LIWC can thus give us an initial range of intuitively meaningful interpretive categories to build on as well as the lexica upon which those categories may in part be based. They should not be taken at face value, but looked into, as with all semantic fields. They offer an alternative to the derived linguistic fields that I explored in the previous chapter. No one method suits all questions. Because the dictionaries are transparent in LIWC, users can refine or alter the dictionaries as they see fit, as I have done on occasion here. They can also be combined with other, more customized, features, as I will be doing in the chapters that follow. Building more sophisticated feature sets is ultimately one of the central research challenges of the field.

The Coherence of Fictionality

The question that I want to begin with is: How coherent is fiction as a type of writing? Are there indeed no syntactic or semantic properties, as Searle contends, that allow us to predict whether something is intended fictionally? Is fictionality exclusively a function of communicative context, the intentionality of the writer, and the belief system of the reader? Or are there features that appear with a high degree of regularity in fictional texts that do not appear in nonfiction, such that even a computer can make accurate guesses as to the nature of the text?

In order to answer these questions, I will use the process known as machine learning to see how accurately a computer can predict a text's given class (I will be using the learning algorithm known as a support vector machine [SVM], which is applied in many text classification scenarios).[24] For those not familiar with this process, a learning algorithm is "trained" on features found in a set of documents for which the classes are already known and then asked to predict which class a group of texts belong to that it has not seen. In this case, I train the algorithm using the LIWC features discovered in a given set of documents and use a process of tenfold cross-validation to make predictions on whether a document is a work of fiction or nonfiction. What this means in practice is that I randomly divide the corpus into a 90–10 split ten times, where 90% of the documents are used to train the algorithm and the unseen 10% are used to test its reliability. (The folds function in the kernlab package ensures that the folds are equally divided between the two categories relative to the overall data.) Doing this ten times allows us to gain a full view of all the documents in the collection, as each document has an opportunity to be in the test set.

n initial range of intuitively meaningful in-
d on as well as the lexica upon which those
sed. They should not be taken at face value,
semantic fields. They offer an alternative to
that I explored in the previous chapter. No
ons. Because the dictionaries are transpar-
ıe or alter the dictionaries as they see fit, as
re. They can also be combined with other,
s I will be doing in the chapters that follow.
feature sets is ultimately one of the central
ıld.

ality

begin with is: How coherent is fiction as a
ıdeed no syntactic or semantic properties,
w us to predict whether something is in-
ılity exclusively a function of communica-
ty of the writer, and the belief system of the
hat appear with a high degree of regularity
ıppear in nonfiction, such that even a com-
ses as to the nature of the text?
questions, I will use the process known as
accurately a computer can predict a text's
ıe learning algorithm known as a support
ı is applied in many text classification sce-
ar with this process, a learning algorithm is
n a set of documents for which the classes
asked to predict which class a group of
ot seen. In this case, I train the algorithm
vered in a given set of documents and use
dation to make predictions on whether a
or nonfiction. What this means in practice
corpus into a 90–10 split ten times, where
ed to train the algorithm and the unseen
ıility. (The folds function in the kernlab
are equally divided between the two cat-
data.) Doing this ten times allows us to
ments in the collection, as each document
e test set.

Table 4.2 Classification results for predicting fictional texts using tenfold cross-validation

Corpus1	Corpus2	Avg. accuracy (F1)	No. docs
Fiction (EN_FIC)	Nonfiction (EN_NON)	0.94	100/100
English Novel (EN_NOV)	Nonfiction (EN_NON)	0.96	100/100
German Novel (DE_NOV)	Nonfiction (DE_NON)	0.95	100/100
English Novel 3P (EN_NOV_3P)	History (EN_HIST)	0.99	95/86
Germ Novel 3P (DE_NOV_3P)	History (DE_HIST)	0.99	88/75
Cont. Novel (CONT_NOV)	Nonfiction (CONT_NON)	0.96	193/200
Cont. Novel 3P (CONT_NOV_3P)	History (CONT_HIST)	0.99	210/200
19C Fiction (HATHI) (Trained)	Cont. Novel (CONT) (Tested)	0.91	21,158/400

Table 4.2 presents the results of this experiment, showing which two data sets were compared and the average accuracy of the predictions on the unseen data. As we can see, not only are the differences between fiction and nonfiction robust across time and languages, but we can use models built in one time period to strongly predict those of another. In the last line, I train a model on nineteenth-century fiction and nonfiction from the Hathi data set and then test it on the contemporary novels and nonfiction. While there is a clear drop in performance when we use nineteenth-century models to predict twenty-first-century novels, we can still see a relatively high degree of performance at work here (around 91% accuracy). There appears to be a notable degree of diachronic stability to fictional discourse over the past two centuries. Indeed, as we will see below, when we examine features that are more indicative of novelistic writing in particular as a subset of fictional discourse, we generally see these features increase over time. The trans-temporal stability of the novel is complemented by an increase in certain types of novel-specific vocabulary that can be traced back to the nineteenth century.

The Phenomenology of the Novel

If fiction is so predictable, what are the features that make it so? Here I use a rank-sum test to examine which features are statistically distinctive of one group compared to another by observing the overall distributions

of each feature in the two groups.[25] The more those distributions differ from one another, the stronger the statistical significance. I then rank the features by the ratios of the median values for each group.[26] The value of doing so is that it preserves information about the overall distribution of a feature in a given population (rather than being driven by a few texts that might have a particular feature in a much higher amount). The disadvantage is that it does not correct for low-occurring features—small features that change considerably will look more important than highly common features that only change slightly. This may not be a disadvantage depending on what one's assumptions are, but it is important to understand the way this assumption is built into the rankings.[27]

To begin, let me review the overall structure of the tables used here to better understand what they can tell us (see table 4.3 as an example).[28] The leftmost column ("Feature") lists the features as defined by LIWC. Some are extremely straightforward ("exclam" refers to the percentage of exclamation marks), while others are more nuanced. "Family," for example, refers to a dictionary of words all related to family members, while "social" relates to words having to do with social experience, which can include pronouns (a choice that effectively duplicates the pronoun categories because they are so much more common than other words). The former is arguably much more straightforward than the latter, and thus we need to be cautious when we encounter a dictionary that is more semantically ambiguous (though even a single word like "you" may have different kinds of functions in novels). The second column ("Category") lists the category to which the feature belongs, a slightly more general framework for understanding the individual features. The next two columns (Fiction %, Nonfiction %) present the median frequency of that feature in each corpus as a percentage of all words. This allows us to see which features are more prevalent relative to other features.

Because percentages are somewhat opaque in terms of a reader's experience, I will generally be translating these numbers into page and work equivalents in the discussion that follows. This allows us to imagine our way into a reader's experience and surmise which features occupy more of a reader's attention. Exclamation marks, for example, comprise on average about 0.45% of a given work of fiction in the nineteenth century. If we assume an average novel length of about 100,000 words (or 500 words per page across 200 pages), this means that there is one exclamation mark for about every 200 words, or 2–3 per page, or roughly 500 total per novel. Personal pronouns, on the other hand, occur about 10% of the time in fiction, which means once every 10 words, or *50 times* per page (and 10,000 times per novel).

Table 4.3 The top ten features with the greatest increase in fiction compared to nonfiction. Values represent median percentages in the nineteenth-century canon collection.

		Fiction vs. Nonfiction 19C canon (English)				
Feature	Category	Fiction (%)	Nonfiction (%)	Ratio	Sample rank	Hathi rank
exclam	linguistic	0.40	0.04	10.00	1	2
you	linguistic	1.34	0.16	8.34	2	1
q-mark	linguistic	0.41	0.08	5.13	3	6
I	linguistic	2.41	0.49	4.92	4	7
quote	linguistic	2.59	0.65	3.98	5	5
assent	social	0.12	0.03	3.83	6	4
family	social	0.56	0.16	3.58	7	10
hear	perception	1.15	0.38	3.01	8	9
shehe	linguistic	4.86	1.90	2.56	9	8
ppron	linguistic	10.73	5.01	2.14	10	14

p-value < 0.0001

The fifth column, "Ratio," lets us see how much more prevalent the feature is in one collection over another by comparing the ratio of its median value in the two collections. Exclamation marks appear almost ten times as often in fiction as in nonfiction. This is a massive difference, but we are still only talking about something that occurs relatively infrequently compared to other features. Personal pronouns, on the other hand, only appear a little more than twice as often in fiction (still a very large difference), but this increase is based on a much larger linguistic aspect of texts. Twice as many pronouns means roughly 5,000 more pronouns per work, or about 25 more *per page*. While I privilege ratio here in my interpretation of the results, we will want to keep our eye on both of these aspects, from the overall prevalence of the feature to the relative increase from one population to another.

Finally, the columns "Sample rank" and "Hathi rank" refer to the respective rankings of a given feature in the small canonical data set of about 100 novels in English versus the much larger collection of over 9,000 documents in the Hathi Trust collection. The idea here is to try to understand the extent to which the smaller sample serves as a decent approximation of the much larger set.

As I proceed, I will be translating these tables into word clouds in order to render the information in a more digestible way. Word size refers to the ratio value, where larger equals a **higher** prevalence in fiction, while color (words in black) refers to a **particular** category of interest.

4.2 Distinctive features for nineteenth-century fiction. Words in black correspond to linguistic categories such as punctuation, pronouns, and verb tense.

Several of the tables are included in the appendix, and all are available in the supplementary data for further review.

Beginning with the baseline comparison of fiction and nonfiction writing using both our canonical sample and the larger collection of Hathi Trust writings from the nineteenth century (fig. 4.2), we see how the features that are most indicative of fictionality are driven by dialogue—exclamation marks, question marks, quotation marks, first- and second-person pronouns like *I* and *you*, assent words like *yes*, *okay*, and *oh*, and finally the word *said* (which is labeled as an auditory verb by LIWC). Importantly, we also see very strong alignment between the nineteenth-century sample and the larger population of Hathi Trust documents, with some notable exceptions around the "social" category and potentially *family*, *home*, and *ingestion*. If we compare these groups directly, we see that only *family* and *ingestion* are somewhat inflated in the canonical sample (by about 10–15%).[29] In other words, while there are interesting variations that are worth exploring, on the whole, the smaller sample does a good job of capturing the same information as the larger collection.

Taken together, these features suggest a relatively unambiguous way in which fictional writing has a uniquely dialogical construction when compared with nonfiction. While this may not be "news," it does help us build a taxonomy of the distinctions that make this kind of writing socially significant. Imagining people talking to each other appears to be one of fiction's primary cultural functions.

' fiction. Words in black correspond to linguistic
, and verb tense.

ı the appendix, and all are available
er review.
mparison of fiction and nonfiction
mple and the larger collection of Ha-
th century (fig. 4.2), we see how the
fictionality are driven by dialogue—
quotation marks, first- and second-
ent words like *yes*, *okay*, and *oh*, and
d as an auditory verb by LIWC). Im-
ılignment between the nineteenth-
ıtion of Hathi Trust documents, with
'social" category and potentially *fam-*
re these groups directly, we see that
ıat inflated in the canonical sample
while there are interesting variations
e, the smaller sample does a good job
the larger collection.
ggest a relatively unambiguous way
quely dialogical construction when
ıis may not be "news," it does help
ions that make this kind of writing
e talking to each other appears to be
ctions.

Indeed, imagining people *as people* may be fiction's most important role. If we remove dialogue from the sets above, including the pronominal expressions that accompany them (she said, he cried, etc.), we see how third-person pronouns emerge as one of the strongest indicators of fictionality, along with references to family members and bodies (fig. 4.3, table A.6).[30] There is over a threefold increase in the average number of she/he pronouns in fiction versus nonfiction outside of dialogue, with just these two words alone accounting for more than 5% of all words in the text (or roughly 5,000 instances for a medium-length novel).

This is especially remarkable considering that on average, works of history, for example, use considerably more proper names than works of fiction (an estimated more than twice as many).[31] The lower number of people in fiction is compensated for by a more expanded durational existence on the page for which pronouns become key linguistic markers. People seem to have more extended identities in fiction, though this is not necessarily to be confused with a more "expansive" identity, that is, one that is more semantically rich. The pronominal frequency of characters is not the same as the linguistic diversity surrounding these characters (a point to which I will turn in my next chapter). Nevertheless,

4.3 Distinctive features for nineteenth-century fiction with dialogue removed. Words in black correspond to social and biological categories.

this gives us a first indication of the ways in which fiction performs the process of identification as a repetitive and extensive act of naming the same person.

The prevalence of family and friend vocabulary in fiction also suggests what type of people are more distinctive of the genre, just as the setting of home gives us an idea of where they are most active. Broadly speaking, when we read fiction in the nineteenth century, what is novel, that is, different from other kinds of texts that purport to be about real things, is a focus on family and the familiar. Travel, adventure, work—these can be experienced elsewhere in ways that documentation of family life and the extended agency of individuals cannot. Family is the dominant social imaginary of nineteenth-century imaginative writing.

The stakes of this attention will become even clearer when we focus on a particular type of fiction (novels with external narrators) and a particular type of nonfiction (history writing) (fig. 4.4, table A.7). What rises to the top here (seen in the top cloud) are a host of perceptual categories (seeing, hearing, feeling) that construct the phenomenological reality of an experiencing individual (all more than three times more likely in fiction in the nineteenth century).[32] And the greater prevalence of body words (again about 2.5 times higher in the nineteenth century) gives us an indication of where that attention most often lies. It is knowledge, not just of otherness, but of another *embodied* individual, that most consistently frames the epistemological horizon of the novel from a quantitative point of view.

Interestingly, when we look at the German and contemporary data sets, we see some slight nuances to this story (fig. 4.4, middle and bottom).[33] Without being able to reliably remove dialogue from the German texts, those dialogical markers of pronouns and punctuation marks dominate, but just beneath these are once again the body and sense perception words at work (with some more emphasis on affect than in the English corpus). In the contemporary novels, we see how embodiment has become as important as sense perception—it is now over 4.5 times more likely to be present in fiction than in history. In the contemporary novel, there is an even stronger focalization effect taking place between sensory experience and the observed human body.

These results pose an interesting challenge for "theory of mind" approaches that argue that fiction's primary purpose is the enactment of another human consciousness.[34] While we will see an area where this hypothesis does make sense in the next test, in terms of understanding the novel's distinctiveness compared to nonfictional writing, the mind-body distinction that underlies theory of mind models does not hold up

e ways in which fiction performs the
tive and extensive act of naming the

end vocabulary in fiction also suggests
ıctive of the genre, just as the setting
ney are most active. Broadly speaking,
eenth century, what is novel, that is,
that purport to be about real things,
iliar. Travel, adventure, work—these
ays that documentation of family life
duals cannot. Family is the dominant
tury imaginative writing.
become even clearer when we focus
els with external narrators) and a par-
writing) (fig. 4.4, table A.7). What rises
oud) are a host of perceptual catego-
onstruct the phenomenological reality
more than three times more likely in
[2] And the greater prevalence of body
r in the nineteenth century) gives us
n most often lies. It is knowledge, not
nbodied individual, that most consis-
ıorizon of the novel from a quantita-

the German and contemporary data
this story (fig. 4.4, middle and bot-
ably remove dialogue from the Ger-
of pronouns and punctuation marks
e once again the body and sense per-
more emphasis on affect than in the
ary novels, we see how embodiment
perception—it is now over 4.5 times
than in history. In the contemporary
alization effect taking place between
ed human body.
g challenge for "theory of mind" ap-
primary purpose is the enactment of
While we will see an area where this
next test, in terms of understanding
ed to nonfictional writing, the mind-
ory of mind models does not hold up

4.4 Distinctive features for third-person novels compared to histories for nineteenth-century English (top), nineteenth-century German (middle), and contemporary English (bottom). Words in black correspond to the categories of sense perception and embodiment.

4.5 Distinctive features for nineteenth-century novels compared to other fiction from the period. Words in black correspond to words related to cognitive processes.

well in light of the novel's strong emphasis on sensorial input and embodied entities. The sensual experience of a sensing being: this is what appears to be uniquely reiterated in the imaginative work of novelistic writing when compared to the nonfiction of historical writing.

As a final way to understand, and bring into sharper relief, the significance of what I am calling the phenomenological orientation of the novel, I compare the nineteenth-century novel with a particular subset of fiction that excludes novels published during the same time period (fig. 4.5, table A.8). Non-novelistic fiction in this case refers to a broad mixture of fictional writing that would have been very present to nineteenth-century readers, including classical epics translated into prose (*The Iliad*, *Odyssey*, *Edda*, *Nibelungenlied*), classic works of prose fiction (*The Tale of Genji*, *The Decameron*, King Arthur tales, and Rabelais), fairy tale collections from around the world (drawn from Irish, German, Danish, Japanese, and Indian sources), contemporary novella collections (novellas by Hoffmann, Tolstoy, Dickens, Maupassant, Hawthorne, and Washington Irving), as well as a variety of "tales" collections (*Tales of Former Times*, *Tales of Domestic Life*, *Moral Tales*). This data set is meant to represent a range of prose fiction that would have been widely read

tury novels compared to other fiction from the words related to cognitive processes.

g emphasis on sensorial input and em-
erience of a sensing being: this is what
l in the imaginative work of novelistic
onfiction of historical writing.

and bring into sharper relief, the sig-
e phenomenological orientation of the
n-century novel with a particular sub-
ls published during the same time pe-
ovelistic fiction in this case refers to a
ng that would have been very present
ncluding classical epics translated into
Nibelungenlied), classic works of prose
ameron, King Arthur tales, and Rabelais),
l the world (drawn from Irish, German,
urces), contemporary novella collections
Dickens, Maupassant, Hawthorne, and
a variety of "tales" collections (*Tales of*
Life, *Moral Tales*). This data set is meant
tion that would have been widely read

and known to nineteenth-century Anglophone readers but would not have been considered a "novel." While the material dates from different epochs, the publications (and translations) are all contemporaneous with the period as a whole.

Three interesting features initially stand out. First, the ratios are much lower when compared with nonfiction (ranging from 65% to as low as 10% increases, as can be seen in table A.8, which is far below the 200–400% increases we were seeing above). While these groups are similarly well differentiated when compared to nonfiction, when compared to each other the overall distinctiveness drops considerably. If we run the same classifier as above, we can predict novels with about 68% accuracy, which is just below the threshold of statistical significance ($p = 0.018$). If we use a slightly larger collection of novels from the Hathi Trust collection (428 to mirror "other fiction"), accuracy will increase slightly, to 74% ($p = 7.23e\text{-}05$). While this is well above random, it is nevertheless considerably lower, for example, than the ability to predict novels from different genres. As Ted Underwood has shown, it is possible to predict detective fiction and science fiction across a 150-year span with between 88 and 90% accuracy.[35] And as I've shown elsewhere, using a similar-sized collection of contemporary novels, we can predict romances versus more general popular novels with about 98% accuracy, science fiction with about 87% accuracy, and mysteries with about 85% accuracy.[36] The broad category of "other fiction," then, is not as differentiated from novels as particular novel genres are from each other.

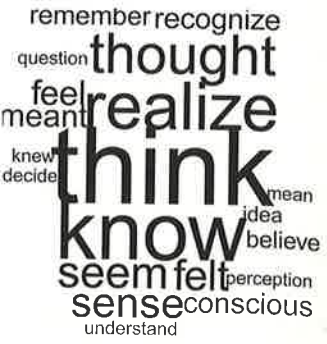
Second, while we see some of our more familiar linguistic fictional markers, such as pronouns and dialogue, we also see a new feature in the category of verbs. There are more verbs overall, as well as more varied tenses (past, future, and present, in addition to auxiliary verbs). In other words, there appears to be greater temporal complexity to novels than can be found in fiction more generally. While this deserves its own study, it suggests an initial insight into one of the key ways that novels differentiate themselves from other kinds of imaginary writing in the nineteenth century.[37]

Finally, we also see a new category emerge here that we have not seen before, one that falls under the heading of "cognitive process." These are the dictionaries that LIWC labels "discrepancy," "negation," "tentativeness," and "insight." If we examine the words in those dictionaries that are most distinctive of novels (and here I rank by log-likelihood ratio), we can see the extent to which these are words that tend to mark out moments of self-reflection, doubt, and hesitation—a kind of testing relationship to the world (fig. 4.6, table A.9).[38] It suggests that where fiction

4.6 Distinctive words in novels compared to other kinds of fiction across four categories (moving clockwise from the top left): "discrepancy," "insight," "negation," and "tentativeness."

overall invests in sensorial embodiment, the novel's signature is more oriented toward cognition.

Modal verbs in particular are extremely prevalent here (*could, would, must, might,* and *should,* as well as their negative contractions), and so too is the act of negation more generally (*don't, can't, didn't, not, never, nobody*). As the presence of "if" suggests, these groups offer different ways of expressing conditionality or even impossibility. At the same time, indefinite words such as *something, somebody, anything,* and *anybody* are more prevalent, along with a more specific vocabulary of hesitation (*perhaps, chance, hope, possibly, guess, maybe, doubt, uncertain*). In between the conditional and the impossible language of the novel, there lies a considerable amount of potentiality—chance, but also skepticism.[39]



inds of fiction across four categories
ıcy," "insight," "negation," and

t, the novel's signature is more

ely prevalent here (*could*, *would*,
egative contractions), and so too
't, *can't*, *didn't*, *not*, *never*, *nobody*).
roups offer different ways of ex-
llity. At the same time, indefinite
ing, and *anybody* are more preva-
ry of hesitation (*perhaps*, *chance*,
ain). In between the conditional
there lies a considerable amount
sm.[39]

Finally, we see how novels are marked by a much stronger use of mental states, captured in major verbs such as *know*, *feel*, *think*, *remember*, and *believe*, along with a second layer of less frequent, but similarly distinctive, complex cognitive verbs such as *admit*, *ponder*, *imagine*, and *forgive* (the latter not shown). This is the ground of the novel's reflectiveness, that which binds together doubt and conditionality into a consistent mental state. Indeed, the combination of *seem* and *feel*, both of which appear 30% more often in the novel, gives us a particular indication of what I am calling the novel's phenomenological orientation. Not the world itself, but a person's encounter with and reflection upon that world—the world's *feltness*—is what marks out the unique terrain of novelistic discourse when compared with other forms of classical fiction. It is this combination of sense perception plus cognitive skepticism that seems to bring out the novel's contribution to fictional discourse. The novel professes its uniqueness in the way it offers extended reading experiences of the human assessment of the world's givenness.[40]

The Great Reversal

I have over the course of this chapter tried to support three distinct arguments about the nature of fictional writing since the nineteenth century, with a particular emphasis on the novel. I call these arguments the coherence hypothesis, the immutability hypothesis, and the phenomenological hypothesis, respectively. As mentioned at the outset, I use the term *hypothesis* because these positions are still tentative. They need to be tested with different historical samples, on different kinds of subgenres, using different kinds of features and especially across more cultural spaces. As I have said repeatedly throughout this book, this is just the beginning.

But when we begin to look at the nature of fictional discourse from a quantitative perspective, there does appear to emerge a relatively clear story about its larger social function and the ways it distinguishes itself from purportedly "true" writing. Using the approach of machine learning, we are able to see how coherent fictional writing is when compared to a number of different kinds of nonfictional writing in different periods and in different languages. Seen in this way, and following other work in the field on the coherence of specific fictional genres,[41] notions of the indefiniteness or openness of literature look profoundly overstated. There is an underlying consistency or integrity to fictional discourse