DAVID L. ULIN AND JOSHUA COMER

By the Numbers

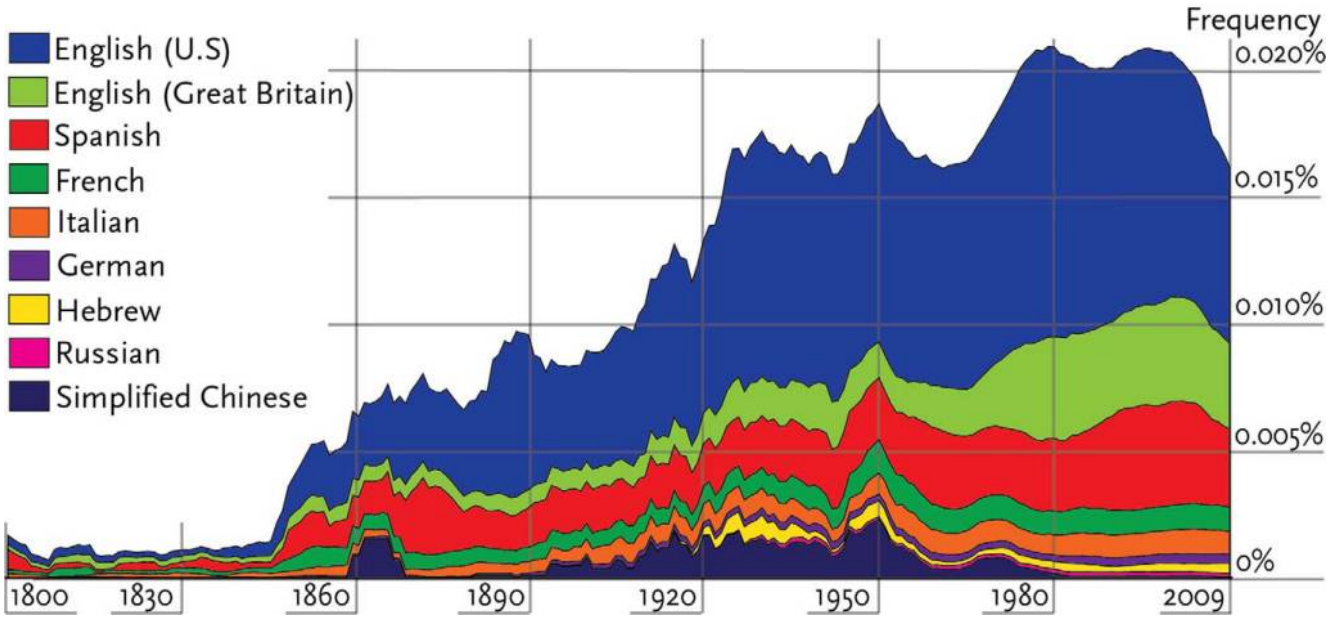
California in world literature

Editor's Note: How is California represented in world literature? There are certainly many qualitative answers to that question. But it is also possible to answer this question quantitatively by analyzing the millions of books digitized by Google in eight different languages. This represents an incredible corpus that can now be used to explore trends over time in words and ideas that have been published from 1500 to the present. We call this "world literature" in an expansive sense of the term because this corpus includes everything from atlases to government documents, poetry, and 'zines. Google makes available a database of all of the words in millions of these digitized publications in multiple languages at its Google Ngram Viewer website. Here we present two preliminary views of California in this corpus—one by a literary critic, the other by a digital humanities scholar—in an ongoing exploration of this question. You can also explore an interactive visualization of this data, our findings, and analysis, and join in the conversation about what it all means online at boomcalifornia.com.

David L. Ulin writes:

California owes its name to the written word. The source is the fictional Queen Califia, whose story comes from the Spanish writer Garci Rodríguez de Montalvo's 1510 romance *The Adventures of Esplandián*. "Know ye," Rodríguez de Montalvo wrote, "that at the right hand of the Indies there is an island called California, very close to that part of the Terrestrial Paradise, which was inhabited by black women without a single man among them, and they lived in the manner of Amazons. They were robust of body with strong passionate hearts and great virtue. The island itself is one of the wildest in the world on account of the bold and craggy rocks." Is it any wonder, then, that when

BOOM: The Journal of California, Vol. 4, Number 1, pps 46–53, ISSN 2153-8018, electronic ISSN 2153-764X.
 DOI: 10.1525/boom.2014.4.1.46.

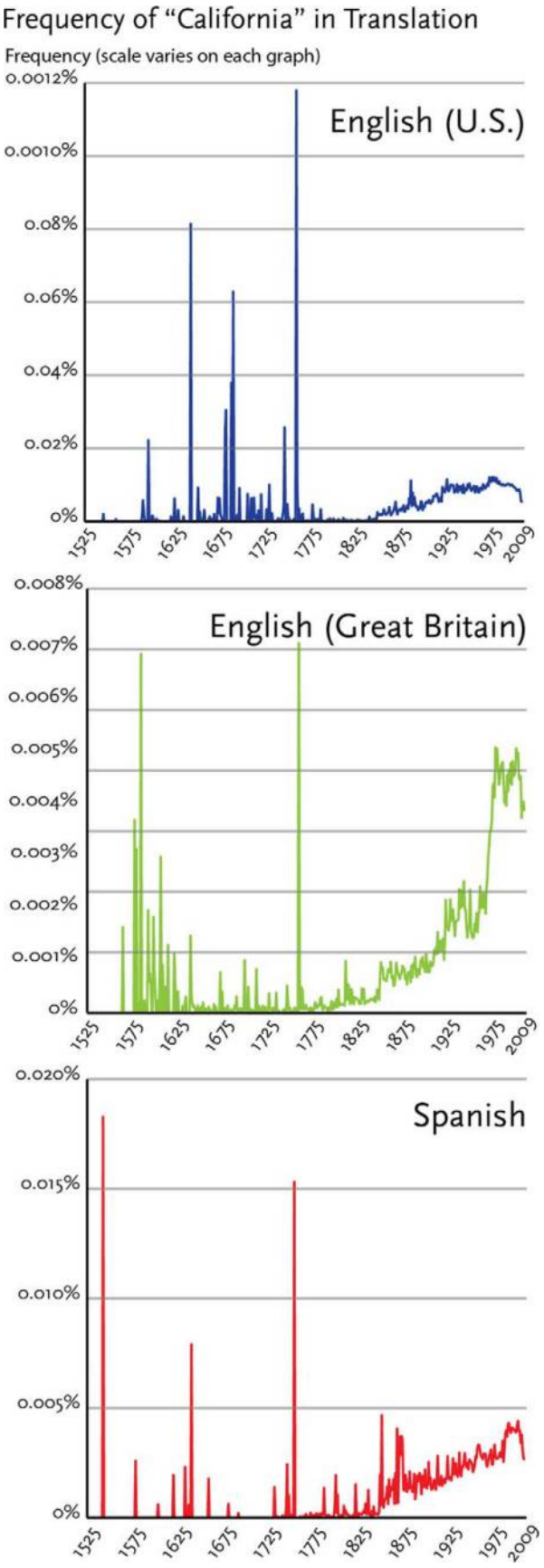


Diego de Becerra and Fortún Ximénez landed at the southern tip of Baja in 1533, they chose to name the place the Island of California, as if they had discovered their own heaven on earth?

I think about the Island of California when I look at Joshua Comer's data analysis graph. His task—to track the word "California" (and related phrases) through millions of books published across nine languages and several centuries—appears simple enough, but what it yields is something else again. To me, it looks like a voiceprint, or a series of overlapping voiceprints, the residue of a conversation we've been having without ever really calculating it, from continent to continent and year to year. It may start with Rodríguez de Montalvo, but it's the proliferation that's important . . . or, better yet, the cacophony.

Cacophony? Yes, the cacophony of California, which is itself made up of voiceprints, languages interrupting one another, each reading (and writing and speaking) the place through its own filter, its own point-of-view. Such an idea comes embedded in the very heart of Comer's research, which seems to address the state as both myth and landscape, manifest and historical destiny, demographic and promised land. I'm not even going to try to summarize his findings; to be honest, I don't think I could do them justice, and anyway, I'm less interested in the data than in the effect. Still, for all that his graphs reveal the fate of references to the state and some of its most essential tropes (the "California dream," for instance, or "Californian gold"), what they also do is suggest that this is just the beginning of the story, that we are looking at the expression of California as idea.

As to why this is important, California has always existed as part of both the real world (whatever that is) and the imagination, a territory we occupy and one we also dream. Read Comer's image one way and you get the former; read it another, and it's the latter that bleeds through. Just look at the spikes—the earliest right around 1850, denoting statehood and the Gold Rush, the next near 1870 (the completion of the continental railroad), and a third in the late 1880s and early 1890s, the time of the first great Los Angeles real-estate boom. This was when the "California Dream" was invented, although its substantiality has long been a subject of debate. After the boom imploded, Carey McWilliams once noted, sixty-two out of "more than a hundred towns platted in Los Angeles County" ceased to exist, if they had ever: "The town of Carlton had 4,060 lots and not a single resident; Nadeau had 4,470 lots but no settlers; Manchester had 2,304 lots, but no inhabitants; Santiago had 2,110 lots,



a few houses, but no occupants for the houses; Chicago Park, laid out in the wash of the San Gabriel River, had 2,289 lots and one resident; while the town of Sunset had 2,014 lots and a watchman." Against such a template, it is tempting to think about California as a blank slate, an empty canvas, with no heritage or history—and yet, the graphs insist otherwise.

How does this work? Let's return to those peaks again, not just collectively but within each of the languages (and one nation, the United States) that Comer charts. In American books, California seems to represent, as we might suspect, its own country: "all that raw land," to borrow a phrase from Jack Kerouac, "that rolls in one unbelievable huge bulge over to the West Coast, and all that road going, all the people dreaming in the immensity of it." Kerouac was a romantic, but what's interesting about the graph is that it is romantic also—or more accurately, romance quantified. It is a portrait of our collective fascination with California, all our arguments and denigrations, as well as the boosters and the hucksters, a portrait in sheer data of the state and what it means. This extends beyond American English, to Spanish, British English, Italian, German, Hebrew, French. References in simplified Chinese rise dramatically between the 1920s and the late 1940s, a function, in part perhaps, of the smallness of the sample, although I prefer to read it in the context of miscegenation laws. Those were dismantled in California in 1948, five years after the repeal of the Chinese Exclusion Act and a year before Mao's revolution, and I would like to think these things are related, reflected in the upward movement of the graph.

Regardless of the vagaries of language, references to the state hit their zenith between the 1970s and the 1990s, and fall off at the new millennium. What this means, I couldn't tell you, but perhaps it indicates a shift beyond what let's call, as McWilliams did, California exceptionalism. This has long been the albatross of every Californian: the burden of its mythos, which I once rejected and then embraced, and now regard with an uneasy ambivalence. The thing about myths, or tropes, is that there has to be some truth to them or else they wouldn't linger as they do. All the same, they obscure a larger truth, a way of thinking, our ability to see this place, any place, for what it is. That, too, is the legacy of Garci Rodríguez de Montalvo and Queen Califia, the legacy of *The Adventures of Esplandián*.

Born out of romance, born out of literature, is it possible that California has finally become, after half a thousand years in the imagination, just another setting in the world? "A city no worse than others," Raymond Chandler observed of Los Angeles, "a city rich and vigorous and full of pride, a city lost and beaten and full of emptiness." What else could he be describing if not California itself? Chandler was a romantic also, but he understood that romance only goes so far. "It all depends on where you sit," he wrote in *The Long Goodbye*, which sits in my mind like a bookend to the legend of Queen Califia, "and what your own private score is. I didn't have one. I didn't care. I finished the drink and went to bed."

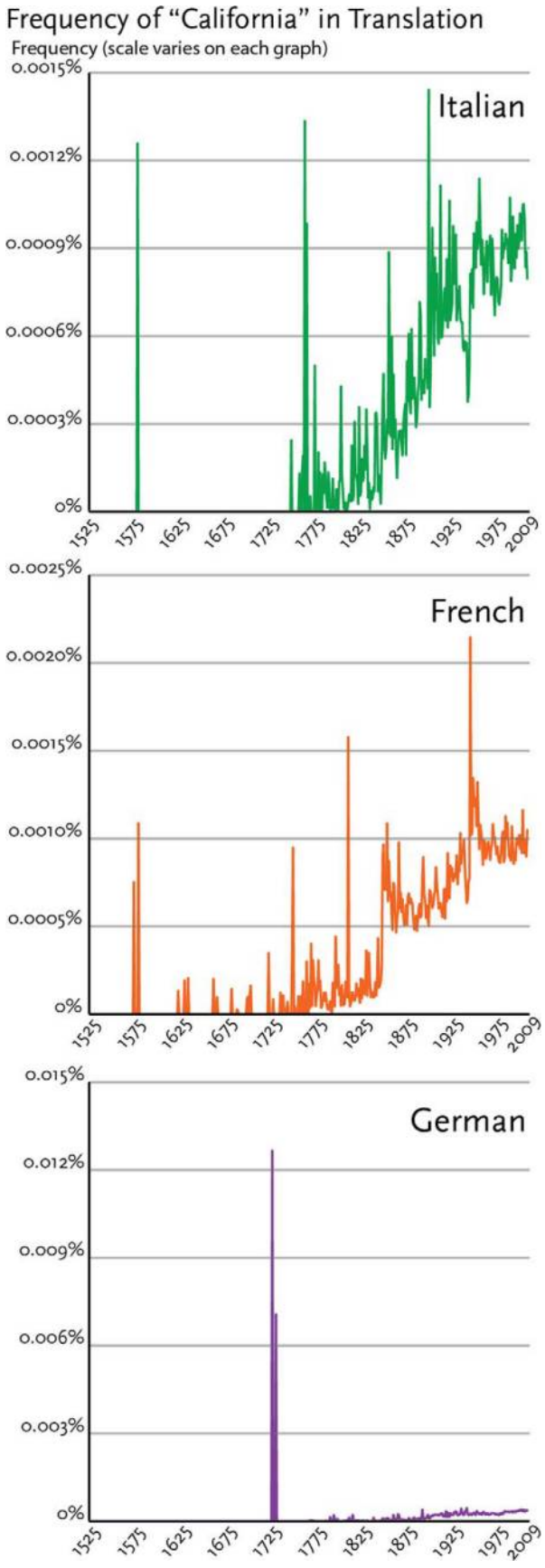
Joshua Comer writes:

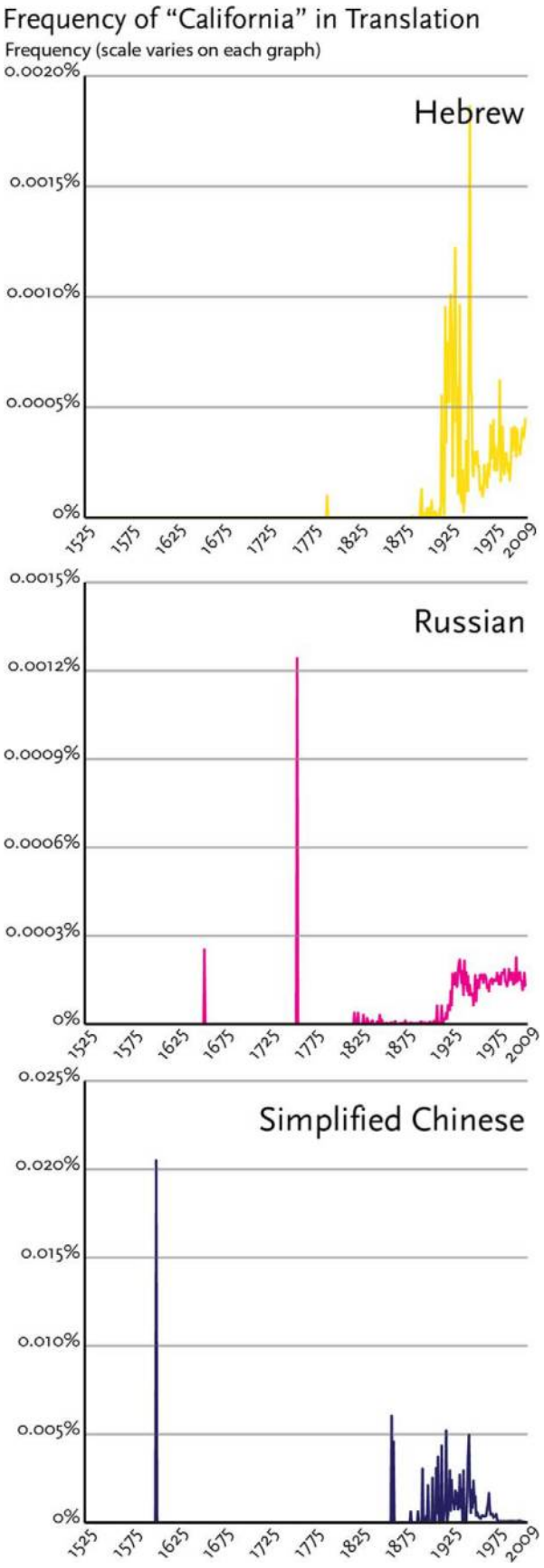
Our analysis of California's place among the billions of words from millions of books amassed at the Google Ngram Viewer web site begins with the "n-gram." An "n-gram" is simply a string of a certain *n* number of words. A one-gram or unigram is one word, such as "California." A two-gram or bigram is a string of two words, such as "California dream," a trigram is a string of three words, and so on.

In our analysis, we looked at the occurrence of unigrams for "California" in English and other languages. We also looked at the occurrence of bigrams for "California" and words that occurred immediately before "California," such as "northern" and "southern," and immediately after, such as "dream," as well as sentences that begin and end with "California." We also looked at bigrams using the adjectival form "Californian."

In this analysis, we measured the frequency of each n-gram among all of the words published each year in books in Google's digitized corpus. In other words, we took the number of times the word "California" occurred in published works that year and divided it by the total number of single words or unigrams published that year. Focusing on uses of "California" between 1800 and 2009, our single-word analysis considers over 36 million appearances of "California" drawn from over 873 billion words. For bigrams, we divided the annual total of each bigram, such as "California dream," by the number of words published that year.

Across languages, we found a fairly regular increase in frequency in the unigram of "California" with noticeable





peaks shortly before and following statehood in 1850, in the final decades of the nineteenth century, in the late 1970s and early 1980s, and a small rise again in the 1990s. The statistical significance of such changes in frequency within individual languages was evaluated by comparing the sometimes volatile year-to-year changes in frequency observed in each language with the proportion of the language in the total global corpus that year, and the total uses of "California" across all languages that year. For example, the Spanish language declines from approximately 25 to 20 percent of all published words in the Google corpus for the years 1945 and 1946. In turn, we would expect Spanish uses of "California" to decline proportionally in relation to total global uses of "California." However, the frequency of Spanish uses of "California" actually remained steady during those years. That makes 1945 and 1946 years in which Spanish use of "California" was significantly different from what one would expect statistically.

In some of our graphs, we used a technique called "smoothing," which creates a moving average across seven years—three before and three after the year in question—to smooth out annual spikes in the data that might be caused by a small number of books or even one single book that contains many instances of "California" in a year in which few books have been digitized by Google. In other graphs, we did not smooth the data because we thought readers might be interested in seeing those annual spikes, particularly in early years.

Smoothing, in addition to giving an aesthetically pleasing and legible wave structure to otherwise noisy data, also says something about our hypotheses about how language works in relation to the published word. If John Steinbeck helped solidify an idea of California nationally and globally toward the end of the Great Depression by publishing *The Grapes of Wrath*, using a three-year moving average in the graph of "California" in the US unigram corpus based on its year of publication, 1939, assumes that Steinbeck's etching of the state's name into the national imagination is connected to three preceding years of writing on California, and its direct impact is registered over the three years following the book's publishing, or 1936 to 1942.

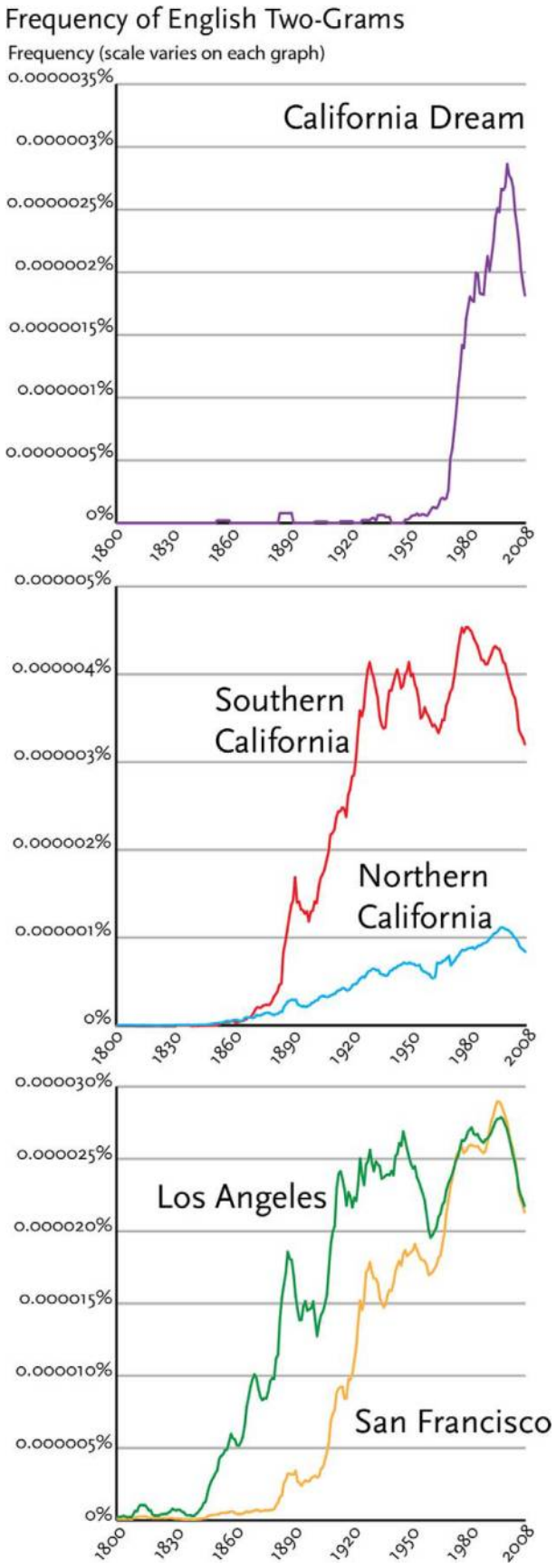
The leveling effect of this quantitative approach is open to criticism. A wildly popular book set in California could, for instance, obscure previous years in which not much was published about the state. Scholars interested in a close

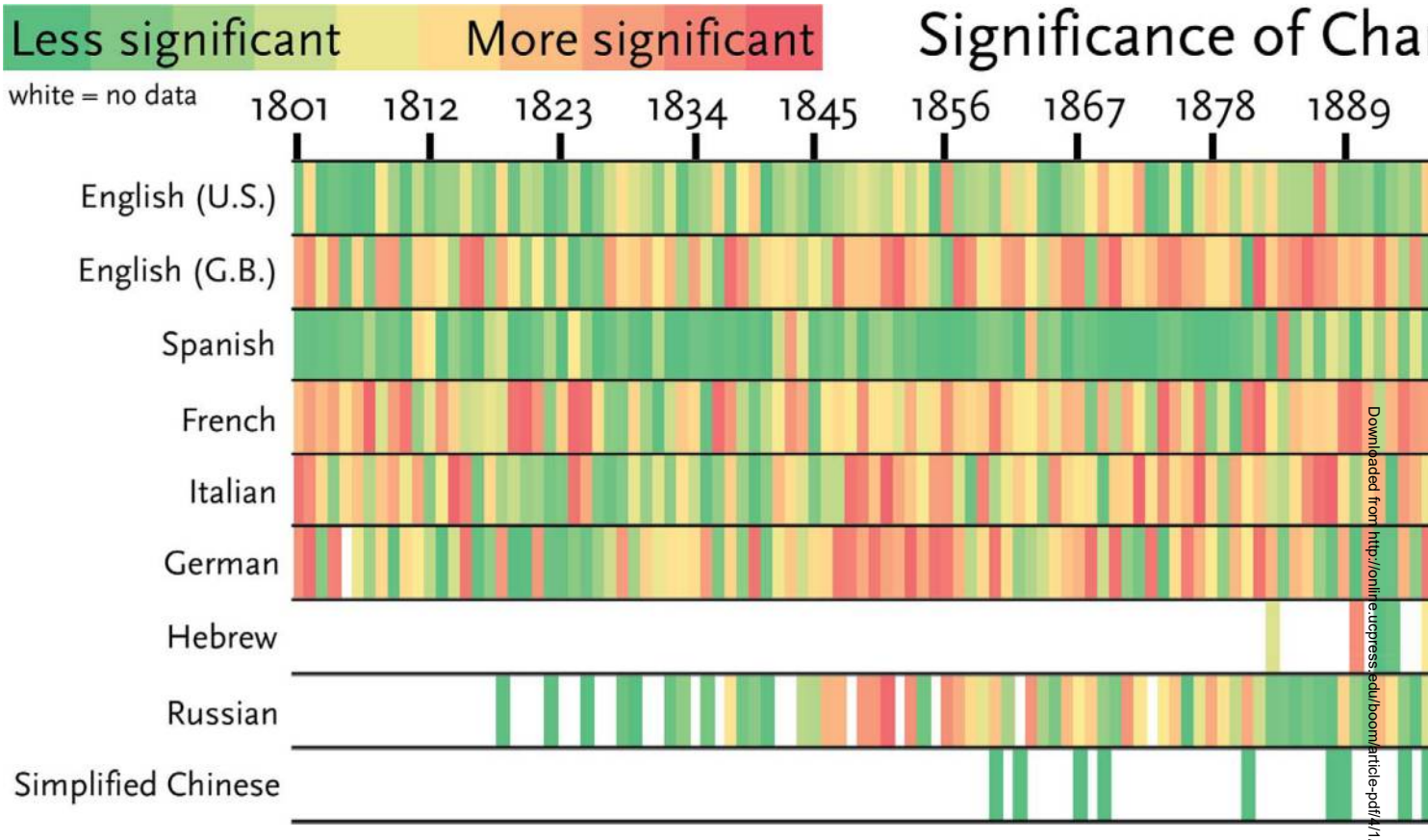
reading of Steinbeck's work might object to the way the lasting influence of *The Grapes of Wrath* is foreshortened to three years. The use of the three-year moving average allows us to begin to measure the influence of "California" in writings such as *The Grapes of Wrath*, but those uses—no matter how many copies of the book were sold, whether a movie was made, or the author won a Nobel Prize—are not directly measured in the smoothed graphs presented here after three years.

These techniques have been called "distant reading"—in that they analyze patterns in a large corpus of text at a great remove from individual texts, let alone specific passages—in contrast to "close readings" of the construction and meaning of individual books and individual passages of text.[I] Both techniques have their value.

Overall, "California" appears in books in English published in the United States almost twice as frequently as in British English books. "California" appears in Spanish-language books at about one-quarter the rate it appears in American books. "California" and "Californie" appear in French books half as often as "California" appears in Spanish books. Italian books feature "California" about as often as French books. From there the occurrence of "California" and its equivalents in different languages falls off in the remaining languages in the Google corpus: German, Hebrew, Russian, and simplified Chinese. In simplified Chinese, occurrences of translations for "California" are generally only one-hundredth the rate of Spanish occurrences of "California," except for some curious peaks, which likely represent simplified Chinese books in the digitized corpus about California, published in years in which very few books in simplified Chinese have been digitized. "California" begins a slow downward trend after 1995 culminating in a significant plunge in frequency of occurrences across all languages, with 2009, the latest year in the corpus marking California's lowest rate of appearances in American books since 1913.

Even the most frequent bigrams using "California" and "Californian" are several orders of magnitude more rare than unigrams—that is, the unigram "Californian" appears at rates a thousand times more frequent than the most frequent bigrams such as "Californian gold"—which makes sense because all bigrams containing "Californian" would also be counted as unigrams. The frequencies of these bigrams are in the low hundred-thousandths of a percent



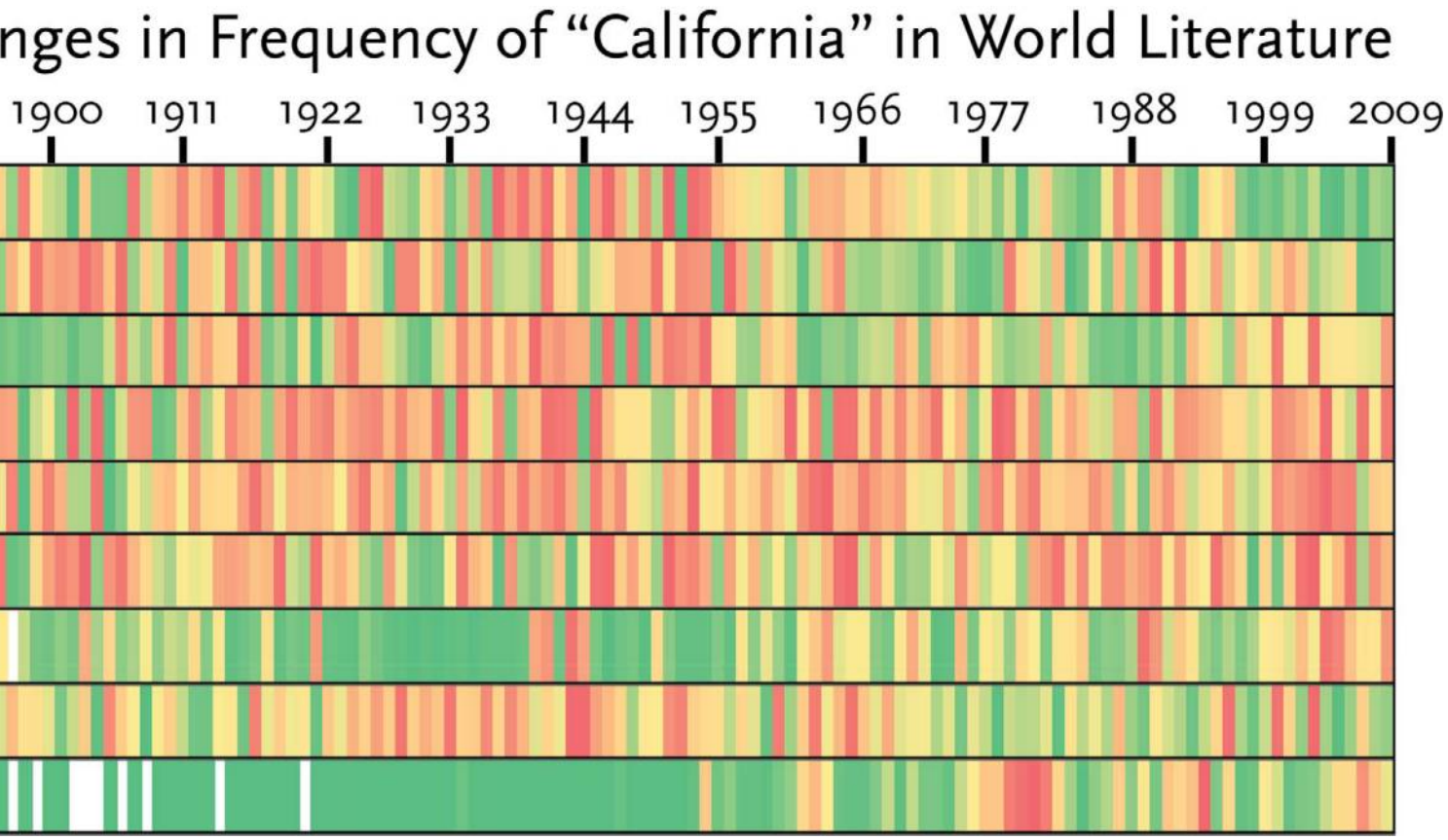


(around 0.00001 to 0.00003 percent) of all bigrams published in any given year.

"Californian Gulf" is the most common bigram of the adjectival form between 1843 and 1848, when it is overtaken by "Californian gold," which is prevalent from around 1850 through 1900, after which it declines sharply. "Californian Indians" and "Californian tribes" are used fairly consistently from around 1850 through 1930. "Californian species" occurs consistently and relatively frequently between 1860 and 1980. "Californian coast" had been consistently and relatively frequently used since 1840, with a large spike around 1890. "Native Californian" and "old Californian" stand out from other pairings due to a bump around 1890. "The Californian" also surges around 1890.

The bigrams "Southern California" and "California Press" stand out among bigrams using "California" rather than "Californian." The presence of "California Press" among the most frequent bigrams indicates the importance of the University of California Press, which publishes *Boom*, in the corpus of published books as a publisher as well as in citations in works published by other presses. When looking at the trigram "University of California," we also found that the University of California makes up approximately 80 percent of all recent occurrences of trigrams extending from the bigram "of California" and about .0015 percent of recent occurrences of the unigram "California" in the American English corpus. The UC system clearly influences what is published about California, directly and indirectly, in many ways.

Despite how easily "California dream" comes to mind when we think of bigram phrases involving California, it is found relatively infrequently in the English corpus. At its highest annual frequency in 1992, "California dream" is found around 4,500 times in fewer than 100 books. If it were not the 1990s, those numbers might still prove the prominence of the phrase, but the enormity of the corpus in



those years puts the dream's highest frequency in the ten millionths of a percent of all bigrams published in those years (.0000009 at its peak using the three-year smoothing technique). If you were to look at all bigrams of that frequency since the 1800s that incorporate "California," you would have to sort through hundreds of bigrams that have been published just as frequently. So much for the California dream? **B**

Note

Franco Moretti, *Distant Reading* (New York: Verso), 2013.