# Perceptions of Palestine: The View from Large Linguistic Datasets

TERRY REGIER

**Cultural norms and trends are often reflected in patterns of language use. This article explores cultural perceptions of Palestine and Palestinians in the English-speaking world, through two analyses of large linguistic datasets. The first analysis seeks to uncover current conceptions of participants in the Israel-Palestine conflict, by identifying words that are distinctively associated with those participants in modern English usage. The second analysis asks what historical-cultural changes led to these current conceptions. A general theme that emerges from these analyses is that a cultural shift appears to have occurred recently in the English-speaking world, marked by greater awareness of Palestinian perspectives on the conflict. Possible causes for such a cultural shift are also explored.**

WORDS, THEMES, AND LINGUISTIC FRAMING can promote particular perspectives on reality at the expense of others. Thus, the debate over how we *speak* about the Israel-Palestine conflict is also ultimately a debate over how we *think* about it: over how the conflict should be conceptually represented, perceived, and construed, and over which aspects of it deserve highlighting and emphasis.

Such ideas and concerns are reflected in a number of recent remarks from a range of commentators. Palestinian diplomat Afif Safieh has argued that mainstream media coverage of the Israel-Palestine question is fundamentally distorted—to the extent, he claimed, that it would make Stalin-era Soviet media blush with envy.[1] Political analyst Yousef Munayyer has singled out,[2] as an element of such distortion, specific expressions such as *relative calm*: he argues that this benign-sounding expression is often used to describe periods of time in which Israelis are no longer being harmed but Palestinians still are. M. J. Rosenberg, presently a political analyst and previously editor of the weekly newsletter of the American Israel Public Affairs Committee (AIPAC), has recently encountered pressure in the United States for persistently using the term *Israel-firster*, which carries the charge of inappropriately prioritized loyalty.[3] Finally, Frank Luntz, a U.S. political consultant, was commissioned by the Israel Project in 2009 to undertake a study of "words that work" in advocating for Israel.[4] That study advised pro-Israel advocates to adopt a tone of empathy for Palestinians, for instance through formulations such as *the Palestinians need books, not bombs*, to avoid alienating undecided audiences. These examples suggest that people with varying political

*Journal of Palestine Studies* Vol. XLV, No. 2 (Winter 2016), p. 41, ISSN: 0377-919X; electronic ISSN: 1533-8614.  DOI: 10.1525/jps.2016.45.2.41.

backgrounds attend closely to language use, and consider it important because of its potential to influence perception and thinking about these issues.

What then are the broad patterns of language use concerning this conflict? What linguistic forms play privileged roles in representing the conflict, and what is the history of these forms? What does language use reveal about general cultural conceptions of the participants in the conflict? My goal here is to begin exploring these questions through the lens of large linguistic datasets. Such datasets capture samples of language use, in both elite and non-elite contexts. I will argue that these widely available resources can help to illuminate perceptions of the conflict, by identifying ideas that are latent in our linguistic environment.

In what follows, I first briefly describe the linguistic datasets on which I rely as well as recent work that has used such resources to explore cultural trends quantitatively. I then use these datasets to explore the specific case of interest, namely language use in the debate over Israel-Palestine. I present two analyses. The first seeks to uncover current cultural conceptions of participants in the conflict, by identifying words that are distinctively associated with those participants in modern English usage. The second analysis explores the historical-cultural changes that led to these current conceptions. A general theme that emerges from these analyses is that a cultural shift appears to have occurred recently in the English-speaking world, marked by greater awareness of Palestinian perspectives on the conflict.

## Large Linguistic Datasets

We increasingly live in an age of "big data"—that is, an age of massive pre-collected datasets that may contain information of interest but are of such size and complexity that they require computational tools for their analysis. Many such datasets contain either primarily or exclusively linguistic data. Examples are the news archives of LexisNexis and Google News, the digitized books of Project Gutenberg, and the digitized books of the Google Books corpus, which is based on the texts of about 4 percent of books ever printed, including texts in English, Chinese, French, German, Hebrew, Italian, Russian, and Spanish.

Such large linguistic datasets are often structured in terms of *n-grams*. An n-gram is a sequence of *n* discrete items, for example words, which appear contiguously in a text or other sequential data stream. For example, the phrase *stand shoulder to shoulder* is a 4-gram because it consists of four words in a row. Some datasets consist not of full text, but instead of n-gram counts—that is, a specification of how often each possible n-gram appears in the full text, for $n \leq 5$ or some similarly small number. For example, such a dataset might contain an entry of the form [*stand shoulder to shoulder* : 4335], encoding the information that the 4-gram *stand shoulder to shoulder* appears a total of 4335 times in the full text on which that dataset is based. Examples of such datasets are Google's Web 1T 5-gram corpus, and the publicly available portion of the Google Books corpus, both of which I use in the analyses below.

Such datasets have the potential to shed light on general cultural tendencies and trends, as reflected in patterns of word use.[5] For example, a recent study drew on the Google Books corpus to analyze cultural trends by tracking the frequency of key words over historical time.[6]

That study reported quantitative data supporting several interesting claims about culture over the past two centuries, including: (1) that we live increasingly in the present—that is, that newly appearing words (and presumably their corresponding concepts) fade into obscurity more quickly now than such words did in the past; (2) that Western culture is decreasingly religious and decreasingly sexist; and (3) that governments suppress mention of individuals they deem undesirable.

To my knowledge, such large linguistic datasets have not previously been brought to bear on the study of the Israel-Palestine conflict. Yet the application seems a natural one. The conflict is highly ideologically polarized and, as we have seen, the use of language in this conflict is itself contested. For this reason, it may be helpful to examine large samples of language use, collected independently and without any known stake in the conflict, and analyzed as objectively as possible, with the goal of understanding what general patterns of language use exist with respect to this topic, and what those patterns reveal about underlying cultural norms and trends.

## Word Images

Many phrases appear to play a central role in the cultural construction of the conflict, and these phrases often carry conflicting perspectives. Examples include *Israeli apartheid*, *Palestinian terrorism*, *the Arab street*, *the cycle of violence*, *relative calm*, and *the Zionist entity*. Such phrases often attract attention because of the assumptions they rest on, and the ideas they convey. In a recent example, repeated Israeli military actions in Gaza that resulted in many deaths have been referred to simply as *mowing the grass*;[7] this metaphor casts deadly violence as a mundane but regularly needed domestic chore, and thus invites the listener to adopt the same perspective. Analyses of such individual phrases can be helpful in illuminating the role of language in this context, and the assumptions on which it rests.[8]

At the same time, it would be desirable to supplement such analyses with something that extends beyond a single pre-specified target phrase or n-gram, and seeks to uncover broader culturally shared conceptions of the region and its inhabitants. Such cultural conceptions of the Middle East have long been the focus of analytical and often critical attention, at least since Edward Said's *Orientalism*.[9] In turn, those analyses have themselves been criticized, in part due to a perceived subjectivity in approach.[10]

Many things have changed since these ideas first gained prominence, and one important change is the recent advent of large linguistic datasets. In this section, I ask what these datasets reveal about the English-speaking cultural imagination as regards the Middle East, and particularly the Israel-Palestine conflict. Concretely, I ask which words or clusters of words emerge as distinctively associated with Israelis, with Palestinians, and with others, in modern English usage. The analyses here seek to extract such "word images" in a data-driven way, in an attempt to ward off possible concerns over partisan filtering or manipulation of the data.

I take a word $w$ to be distinctively associated with a topic $T$ to the extent that $w$ appears frequently in text concerning topic $T$ and infrequently in text concerning other topics. For example, the word *intifada* is distinctively associated with the topic of Palestinians, because this

word appears with some frequency in text concerning Palestinians and infrequently elsewhere. This idea of distinctive association is naturally captured in terms of a likelihood ratio:[11]

$$L(w|T) = \frac{P(w|T)}{P(w|\neg T)} \quad (1)$$

Here, $P(w|T)$ is the probability of encountering word $w$ in text concerning topic $T$, and $P(w|\neg T)$ is the probability of encountering the same word $w$ in text concerning any other topic; the ratio of these two quantities is the likelihood ratio $L(w|T)$. I take a word $w$ to be distinctively associated with a topic $T$ to the extent that $L(w|T) > 1$. In the analyses pursued here, the topic $T$ is chosen based on investigator interest, and words $w$ that are distinctively associated with that topic $T$ are retrieved.

As an illustration and preliminary test of this idea in a context unrelated to the Middle East, I first sought to determine which words were identified by equation 1 as distinctively associated with the Sherlock Holmes genre.[12] The fifty words that are most distinctively associated with this genre are shown in table 1.[13] This list of words captures such relevant themes as puzzle solving (*investigation, questioning, solved, data, observe, theory*) and danger or dread (*dreadful, horrible, evil, strike, pistol, deadly*). The list as a whole provides a condensed "word image" of the Sherlock Holmes stories, which picks up on certain conceptual elements that are distinctively associated with those stories.[14]

What would corresponding word images look like for the major actors in the Israel-Palestine conflict, if derived from statistics of word usage in everyday modern English? To find out, I applied the same ideas to a different empirical resource: the Web 1T 5-gram Version 1 corpus from Google. This corpus consists of observed English n-grams, for $n$ = 1 to 5, together with frequency counts for those n-grams. The corpus was "generated from approximately 1 trillion word tokens of text from publicly accessible Web pages,"[15] and was released in 2006; it seems an appropriate resource because of its broad and relatively recent coverage. Since this corpus is based on text from Web pages, it is not restricted to elite discourse, and instead represents a broad spectrum of opinion, including what can be thought of as the "electronic street." Because the corpus contains only n-gram counts, I searched for bigrams (2-grams) in which the first word is a national adjective such as *Palestinian*, *Israeli*, *French*, *Korean*, or the like,[16] and the second word is any word that follows such a national adjective in the corpus. This search retrieved such bigrams as: *Palestinian cameraman*, *Israeli hotels*, *French economy*, *Korean invitation*, and so on. Given such bigrams and their associated frequency counts, I asked which words were most distinctively associated with which preceding national adjectives. That is, for a given national adjective such as *Palestinian*, I used equation 1 to determine which words $w$ appear frequently following that

**Table 1. Words distinctively associated with the Sherlock Holmes genre**

| |
|---|
| goose cab inspector pounds advice finger locked dreadful trust key doctor obvious wall colonel horrible band trap investigation traces formed evil questioning solved data west lantern profession shown ship bag crop metal observe imagine strike bureau medical pistol pool sofa deadly sunk thoroughly theory notes wrist news wooden commonplace smell |

national adjective and infrequently following others. Effectively, this process retrieves words *w* that appear disproportionately often in a target context defined by a given national adjective, such as *Palestinian* ___, for example *Palestinian intifada* or *Palestinian territories*.[17]

Table 2 shows the fifty words most distinctively associated with the context *Palestinian* ___ in this analysis, shown in order of decreasing distinctiveness of association. Several themes emerge from this list. One unfortunate but predictable theme is violence (*detonated, gunmen, suicide, assailant, assassinated*). This should not be surprising: the conflict is a violent one, and the specific character of that violence is salient in media and other representations—at the same time, it is noteworthy that the word *nonviolent* also emerges as distinctively Palestinian. Other themes that emerge are those of deprivation and difficulty (*hardship, bereaved, malnutrition, statelessness, dispossession, desperation*) and chaos (*disarray, discord*). There are also words that do not clearly cohere into general themes but that nonetheless seem significant. As expected, *intifada* and its English equivalent *uprising* both emerge as distinctively Palestinian, as does the word *contiguity*, a geographic concern that is highly relevant in the Palestinian case and less so elsewhere, also reflected in the word *bantustans*. Finally, *steadfastness* (the English equivalent of Arabic *sumud*) is distinctively associated with the context *Palestinian* ___.

Table 3 shows the corresponding list for the context *Israeli* ___. Again, several themes emerge, and again one of them is violence, this time of a character associated with organized military occupation (*bulldozer, crossfire, crackdowns, demolitions, gunships*). As before, this should not be surprising, because the conflict generally and the occupation in particular are violent and widely represented as such. Also as before, there are words that signal dissent from the general theme of violence (*peaceniks, refuseniks*). Another salient theme, also associated with the occupation, and highlighting the asymmetrical character of the conflict, is prevention of movement (*checkpoints, roadblocks, fence, curfew, closure, barrier, obstacles, apartheid*). Finally, and to my surprise, there is an apparent theme of repetition (*redeployment, reoccupation, retaliations*). The implication appears to be that according to the statistics of this corpus, other countries may also deploy and occupy, but it is distinctively Israeli to *re*deploy and *re*occupy. This theme of repetition is also

**Table 2. Words distinctively associated with the context *Palestinian* ___.**

| |
|---|
| sacking pivot groves disarray electioneering rejectionist intifada detonated contiguity beautification hardship vehicular bereaved statehood legitimate gunmen suicide ambulances fedayeen malnutrition nonviolent livelihood bantustans pollsters statelessness vetoed assailant territories passerby preventive steadfastness discord awe refusing homicidal assassinated helm optimists cashier dispossession fixers uprising negotiator legislative exhaustion frustrations balloting stabs disguised desperation |

**Table 3. Words distinctively associated with the context *Israeli* ___.**

| |
|---|
| bulldozer shekels kibbutz retaliations crossfire assassinations hints ashamed sonic disengagement checkpoints hybridization redeployment callousness doves roadblocks foreknowledge unclear apache procrastination fence crackdowns curfew reoccupation preemptive escapism demolitions warn closure occupying cyberwar counterterror barrier wiretapping pullout reservist gunships upgraded yeshivot escalation couscous watchtower pilotless pullback peaceniks aggressions obstacles apartheid clampdown refuseniks |

present in the recent metaphor *mowing the grass*, referenced above. Finally, again in connection with this theme, the appearance of the word *retaliations* is interesting because its presumed distinctively Israeli character has been called into question by recent quantitative research.[18] That research investigated the prior claim that Israelis deploy violence in response to Palestinian attacks whereas Palestinians are "uncontingently violent," in that their violence is not predicted by Israeli attacks—and found that the claim was not supported.[19] Instead, the data examined in that research suggested that both sides deploy violence in response to attacks by the other; that is, both sides retaliate. This underscores the more general point that these word images are not necessarily accurate representations of what is distinctive and exceptional about the actors themselves—they are instead an attempt to capture culturally shared conceptions of those actors, whether or not those conceptions are accurate.

Finally, table 4 shows words that are distinctively associated with the context *Lebanese* ___. Lebanon was chosen for comparison because it is geographically and culturally very near the Israel-Palestine conflict, has been regularly affected by the conflict, and has a history of recent violence that overlaps with but is not identical with the conflict. Interestingly, despite these similarities, the word image for *Lebanese* is rather different in overall tone from those for *Palestinian* and *Israeli*. Two major themes that emerge here are food (*hummus, mezze, falafel, pita, cucumbers, sweets*) and dispute (*denounce, demonstrating, blame, accuse, resent, demanding, protesting, insult*)—on the whole a rather Mediterranean feel. There are also traces of the theme of violence (*hijacking, guerrilla*), but these are less prominent than in the other two lists we have seen—despite the extreme violence of the Lebanese civil war and periodic hostilities since then.[20]

These word lists provide verbal snapshots that are intended to capture distinctive features of verbally transmitted cultural conceptions of major actors in the Middle East. Similar analyses can also easily be applied to other nationalities, to explore national stereotypes of other actors in the Middle East and elsewhere.

## Cultural Change

I next turn to exploring the conflict and its participants in historical perspective, by drawing on the Google Books corpus. Google has digitized the texts of millions of books published in several languages from the 1500s to 2008, and has made n-gram counts from these texts publicly available, time-stamped by year of publication. The analyses reported here are based on the English section of this corpus. To the extent that books can be seen as reflecting the linguistic usage of cultural elites, the results of the following analyses emphasize elite over non-elite usage.

**Table 4. Words distinctively associated with the context *Lebanese* ___.**

| |
|---|
| denounce webmaster hummus forte intelligences hijacking demonstrating mezze livre blame flooded mourn unite massed allele flocked falafel accuse reside cedars resent cronies demanding hash stay clairvoyant pita confessional diver middleman protesting tracks proudly insult cucumbers gangsta loop responsible poured surprised proxies oud psychic hoped detained sweets gathered grooms bubbly guerrilla |

The words *Israel* and *Palestine* themselves constitute the obvious starting point for a historical exploration of the Israel-Palestine conflict. Figure 1 shows the frequency profiles for these two words over the years 1800–2008.[21] In this and subsequent graphs, the y-axis shows the relative frequency of a given word, or the proportion of all text that it represents, over time. For example, in figure 1, the relative frequency of the word *Israel* reaches an upper limit of 0.00005915 in the year 1984; this means that in the corpus for that year, the word *Israel* occurred around 59.15 times per million words of text.

Aspects of the profiles for *Israel* and *Palestine* cohere naturally with the history of the conflict. *Palestine* begins as a relatively low-frequency word in the early 1800s, when Palestine was part of the Ottoman Empire, and was of interest in the English-speaking world primarily as the Holy Land, but not yet as a major target for British political power. Its relative frequency then rises during World War I and the subsequent Mandate period, when Palestine did become a target, and then an acquisition, of the British. Its prominence peaks around 1948 (marked by the gray vertical line), the year in which British administration of Palestine ended and the state of Israel was established. After that date, the frequency of *Palestine* falls off to roughly Ottoman-era levels of obscurity. The profile of the word *Israel* post-1948 is also explained by these events: with the establishment of the state, use of the word naturally increases. However, the profile of *Israel* prior to 1948 is not explained by these political events. *Israel* was a relatively common word in the early 1800s; following Biblical usage, it generally referred to Jews as a people, for example "the sons of Israel,"[22] "the children of Israel,"[23] as well as "the dispersion of Israel"[24] and "the conversion of Israel."[25] Interestingly, the word *Israel* begins to lose prominence around the beginning of active Zionist immigration to Palestine in the 1880s, and continues to drop in frequency throughout the pre-state period, until just before 1948.

An interesting contrast with the words *Israel* and *Palestine* is provided by the corresponding derived forms, *Israeli* and *Palestinian*.[26] Their frequency profiles are shown in figure 2. The post-1948 profile for *Israeli* roughly mirrors that for *Israel* after the founding of the state, but the pre-1948 era shows minimal use of *Israeli*, presumably because that term is associated primarily with the state, which did not yet exist. Notably, the word *Palestinian* eventually becomes almost as frequent as the word *Israeli*, in sharp contrast with the low frequency of *Palestine* relative to

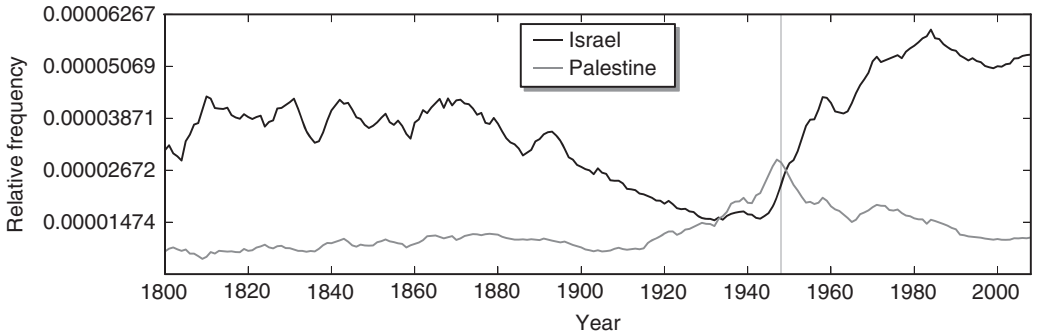


**Figure 1.** The relative frequency of the English words *Israel* and *Palestine*, for each year from 1800 to 2008. Smoothing = 3 years. The gray vertical line marks 1948. All historical data in this section are from https://books.google.com/ngrams/.

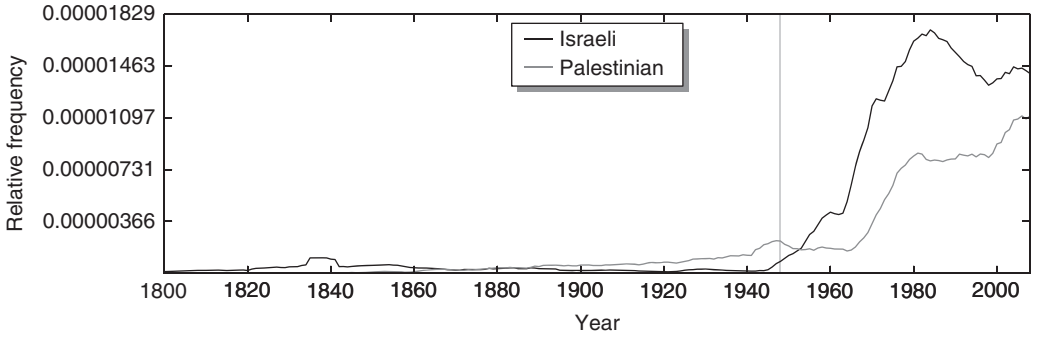


**Figure 2.** The relative frequency of the English words *Israeli* and *Palestinian*, for each year from 1800 to 2008. Smoothing = 3 years. The gray vertical line marks 1948.

*Israel* (compare the rightmost ends of figure 1 and figure 2). This makes sense: *Palestine* today references primarily a memory of the past and an aspiration for the future, whereas *Palestinian* references a presently active social and political force. Moreover, the profile of *Palestinian* recapitulates important aspects of the history of the Palestinian national movement. As with *Palestine*, we see a slow rise in the frequency of *Palestinian* from the 1800s through 1948, correlating with British colonial interest in the region. After 1948, the prominence of *Palestinian* decreases somewhat and stays relatively low for the period in which the conflict with Israel was dominated by the Arab states (1948 through the mid-1960s) rather than by Palestinians themselves. During this period "the Palestinians seemed to many to have disappeared from the political map."[27] This period ended when Palestinians began to assume a more central role in the conflict, through the establishment of the Palestine Liberation Organization (1964), the discrediting defeat of the Arab states in war (1967), and the Battle of Karameh (1968): this political change is reflected in the rise in frequency of the word *Palestinian* at that time, peaking in the late 1970s and early 1980s, a time of intense Israeli-Palestinian hostilities that culminated in the 1982 Israeli invasion of Lebanon. After this, the prominence of the word *Palestinian* remains high, to rise still further at the beginning of the second intifada in 2000.[28] The overall picture is one of increasing recent prominence for the word *Palestinian*—despite the low profile of the word from which it is derived: *Palestine*. This pattern inverts the earlier preference of certain Arab states, which were happy to support the cause of Palestine as long as the status of Palestinians in their own countries was left unaddressed—a stance that has been summarized as "Palestine yes, Palestinians no."[29] In contrast, modern English usage may be summarized as: *Palestinian* yes, *Palestine* no.

With increasing prominence for the word *Palestinian*, one might expect a corresponding recent increase in prominence for words or phrases that convey a Palestinian perspective on matters. One such phrase is *the Nakba*, a borrowing into English of the Arabic *al-nakba* (the catastrophe), referring to the loss of Palestine in 1948.[30] This most unambiguously Palestinian of terms has recently seen a remarkable increase in English usage, as shown in figure 3.

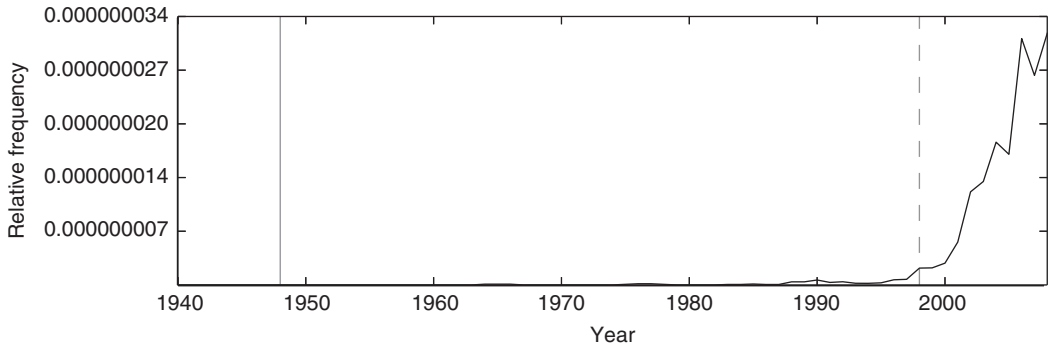


**Figure 3.** The relative frequency of the English phrase *the Nakba*, for each year from 1940 to 2008. Smoothing = 1 year. The solid gray vertical line marks 1948, and the dashed gray vertical line marks 1998.

The expression *the Nakba* was unknown in English at the time of the event itself (1948, marked by the solid gray vertical line), and it appeared at negligible frequencies until the fifty-year commemoration of the event, in 1998 (marked by the broken gray vertical line). Shortly after that, the expression's frequency rose sharply, yielding a 91-fold overall increase from 1990 to 2008. This is of course an increase from near zero, and *the Nakba* is still a very low-frequency term—but it is now an established element of the English language, with frequency comparable to the word *autodidactic*.[31] Although not shown here, some other terms that also assume a Palestinian perspective show a similar (although less extreme) post-1998 increase in frequency: *Palestinian right of return* increases roughly 4-fold in relative frequency from 1997 to 2008, and *Israeli apartheid* roughly 6-fold. A potentially significant contrast is *Deir Yassin*, the site of an infamous massacre during the Nakba: unlike the perspective-adopting terms considered above, the term *Deir Yassin* has appeared regularly (although at low frequency) in English since 1948, and it shows no clear increase in frequency post-1998. It may be relevant that *Deir Yassin* is a place name and therefore must be mentioned in any detailed reference to the event that occurred there, whereas *the Nakba* and the other perspective-adopting terms mentioned above are avoidable expressions, the use of which suggests deliberate alignment with, or at least acknowledgment of, that perspective on events. Although these data are sketchy and initial, they suggest a recent shift in English usage toward greater openness to Palestinian perspectives.[32]

If such a shift has happened, what brought it about? It is impossible to be certain, but some hints may be gathered by seeing which other terms pattern like *the Nakba*. To identify such terms, I searched through the Google Books English corpus, focusing on the years 1990–2008, and extracted those 1-grams (words) that pattern like *the Nakba* in that they also show a sharp upswing in frequency at the same time.[33] Table 5 shows the fifty most frequent such words, appearing in order of decreasing frequency.

Several general themes emerge from this set of recently prominent words. Some of these themes—for example nanotechnology (*nanoparticles, nanotubes*), genetics (*microarray, genomics*)—do not seem especially relevant to perceptions of Palestine, and these are most likely developments that merely happened to rise to prominence at this time. However there are two other themes that do seem potentially relevant. One of these is Internet-based communication (*website, Google, blog,*

**Table 5. The fifty most frequent words that, in parallel with *the Nakba*, show a sharp upswing in frequency over the period 1990–2008. Words are listed in order of decreasing frequency.**

website Retrieved Google Qaeda websites eBay outsourcing blog Accessed workflow SharePoint Katrina Dummies nanoparticles Hamas Palgrave blogs Putin neoliberal Islamist spam Rumsfeld microarray Mahwah ontologies Starbucks DVDs doi emails nano Hezbollah VoIP Ashgate nanotubes nanotechnology fMRI euros genomics Sarbanes neoliberalism Ghraib InDesign ICTs transgender Mbeki Kolkata microarrays outsourced Musharraf podcast

*emails*). This may be relevant to perceptions of Palestine because it signals new means of transmitting and acquiring information, and a leveling of the informational playing field such that information transmission concerning current events and the history behind them is no longer dominated by a few well-established outlets, or at least not as much as it was previously. The other potentially relevant theme is terrorism and the war on terror: *Qaeda* presumably from *al-Qaeda*; *Hamas*; *Hezbollah*; *Ghraib* presumably from *Abu Ghraib*; and *Rumsfeld*.[34] The 9/11 Commission report noted that "the principal architect of the 9/11 attacks" claimed to have been motivated by "his violent disagreement with U.S. foreign policy favoring Israel,"[35] and this observation has motivated increased post-9/11 American attention to the Israel-Palestine conflict.[36]

Correlation is not necessarily evidence of causation, and we have no firm basis for concluding that these two developments—Internet-based communication and the war on terror—are in fact the causes of the apparent newfound awareness of Palestinian perspectives in the English-speaking world. Moreover, if there is a causal link, we have no reason to imagine that it would be a straightforward or direct one. Instead, if a link exists, it seems more likely that it may have resulted from a complex interaction involving an urgent shift of attention to the Middle East brought about by the 9/11 attacks and their aftermath, a strong drive among some Americans to understand the root causes of these events,[37] heightened by the events of the second intifada (2000–2005), and facilitated by greater and speedier access to a wide range of Web-based information.

A related possibly relevant consideration is that as 9/11 assumed the status of a culturally defining tragedy for Americans, the attention given to earlier culturally defining tragedies may have diminished somewhat, as suggested by figure 4.

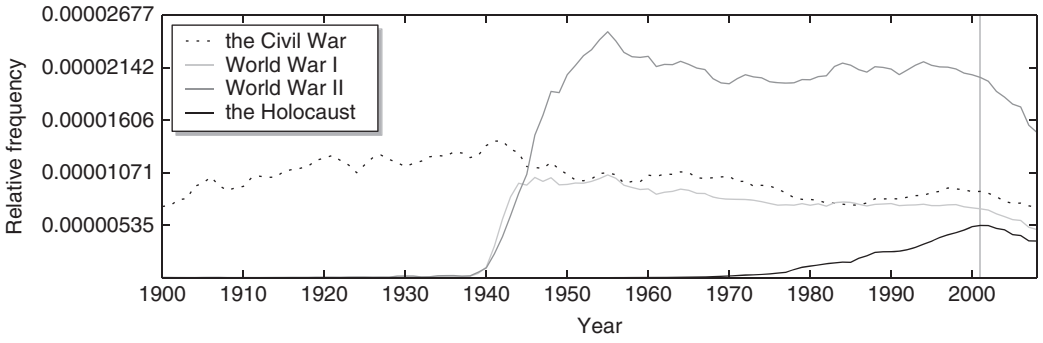


**Figure 4.** The relative frequency of the terms *the Civil War, World War I, World War II*, and *the Holocaust*, for each year from 1900 to 2008. Smoothing = 1 year. The gray vertical line marks 2001.

The relative frequencies of *the Civil War* (dotted line), *World War I* (light gray line), *World War II* (dark gray line), and *the Holocaust* (black line) all show a decline during the 2000s. In some cases, the decline began shortly before this period, but at least in the case of *World War II*, the decline appears to have accelerated post-9/11. Notably, in the case of *the Holocaust*, the frequency profile climbs until the year 2001, peaks then, and declines thereafter.[38] These data suggest that after 9/11, earlier culturally-defining tragedies such as those of World War II may have become slightly less prominent in the cultural imagination; it is possible that as that happened, the standard views of the post–World War II order also receded somewhat, permitting new perspectives on old topics.

These causal scenarios are only conjectures, and the actual cause of the apparent new openness to Palestinian perspectives may lie elsewhere. It is also important not to overstate the extent of this openness: nothing we have seen suggests a complete break with earlier views of Palestine and Palestinians. Still, however incremental the change, and whatever its cause, the evidence we have seen here does suggest that some sort of cultural shift has taken place recently in the English-speaking world that has in some way enabled a greater openness to, if not necessarily acceptance of, Palestinian perspectives.

## Conclusions

We have seen that large linguistic datasets can identify patterns of language use concerning the Israel-Palestine conflict, and can thus help to illuminate the cultural conceptions that presumably lie behind that language use. Specifically, we have seen that exploration of such datasets can reveal current cultural assumptions about specific countries and nationalities, by distilling out words that are distinctively associated with them, and can also help to uncover the historical-cultural changes that led to the present cultural state of affairs. The analyses presented here suggest that English usage has recently come to reflect Palestinian perspectives on the conflict to a greater degree than it previously did, suggesting a greater awareness of those perspectives in the English-speaking world.

Much of the data on which these analyses were based are available to anyone with an Internet connection. Moreover, the amount of computation required to conduct these analyses ranged from negligible (a few seconds interacting with the Google n-gram viewer website, in the case of tracking a single key word over time), to modest (a few hours on a midrange laptop, when searching for all words that match a specific frequency profile). Thus, such analyses are relatively straightforward to run, and I hope to have encouraged the use of such tools with the analyses I have presented here. There are many subsequent analyses that suggest themselves for future research, including more thorough tests in English, and analogous tests in other languages.

At the same time, there are still other analyses that cannot yet be conducted. A natural vision for future work is to sweep these techniques across cultures as well as across time, to explore vanished but relevant cultural worlds, such as that of the Ottoman era. What rhetorical tools, what cultural conceptions of specific actors, and what changes in such ideas over time can be found in that and other historical eras? We cannot presently apply the ideas explored here to answer those questions because the required data are not yet electronically available. However, that is changing. As more and more historical linguistic data are compiled into large linguistic datasets and Web-posted, such approaches to these questions may become increasingly feasible.

**About the Author**

Terry Regier is a professor of linguistics and cognitive science at the University of California, Berkeley.
The ideas in this article were first presented in December 2013 at the *Disputed Words: Palestine, Language and Political Discourse* conference, organized by the Institute for Palestine Studies and the American University of Beirut. The author wishes to thank the organizers, participants, and audience of that workshop for their helpful feedback.

## ENDNOTES

1 Afif Safieh, remark during the question and answer period following his address to the World Affairs Council of Northern California, San Francisco, 1 August 2006. Available at "Ambassador Afif Safieh," fora.tv, http://fora.tv/2006/08/01/Ambassador_Afif_Safieh.

2 Yousef Munayyer, "The Top 7 Terms That Distort Israel/Palestine," *Permission to Narrate* (blog), Jerusalem Fund, 24 March 2011, http://blog.thejerusalemfund.org/2011/03/top-7-terms-that-distort.html.

3 M. J. Rosenberg, "The Israel-Firster Brouhaha and Why I Left Media Matters," *The Blog*, *Huffington Post*, 7 April 2012, http://www.huffingtonpost.com/mj-rosenberg/the-israel-firster-brouha_1_b_1409931.html.

4 Frank Luntz, "The Israel Project's 2009 Global Language Dictionary," The Israel Project, 2009, Web-archived, available at http://web.archive.org/web/20090730194258/http://http://www.newsweek.com/media/70/tip_report.pdf.

5 This idea should be approached with caution. Big data can usefully supplement, but cannot replace, other means of treating these questions. In this connection, see David Lazer et al., "The Parable of Google Flu: Traps in Big Data Analysis," *Science* 343, no. 6176 (2014): pp. 1203–5; see especially p. 1203 on avoiding "big data hubris." Moreover, there is inevitably some noise in the data. For example, the Google Books database has been criticized for incorrectly recording the date of publication for some of the texts it contains (Geoffrey Nunberg, "Google's Book Search: A Disaster for Scholars," *Chronicle of Higher Education*, 31 August 2009). Fortunately, "a large proportion" of these errors appear to have been subsequently corrected (Geoffrey Nunberg, "Counting on Google Books," *Chronicle of Higher Education*, 16 December 2010).

6 Jean-Baptiste Michel et al., "Quantitative Analysis of Culture Using Millions of Digitized Books," *Science* 331, no. 6014 (2011): pp. 176–82.

7 Efraim Inbar and Eitan Shamir, "Mowing the Grass in Gaza," *Jerusalem Post*, 22 July 2014.

8 For example, Julie Peteet, "Beyond Compare," *Middle East Report* 39, no. 253 (2009): pp. 16–25; Terry Regier and Muhammad Ali Khalidi, "The Arab Street: Tracking a Political Metaphor," *Middle East Journal* 63, no. 1 (2009): pp. 11–29.

9 Edward Said, *Orientalism* (New York: Pantheon, 1978).

10 Bernard Lewis, chap. 6, "The Question of Orientalism," in *Islam and the West* (New York: Oxford, 1993).

11 This follows a recent proposal in cognitive science that characterized the cognitive notion of representativeness as the logarithm of the likelihood ratio. See Joshua B. Tenenbaum and Thomas L. Griffiths, "The Rational Basis of Representativeness," *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (2001): pp.1036–41.

12 For each word *w* that appears in the full text of *The Adventures of Sherlock Holmes* (from Project Gutenberg, http://www.gutenberg.org), I let $P(w|T)$ be the relative frequency of *w* in that text, and I approximated $P(w|\neg T)$ by the relative frequency of *w* in a specific text from another genre, *The Best American Humorous Short Stories*, also from Project Gutenberg. I excluded words that appeared only in the Sherlock Holmes text, and words that began with capital letters, such as proper names.

13 For further (for example, quantitative) details concerning this and other analyses in this paper, please contact the author directly.

14 In these results, and those reported below, there are also a number of retrieved words the relevance of which is not immediately obvious, for example *west* in the Sherlock Holmes context.

15 This corpus is described in detail at https://catalog.ldc.upenn.edu/LDC2006T13 and is available through the Linguistic Data Consortium (LDC), an organization hosted by the University of Pennsylvania that creates and distributes language data resources including corpora.

16 I obtained a list of national adjectives by searching for all words *w* that begin with a capital letter and appear in a bigram of the form *w nationalism* (for example *Japanese nationalism*) in the corpus. I removed by hand those (relatively few) words that were not national adjectives or that referred to a national identity only indirectly through a political ideology (for example *Kemalist, Baathist, Zionist*). This produced a list of 220 adjectives that specified national or other ethnic identities, including, for example, *African, Hawaiian, Spanish*, and, of course, *Palestinian* and *Israeli*.

17 As in the Sherlock Holmes analysis, I excluded words that appeared only following the target national adjective, and words that began with capital letters, such as proper names. I also excluded variants of words already on the list (for example if *cucumbers* was already on the list I excluded *cucumber* as a separate item). I also excluded obvious misspellings such as *electio, seige*. A given national adjective, for example *Palestinian*, is also often used to denote a person of that nationality; these different uses were treated identically.

18 Johannes Haushofer, Anat Biletzki and Nancy Kanwisher, "Both Sides Retaliate in the Israeli-Palestinian Conflict," *Proceedings of the National Academy of Sciences* 107, no. 42 (2010): 17927–32.

19 Haushofer, Biletzki and Kanwisher, "Both Sides Retaliate," p. 17927.

20 It may be relevant that the dataset on which these analyses were based was released in September of 2006, just a few months after the July War of 2006, so it is possible that language use concerning that war did not make it into the corpus.

21 In this figure, data are shown with smoothing of three years. This means that the relative frequency for each year is averaged together with the relative frequencies for three years on either side of that target year. Smoothing allows general trends across time to appear more clearly.

22 An Ausonian [P. Anichini], *A Few Remarks on the Expediency and Justice of Emancipating the Jews* (London: Royal Exchange, 1829), p. 61.

23 Ebenezer Wickes, *A Compendium of the Travels of the Children of Israel from Egypt to the Land of Canaan* (Albany, NY: John B. Johnson, 1823).

24 Isaäc da Costa, *Israel and the Gentiles: Contributions of the History to the Jews from the Earliest Times to the Present Day* (London: James Nisbet, 1850), p. 3.

25 Joseph Samuel Christian Frederick Frey, *Judah and Israel: Or, the Restoration and Conversion of the Jews and the Ten Tribes* (New York: D. Fanshaw, 1840), p. iv.

26 These analyses report frequency for these words without regard to whether they appeared as adjectives or nouns, whether the intended referent was singular or plural, and so forth.

27 Rashid Khalidi, *Palestinian Identity: The Construction of Modern National Consciousness* (New York: Columbia University Press, 2010), p. 178.

28 However, there is no obvious reflection in this frequency profile of such prominent events as the first intifada (1987–91), or the Oslo accords (1993).

29 Edward Said, "Permission to Narrate," *Journal of Palestine Studies* 13, no. 3 (1984): p. 33.

30 Constantin Zureiq, *Ma'na al-nakba* [The meaning of the catastrophe] (Beirut: Dar al-ílm lil-malayin, 1948).

31 Over the years 2006 to 2008, the average relative frequency for *the Nakba* was 2.63 occurrences per hundred million words of text, whereas that for *autodidactic* was 2.36 occurrences per hundred million words of text.

32 One complication for this account is that the word *Palestine* is also a perspective-adopting term, and a very relevant one, yet as we have seen it has not recently increased in frequency, at least not in written text on which the Google Books corpus is based. Informally, my sense is that there has been an increase in spoken use of the word *Palestine* in such expressions as *visiting Palestine*, or for that matter *Israel-Palestine* as a topic. It is possible (but undemonstrated) that such a change in usage occurred in spoken English prior to its appearance in the written language.

33 Specifically, for each word *w* in the corpus, I determined the relative frequency of *w* for the years 1990 to 2008, with smoothing = 1 year, and I then extracted *w* if it met three conditions: (1) the correlation of the relative frequency profile of *w* over the target years with the corresponding profile for *the Nakba* yielded Pearson's $r \geq 0.9$; (2) the relative frequency for *w* in 2008 was more than ten times greater than it was in 1990; and (3) the maximum relative frequency for *w* over the target years was at least as great as that for *the Nakba*.

34 The expression *war on terror* itself, although not included in the above analysis because it is a 3-gram rather than a 1-gram, also correlates strongly with *the Nakba* (Pearson's $r > 0.96$), and shows a greater than 6500-fold increase in relative frequency from 1990 to 2008.

35 "Al Qaeda Aims at the American Homeland," Archive: National Commission on Terrorist Attacks upon the United States, http://www.9-11commission.gov/report/911Report_Ch5.htm.

36 John J. Mearsheimer and Stephen M. Walt, *The Israel Lobby and U.S. Foreign Policy* (New York: Farrar, Straus and Giroux, 2007). The authors cite the 9/11 Commission report in this connection on p. 67.

37 For example, Robert Pape, *Dying to Win: The Strategic Logic of Suicide Terrorism* (New York: Random House, 2005); Mearsheimer and Walt, *The Israel Lobby*, 2007.

38 The earlier portion of the profile for *the Holocaust* is consistent with Peter Novick's contention that the term first rose to prominence in American life in the late 1960s to early 1970s. See Novick, *The Holocaust in American Life* (New York: Houghton Mifflin, 1999).