cultural Analytics* 6.3. [10.22148/001c.25273](#).
- Wright, Lyle Henry (1965). *American Fiction, 1851-1875: A Contribution toward a Bibliography*. Revised. The Huntington Library.