CA Journal of Cultural Analytics

The Canon of Dutch Literature According to Google

Lucas van der Deijl, Roel Smeets, and Antal van den Bosch

09.30.19

Peer-Reviewed By: Timothy Tangherlini

Clusters: Infrastructure

Article DOI: 10.22148/16.046

Dataverse DOI: 10.7910/DVN/T79QZE

Journal ISSN: 2371-4549

Cite: Lucas van der Deijl, Roel Smeets, and Antal van den Bosch, "The Canon of Dutch Literature According to Google," Journal of Cultural Analytics. September 24, 2019.

Literary history is no longer written in books alone.[12] As literary reception thrives in blogs, Wikipedia entries, Amazon reviews, and Goodreads profiles, the Web has become a key platform for the exchange of information on literature. Although conventional printed media in the field—academic monographs, literary supplements, and magazines—may still claim the highest authority, online media presumably provide the first (and possibly the only) source for many readers casually interested in literary history. Wikipedia offers quick and free answers to readers' questions and the range of topics described in its entries dramatically exceeds the volume any printed encyclopedia could possibly cover.[1] While an important share of this expanding knowledge base about literature is produced bottom-up (user based and crowd-sourced), search engines such as Google have become brokers in this online economy of knowledge, organizing information

[1] It should be noted that there are large differences in size and coverage between the different language versions of Wikipedia.

on the Web for its users. Similar to the printed literary histories, search engines prioritize certain information sources over others when ranking and sorting Web pages; as such, their search algorithms create hierarchies of books, authors, and periods.

This article explores these algorithmically constructed hierarchies as cultural representations of what is (and what is not) presented as important to the Google user, taking information about authors from a particular body of national literature as a case study. We examine the relations between a sample of Dutch writers through the carousels of related searches generated by Google's Knowledge Graph. The sample used for this experiment comprises all 2,287 individuals who were labeled on Dutch Wikipedia with the category "Dutch writer" (*Nederlandse schrijver*).[2] On Wikipedia, this is the general category for book authors from the Netherlands and includes Dutch writers from all possible genres, such as literary prose, poetry, (literary) thrillers, fantasy, nonfiction, and cookbooks. The names of authors were fed into the search engine and, for each writer, all entities—either Dutch writers or other individuals—that Google returned under the "People also search for"-function were scraped and stored. By using methods derived from network analysis, we then compiled a "canon" of Dutch literature as it emerged through the relationships established by Google's Knowledge Graph. Furthermore, to evaluate the network constructed in this way, a comparison was made between this canon and the academic preferences concerning literary authors described in Dutch literary historiography. This comparison focuses on the gender balance and occurrences of the 3453 authors mentioned in three volumes (covering the period 1800-2005) from the recent nine-volume series on Dutch literary history *Geschiedenis van de Nederlandse literatuur* published between 2006 and 2017.[3] The results allow an assessment of Google's possibilities for constructing alternative hierarchies of canonicity or confirming the approach to the canon

[2] The list was extracted from Wikipedia on 28 October 2017.

[3] Willem van den Berg and Piet Couttenier, *Alles is taal geworden. Geschiedenis van de Nederlandse literatuur, 1800-1900* (Amsterdam: Prometheus, 2009); Jacqueline Bel, *Bloed en rozen. Geschiedenis van de Nederlandse literatuur 1900-1945* (Amsterdam: Prometheus, 2018); Hugo Brems, *Altijd weer die vogels die nesten beginnen. Geschiedenis van de Nederlandse literatuur 1945-2005* (Amsterdam: Prometheus, 2013). The other volumes (covering the period from the earliest instances of Dutch literature to 1800) are Frits van Oostrom, *Stemmen op schrift. Geschiedenis van de Nederlandse literatuur vanaf het begin tot 1300* (Amsterdam: Prometheus, 2006); Frits van Oostrom, *Wereld in woorden. Geschiedenis van de Nederlandse literatuur 1300-1400* (Amsterdam: Prometheus, 2013); Herman Pleij, *Het gevleugelde woord. Geschiedenis van de Nederlandse literatuur 1400-1560* (Amsterdam: Prometheus, 2007); Karel Porteman and Mieke Smits-Veldt, *Een nieuw vaderland voor de muzen. Geschiedenis van de Nederlandse literatuur 1560-1700* (Amsterdam: Prometheus, 2008); Inger Leemans and Gert-Jan Johannes, *Worm en donder. Geschiedenis van de Nederlandse literatuur 1700-1800: de Republiek* (Amsterdam: Prometheus, 2013); Tom Verschaffel, *De weg naar het binnenland. Geschiedenis van de Nederlandse literatuur 1700-1800: de Zuidelijke Nederlanden* (Amsterdam: Prometheus, 2016).

that remains prevalent within the Dutch literary field.

Google's Knowledge Graph

In 2012, Google developers presented a new technology that allowed their search engine to return a summary of the most relevant information on the searched object, which appears next to the list of results. When searching for "Jane Austen," for example, Google returns a summary with pictures of the author, dates and places of birth and death, movie adaptations, siblings, famous quotes, a list of books and a "carousel" of 19 supposedly related historical individuals ("people also search for"). This so-called Knowledge Graph was first developed as a tool for disambiguation: its task was to figure out which meaning of an ambiguous search term the user intended. This advancement of Google's search functionality mainly relies on existing databases and ontologies such as Wikipedia. By aggregating and connecting information from various existing databases with the search behavior of its users, the Knowledge Graph enables Google to make a more informed guess about what information the user needs. The function's name implies that, on top of that connected pile of data, a more abstract level of understanding emerges: "knowledge" instead of information, "things, not strings."[4]

As a company, Google wants us to believe that the implications are far-reaching in the long run: "This is a critical first step towards building the next generation of search, which taps into the collective intelligence of the Web and understands the world a bit more like people do."[5] It needs to be noted, however, that a decade earlier Tim Berners Lee had already claimed to have made this step when he launched the Semantic Web, then referred to as a "brain for humankind."[6] Throughout the early 2000s, several similar initiatives emerged; it is therefore misleading to present the Knowledge Graph as the first or only implementation of the semantic approach to Web technology. Nonetheless, in this contribution, we focus on Google for the primary reason that the Knowledge Graph is arguably the most visible of these initiatives, lending itself easily to the type of experiment described here. It has reached this status because of Google's market dominance in Web search, allowing it to push products such as the Knowledge Graph alongside the basic search engine.

An important factor in the construction of the collective intelligence that emerges

[4] Amit Singhal, "Introducing the Knowledge Graph: Things, not Strings," Official Blog Google, May 16, 2012.

[5] Singhal, "Knowledge Graph."

[6] Cf. Dieter Fensel, and Mark Alan Musen, "The Semantic Web: A Brain for Humankind," *IEEE Intelligent Systems* 16 (2001) 2: 24-25. doi: 10.1109/MIS.2001.920595

through the Knowledge Graph is the connections established between different, somehow related objects or search items. Those connections are made explicit to the user, through the list under "related searches" or "people also search for." Jane Austen, for instance, is related to other highly canonical nineteenth-century British authors such as the Brontë sisters, Charles Dickens, George Eliot, and Oscar Wilde alongside the inevitable archetype of British authorship: Shakespeare (see Ill. 1). At first glance, the technology seems to perform fairly well: the string "Jane Austen" does indeed refer to a historical individual that is arguably quite similar to the historical individual Charlotte Brontë in terms of profession (novelist), canonical status, nationality, gender, periodization et cetera. Through this algorithmic connection of literary authors based on search behavior and information from existing databases, the search engine not only ranks and sorts online information, it also gives meaning to the Web by prioritizing certain relationships over others. Using those relationships, Google contextualizes a single author with chronological, artistic, or other relationships—even if you still need a literary historian to explain the nature of those relations and why they would (not) be meaningful.

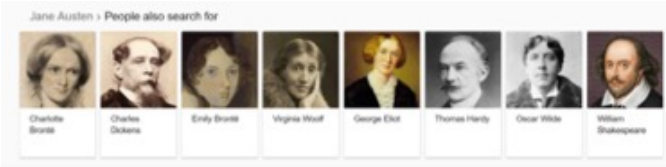

Illustration 1. Google's carousel of related searches for "Jane Austen"

Such carousels of literary authors become subject to the deficits of any other (literary) history or narrative. David Perkins acknowledged that the writing of literary history involves "selection, generalization, organization, and a point of view."[7] Thus, Jane Austen's list of "related searches" raises questions such as: why these authors (selection), what is their common denominator (generalization), why in this order (organization), and what was the rationale behind this grouping (point of view)? In this process of selection and organization, Google is creating hierarchies between authors: a "canon."

Nevertheless, to speak of a "literary canon according to Google" implies a slightly unusual sense of the word "canon." The Oxford Dictionary of Literary Terms defines a literary canon as "a body of writings recognized by authority."[8] In most

[7] David Perkins, *Is Literary History Possible* (Baltimore: John Hopkins UP, 1992), 19.

[8] For the precise definition in the Oxford Dictionary of Literary Terms, see: http://www.oxfordreference.com/view/10.1093/acref/9780199208272.001.0001/acref-9780199208272-e-163?rskey=aw7Jjp&result=163, accessed 4 July, 2019.

(academic) discussions about the canon, the notion often simply refers to "the choice of books in our teaching institutions."[9] The canonical status derived from that recognition and preference is associated with an increased critical and scholarly attention, which has, in turn, resulted in a large body of knowledge about a particular selection of authors throughout the years. The canon according to Google, however, does not necessarily comprise authors recognized (and appreciated) by authority or authors read in teaching institutions. Instead, it includes a selection of authors who occupy the highest positions in the hierarchies constructed by the Knowledge Graph through related searches. Those hierarchies are based on the information about authors available on the Web and the relative amount of search for that information by Google's users. Contrary to the "traditional" canon, the canonical status—or simply "importance"—of a given author in the Knowledge Graph emerges from the volume of searches for information about that author. In other words, the crowd decides through its online behavior.

This contribution is less concerned with the specific technical functionality of the Knowledge Graph's algorithms—which Google is not likely to reveal—than with the relational patterns that become visible when large numbers of authors are queried. Rather, we take a systematic sample of Google's search results and analyze the relationships that Google constructs. Those relationships are analyzed and displayed as networks. The actual relationships are subject to constant change and therefore provide only a snapshot of Google's constantly changing information architecture. They do inform us, however, about Google's function as a broker of information on Dutch literature. The combined related searches give us an impression of what the canon of Dutch literature would look like if it were up to Google. Although this article focuses on the case of Dutch literature, this approach could be replicated for any language field or literature.

The Web and the politics of knowledge

To question the canon of Dutch literature according to Google is relevant for two reasons. The first reason relates to the potential of epistemological and political revolution attributed to the Internet in its early days. Eli Pariser, author of *The Filter Bubble* (2011) recollects that the technological optimism surrounding the emergence of the Internet in the 1990s resulted in the belief that an "inevitable, irresistible revolution was just around the corner, one that would flatten society,

[9] Harold Bloom, *The Western Canon: The Books and School of the Ages* (San Diego, CA: Harcourt Brace 1994), 15.

unseat the elites, and usher in a kind of freewheeling global utopia."[10] That technological optimism resonated with historical media revolutions, from the printing press to radio and television, which had redefined power structures and diminished the position of Church and state as information mediators.[11] The Web provided a cheap and global platform that bypassed conventional communication channels controlled by governments, newspapers, and publishing houses.

Larry Sanger, one of the co-founders of Wikipedia, argued that the revolution of the Web 2.0—its emergence as a participatory, social medium—has introduced a new "politics of knowledge" and an "epistemic egalitarianism" for which he considered Wikipedia the optimal vehicle.[12] It is not difficult to imagine the expected political consequences of that new *episteme*: many were quick to explain contemporary political revolts by pointing to the "democratization of information" that Web 2.0 and social media were said to enable. The Iranian "Twitter revolution" in 2009 and the presumed role of Facebook during the Arab Spring in 2010 and 2011 were regarded as milestones in that development.[13] Others even argued that the Web's democratic potential is rooted in its technological blueprint: the very algorithm that has been crucial to Google's success, PageRank—the mechanism that prioritizes Web pages according to links ("upvotes") to those pages from other important web pages—was deemed to "[utilize] the uniquely democratic structure of the web."[14]

However, this optimistic perspective on the supposed democratizing power of the Web soon became subject to criticism. Pariser pointed out the dramatic effects of the centralization of information distribution at a handful of large companies, Google and Facebook in particular. He argued that the tendency to personalize search results and news feeds according to user data has resulted in a "filter bubble," which effectively blocks information sources that do not confirm what the user already knows. This technological censorship would eventually pose a threat to democratic societies, Pariser insisted.[15] A similar reservation was articulated by Evgeny Morozov, who analyzed the dominant myth about the presumed relationship between democratization and Internet technology: a fallacy he called the "net delusion."[16] In complete opposition to that myth, Mo-

[10] Eli Pariser, *The Filter Bubble: What the Internet is Hiding from You* (London: Penguin Books, 2011), 7.

[11] Pariser, *Filter Bubble*, 37; Larry Sanger, "Who Says we Know. On the Politics of Knowledge," *Edge* 2007, accessed 4 July, 2019.

[12] Sanger, "Who says."

[13] Cf. Evgeny Morozov, *The Net Delusion. How Not to Liberate the World* (London: Penguin Books, 2012), 4.

[14] Cited in Pariser, *Filter Bubble*, 37.

[15] Pariser, *Filter Bubble*,31.

[16] Morozov, *Net Delusion*,xvii.

rozov warned against the repressive use of Internet technology by authoritarian states and underscored the Web's powerful potential for surveillance. That criticism proved to be visionary and reached a wide consensus in the public debate after cases such as the PRISM surveillance program (revealed by Edward Snowden) and the Cambridge Analytica case involving the leak and abuse of Facebook user data. In addition to political abuse of web technology, Shoshana Zuboff recently made clear that surveillance also became key to the commercial strategies of various web companies, treating personal data as their main commodity.[17] Powerful Web companies and political states have become the new information intermediaries that determine who gets to know what. They complicate the once unequivocal idea that the Internet could empower regular users and enforce new knowledge structures.

In contrast to this polarized scheme where Web technology is mainly viewed in either utopian or dystopian terms, various initiatives and studies preferred to evaluate the Web within a critical framework, accounting for both its epistemological opportunities and risks.[18] Organizations like the Association for Progressive Communication monitor and promote Internet access and Internet freedom worldwide, especially for women and minorities, following the association's first axiom that "the Internet is an enabler of human rights, development and social justice, including gender justice" and "a global public resource that has transformed human communications and behavior and that challenges existing structures of power, including gender-based power."[19] Such views inform, and are in turn informed by, critical studies of Web platforms revealing (for instance) the considerable gender gap in Wikipedia's editor base and content,[20] or the gendered patterns of book consumption and recommendation on Amazon.[21] Additionally, in her book *Algorithms of Oppression. How Search Engines Reinforce Racism* (2018), Safiya Umoja Noble convincingly argued that biases—or plain racism and sexism—and the centralization of power among a few large Web companies have shaped the representation and discoverability of women of color on the Web. The Web's potential for social and epistemological change is

[17] Shoshana Zuboff, *The Age of Surveillance Capitalism. The Fight for a Human Future at the New Frontier of Power* (London: Profile Books, 2019.)

[18] E.g.: Merlyna Lim, "Challenging Technological Utopianism," *Canadian Journal of Communication* 43 (2018): 375-379.

[19] 'APC's Theory of Change 2016-2019', accessed 4 July, 2019, https://www.apc.org/en/apcs-theory-change-2016-2019.

[20] Heather Ford & Judy Wajcman, "'Anyone Can Edit,' Not Everyone Does: Wikipedia's Infrastructure and the Gender Gap," *Social Studies of Science*, 47 (2017) 4: 511-527; Claudia Wagner, David Garcia, Mohsen Jadidi, Markus Strohmaier, "It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia," *The International AAAI Conference on Web and Social Media* (ICWSM2015).

[21] Doina Bucur, "Gender Homophily in Online Book Networks," *Information Sciences* 481 (May 2019): 229-243.

now widely accepted; however, platforms like Wikipedia, Amazon, and Google still have much room for improvement with regards to the accessibility and neutrality of their content.

What are the implications of this discussion for the case of the Knowledge Graph's representation of literary history? One thesis could be that the promise of a democratization of knowledge would provide an opportunity for the construction of an alternative literary canon or even literary history. After all, the canon has traditionally been established and reconfirmed by centralized literary and academic institutions. Platforms such as Wikipedia now enable a decentralization and diversification of literary knowledge. The design of Google's Knowledge Graph—the connection of search behavior to various information sources on the web—arguably subjects the judgment of the relevance of information to the popular vote. However, informed by the abovementioned notions of "filter bubble," "net delusion" and "algorithms of oppression," the *antithesis* would object to this promise and argue that Google's Knowledge Graph merely reproduces or even reinforces the same knowledge structures and biases that have informed literary history for decades.

Literary canons contested

The second reason why it is relevant to study the canon of Dutch literature according to Google is that it can offer a new perspective to the specific debates on canonization among Dutch and international literary scholars. There have been various attempts to abandon the canon and study a larger, more diverse corpus of literature. Some scholars employed the notion of "middlebrow" in order to foreground literature that had been neglected for centuries due to the exclusive focus on "highbrow" literary texts.[22] Many have stressed the value of digital methods, digital corpora, and bibliographic collections that are now available for this purpose.[23] Others criticized the canon by pointing at the ethnic and gender

[22] For example: Jaime Harker, *America the Middlebrow: Women's Novels, Progressivism, and Middle-brow Authorship between the Wars* (Amherst/ Boston: University of Massachusetts Press, 2007); Erica Brown, (ed.), "Investigating the Middlebrow," special issue of *Working Papers on the Web* 11 (2008); Erica Brown, and Mary Grover (eds.), *Middlebrow Literary Cultures: the Battle of the Brows, 1920-1960* (London: Palgrave 2011); Alicia Montoya, "Middlebrow, Religion, and the European Enlightenment. A New Bibliometric Project, MEDIATE (1665-1820)," *French History and Civilization* 7 (2017): 66-79.

[23] For example: Franco Moretti, "Conjectures of World Literature," *New Left Review* 1, January-February 2000; Franco Moretti, *Graphs, Maps, Trees. Abstract Models for a Literary Theory* (New York: Verso, 2005); Marc Algee-Witt, and Mark McGurl, "Between Canon and Corpus: Six Perspectives on 20th-Century novels," *Stanford Literary Lab Pamphlet* 8 (January 2015); Joep Leerssen, "De

imbalance on academic reading lists: a quarrel that culminated in the so-called Canon Wars. Toni Morrison, among others, recognized the white and male literary canon and imagination as a reflection of the sexist and racist elements of Western culture.[24] Indeed, Morrison and others have underscored the role of literary institutions that establish and confirm this literary canon. Their criticism elicited a strong reply from defenders of the idea that literary quality is universal and objective. Allan David Bloom warned against the moral ramifications of cultural relativism and postmodernism, and Harold Bloom in turn refuted the instrumentalization of the canon in what he considered to be a "program for social salvation."[25]

While the Canon Wars never reached the same vehemence on the Dutch literary field, the debate did coincide with a scholarly movement that studied the predominant position of (mainly) male and highbrow authors in Dutch literary historiography.[26] These studies questioned the institutions and gatekeepers within the literary system that excluded and devalued female authors. The critical and scholarly debate about the diversity of the literary canon has been on the agenda ever since. Donadio observed that the "multiculturalists" and the feminists are often taken to be the winners of the Canon Wars[27] and yet the various recent Dutch protests about the lack of ethnic and gender diversity and representation complicate that view.[28] In a controversial essay from 1997, Anil Ramdas criticized

Nederlandse canon langs de digitale meetlat," Lecture delivered at the National Library of the Netherlands, May 14, 2018.

[24] Toni Morrison, *Playing in the Dark. Whiteness and the Literary Imagination* (Cambridge, MA: Harvard UP, 1992).

[25] Allan David Bloom, *The Closing of the American Mind* (New York City, NY: Simon & Schuster 1987), 353; Bloom, *American Mind*, 29.

[26] For example: Maaike Meijer, *De lust tot lezen. Nederlandse dichteressen en het literaire systeem* (Amsterdam: Sara/Van Gennep, 1988); Maaike Meijer, and Ernst van Alphen, *De canon onder vuur. Nederlandse literatuur tegendraads lezen* (Amsterdam: Van Gennep, 1991); Riet Schenkeveld, *Met en zonder lauwerkrans. Schrijvende vrouwen uit de vroegmoderne tijd 1550-1850*, (Amsterdam: Amsterdam UP, 1997); Agnes Verbiest, "Een porseleinkast in de jungle. De verwoording van wetenschappelijke teksten over (het werk van) vrouwen," *Tijdschrift voor Nederlandse Taal- en Letterkunde* 111 (1995) 1: 117-126; Marianne Vogel, *'Baard boven baard'. Over het Nederlandse literaire en maatschappelijke leven 1945-1960* (Amsterdam: Van Gennep, 2001); Erica van Boven, Koen Rymenants, Mathijs Sanders, and Pieter Verstraeten, "*Middlebrow* en modernisme," *TNTL* 124 (2008) 4: 304-311; Lenny Vos, *Uitzondering op de regel. De positie van vrouwelijke auteurs in het naoorlogse Nederlandse literaire veld* (Dissertation University of Groningen, 2008); Jacqueline Bel, and Thomas Vaessens (red.), *Schrijvende vrouwen. Een kleine literatuurgeschiedenis van de Lage Landen 1880-2010* (Amsterdam: Amsterdam UP, 2010).

[27] Rachel Donadio, "Revisiting the Canon Wars," *The New York Times*, November 16, 2007.

[28] For example: Lezeres des Vaderlands, "Toverstaf en tijdmachine. Hoe ons literatuuronderwijs vrouwen buitensluit," *De Groene Amsterdammer*, April 20, 2016. Jannah Loontjens, "De literatuur is achtergebleven in het masculiene tijdperk," *NRC Next*, May 19, 2016; Marja Pruis, "En de vrouw, zij schreef voort. Voorbij mannelijke zelfvergroting en vrouwelijke wissewasjes in de literatuur," *De*

the dominant white gaze among Dutch authors in their representation of non-white characters.[29] Almost twenty years later, in 2015, both Karin Amatmoekrim and Ebissé Rouw concluded that not much had changed since Ramdas's publication: the Dutch literary field remained a segregated space where opportunities for publication, literary awards, and attention were less accessible to authors with a migrant background.[30] In addition to these contributions to the public debate, recent studies have pointed to the conditions and the possible effects of a Dutch tradition of gendered differences in literary value on the Dutch literary field. Corina Koolen demonstrated, for instance, by means of a questionnaire distributed among a large group of Dutch readers that works by female authors are less likely to be considered of high literary value: her thesis also rejects the supposed existence of general significant differences between male and female literary style.[31] Secondly, Van der Deijl et al. 2016 alluded to the relationship between the gender of the author and the gender of the characters in the narratives they produce: this, they proposed, was a possible explanation for the unequal gender roles they reported among the characters from a corpus of 170 recent Dutch novels.[32]

Meanwhile, various attempts have been undertaken to establish a new canon of Dutch literature. In 2002, the Society of Dutch Literature (*Maatschappij der Nederlandse Letterkunde*) published a list of 108 authors and 125 works of literature that 299 of its members considered "classics of Dutch literature."[33] The list is populated by many iconic Dutch authors such as Multatuli, Joost van den Vondel and

Groene Amsterdammer, June 13, 2018; "Mogen vrouwen ook iets zeggen als het over vrouwen gaat?," *NRC Handelsblad*, June 17, 2018.

[29] Anil Ramdas, "Moedwil en kwade trouw bij blanke schrijvers. Niemand heeft oog voor het vreemde," *NRC Handelsblad*, March 14, 1997.

[30] Karin Amatmoekrim, "Een monoculturele uitwas. De ondraaglijke witheid van de Nederlandse letteren." *De Groene Amsterdammer*, August 20, 2015; Ebissé Rouw, "Literatuur blijft te wit," *NRC Handelsblad*, May 16, 2015.

[31] Corina Koolen, *Reading Beyond the Female. The Relationship between Perception of Author Gender and Literary Quality* (Dissertation University of Amsterdam 2018), 103. It is important to add that Koolen questions the ethical and methodological validity of gender classification altogether in literary studies.

[32] Lucas van der Deijl, Saskia Pieterse, Marion Prinse, and Roel Smeets, "Mapping the Demographic Landscape of Characters in Recent Dutch Prose: A Quantitative Approach to Literary Representation," *Journal of Dutch Literature* 7 (2016) 1: 20-42. Cf. Ted Underwood, David Bamman, and Sabrina Lee, "The Transformation of Gender in English-Language Fiction," *Cultural Analytics* (February 13, 2018). DOI: 10.7910/DVN/TEGMGI for a similar approach to English fiction. http://culturalanalytics.org/2018/02/the-transformation-of-gender-in-english-language-fiction/ Also see Eve Kraicer and Andrew Piper, "Social Characters: The Hierarchy of Gender in Contemporary English-Language Fiction," *Journal of Cultural Analytics.* January 30, 2019. DOI: 10.31235/osf.io/4kwrg.

[33] For an overview of the selection procedure for the "Canon of Dutch Literature," see http://www.dbnl.org/letterkunde/enquete/enquete_dbnlmnl_21062002.htm#26 (in Dutch), accessed 4 July, 2019.

the so-called "Great Three" (W.F. Hermans, Harry Mulisch, Gerard Reve). This selection of classical works of Dutch literature is again gendered: female authors appear only from the 19th position downwards and the general male-female ratio is out of balance, which the critics did not fail to notice.[34] Another, more recent initiative to establish the Dutch-Flemish literary canon also sparked academic and critical debates.[35] This canon was composed by two Flemish institutes with an important position in the literary field, The Royal Academy of Dutch Language and Literature (*Koninklijke Academie voor Nederlandse Taal- en Letterkunde*) and the Flemish Literature Fund (*Vlaams Fonds voor de Letteren*), and consists of a list of 51 "essential works of Dutch literature."[36] Yet again, one of the main points of criticism raised was the bias in the selection regarding gender and ethnicity.[37] In other words, the scholarly attention, canon criticism and canon revision from the past three decades have not changed the Dutch canon fundamentally.

An additional, quantitatively informed image of the established Dutch canon can be obtained from the nine-volume series *Geschiedenis van de Nederlandse literatuur* (*History of Dutch Literature*) published between 2006 and 2017, which documents the history of literature written in Dutch from the oldest extant sources to 2005. The individual volumes were (co-)authored by leading academic experts, were published by the leading publishing house Prometheus/Bert Bakker, and were commissioned by the *Taalunie*, an organization that promotes and develops policies concerning the Dutch language and literature. The value of this major project within the field of Dutch literature for both academic research and education has been widely acknowledged.[38] Because of its authoritative and synthesizing function, this particular literary history offers a useful source for a reconstruction of canonical hierarchies in Dutch academic literary historiography.

In order to compare the relationships between Dutch authors as determined by Google's algorithms against the information captured in the *Geschiedenis van de Nederlandse literatuur* (hereafter: *GNT*), all 3836 entries that refer to a natural person were extracted from the registers of the three volumes covering the period 1800-2005.[39] For each name, the number of pages featuring the name was

[34] Pieter van Os and Sander Pleij, "Het deprimerende dogma van de canon," *De Groene Amsterdammer*, March 15, 2003.

[35] For an overview of relevant debates regarding this particular canon see http://literairecanon.be/nieuws/debat-over-de-literaire-canon (in Dutch), accessed 4 July, 2019.

[36] All Dutch phrases have been translated to English by the authors.

[37] Laurens Ham, "Sluit de Canon niet op in zijn ark. Een reflectie op de Dynamische Canon van de Nederlandstalige literatuur," *Ons Erfdeel* (2015) 4: 4-13.

[38] C.f. Lars Bernaerts, and Youri Desplenter, "Rijper, wijzer, smaakvoller. Impulsen voor het literatuuronderwijs," *Nederlandse letterkunde* 23 (2018) 3: 223-234.

[39] The registers were kindly made available digitally to the authors by the publisher Prometheus/Bert Bakker. They were extracted from: Willem van den Berg, and Piet Coutte-

used as an indication of occurrence and, by extension, of importance. Removal of duplicates (names occurring in different volumes) resulted in a list of 3453 distinct individuals with a male/female/unknown ratio of 83.4% / 13.8% / 2.8%.[40] Male individuals are both mentioned more often and discussed in more depth *if* they are mentioned: the average number of occurrences (number of pages that mention his / her name) for male individuals equals 4.5 pages (SD = 8.6); female individuals are mentioned on 3.1 pages on average (SD = 4.7). This cursory examination of the canon according to the *GNT* thus confirms the prominence of male authors in conventional literary history.

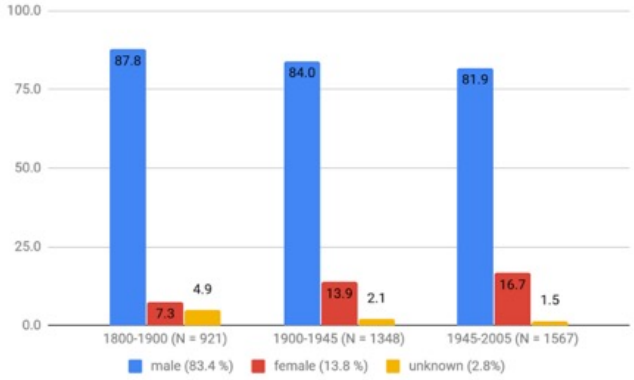

Figure 1. Male-female ratio in the registers of Geschiedenis van de Nederlandse literatuur *vol. 5, 6, and 7 (1800-2005)*

The rankings of most frequently mentioned authors provide another indication of the prominence of male authors in the *GNT* volumes. The top 20 is dominated by male individuals (see Table 1) and the highest ranked female authors range from rank 20/21 to rank 264 (see Table 2).

Rank	Name	Gender	Occurrences
1	Braak, Menno ter	male	116
2	Verwey, Albert	male	100
3	Nijhoff, Martinus	male	97
4	Ostaijen, Paul van	male	80

nier, *Alles is taal geworden. Geschiedenis van de Nederlandse literatuur 1800-1900* (Amsterdam: Prometheus/ Bert Bakker, 2009); Jacqueline Bel, *Bloed en rozen. Geschiedenis van de Nederlandse literatuur 1900-1945* (Amsterdam: Prometheus/ Bert Bakker, 2015); Hugo Brems, *Altijd weer vogels die nesten beginnen. Geschiedenis van de Nederlandse literatuur 1945-2005* (Amsterdam: Prometheus/ Bert Bakker, 2013).

[40] Gender was manually assigned to each individual. For a small portion of the names, the corresponding gender could not be retrieved from either the *GNT* or the Web and therefore remains unknown.

Rank	Name	Gender	Occurrences
5	Buysse, Cyriel	male	79
6	Couperus, Louis	male	75
7	Woestijne, Karel	male	75
8	Kloos, Willem	male	73
9	Claus, Hugo	male	72
10	Conscience, Hendrik	male	72
11	Perron, Eddy du	male	72
12	Eeden, Frederik van	male	67
13	Walschap, Gerard	male	67
14	Gezelle, Guido	male	64
15	Vermeylen, August	male	63
16	Busken Huet, Conrad	male	62
17	Bilderdijk, Willem	male	61
18	Marsman, Hendrik	male	61
19	Boon, Louis Paul	male	59
20	Deyssel, Lodewijk van (Karel Alberdingk Thijm)	male	57

Table 1. Top 20 most frequently mentioned authors in the Geschiedenis van de Nederlandse literatuur *vol. 5, 6, and 7 (1800-2005)*

Rank	Name	Occurrences
21	Roland Holst-van der Schalk, Henriette	57
46	Bruggen, Carry van (Carolina Lea de Haan)	36
72	Loveling, Virginie	27
76	Blaman, Anna	26
105	(Bosboom-)Toussaint, Anna Louisa Geertruida (Truitje)	22
109	Haasse, Hella S.	22
111	Swarth, Hélène	22
132	Wit, Augusta de	20
144	(Ackere-)Doolaeghe, Maria van	18
145	Antink, Margo	18
160	Herzberg, Judith	17
193	Boudier-Bakker, Ina	15
200	Marissing, Lidy van	15
202	Meijer, Maaike	15
216	Naeff, Top	14
235	Mutsaers, Charlotte	13
242	Belpaire, Maria	12
259	Paemel, Monika van	12
264	Vasalis, M.	12

Table 2. Top 20 most frequently mentioned female authors in the Geschiedenis van de Nederlandse literatuur *vol. 5, 6, and 7 (1800-2005)*

An obvious explanation for the male dominance in these volumes can be found in the fact that literary history is (in itself) also a history of gender inequality. Published female authors have been a minority throughout the studied period and even today the scale remains out of balance: Koolen estimated a 60-40% male-female ratio among the authors of all 5,842 Dutch "literary novels" documented between 2007-2012.[41] While female authors arguably became more visible in the most recent periods of Dutch literary history, reflected in the literary historiography (see Figure 1), Lenny Vos estimated that the number of female authors published at major publishing houses have, in fact, declined since

[41] Koolen, *Reading Beyond the Female*, 40-41. The data were provided by the National Library of the Netherlands. The genre classification was based on bibliographic labels called NUR-codes, which are applied by the publishers.

the 1960s.[42] Underwood, Bamman and Lee observed a similar decline regarding English-language fiction published between 1800 and 1960: the emancipation of female authors does not always progress in a linear fashion.[43] Moreover, Maaike Meijer argued that traditional historical accounts have often been preoccupied with a select number of key figures (and literary movements) who are considered to exemplify the literary history of a given period and who are often male.[44] The *GNT* embraced critical and corrective studies like Meijer's[45] but the possibilities for a synthesizing overview like the *GNT* to correct and complement decades of literary historiography, including its biases and preoccupations, are simply limited. In our discussion of the hierarchies constructed by Google's algorithms, this bias in literary history and academic historiography will remain a point of comparison: a knowledge tradition that possibly conditions search behavior by users and the knowledge about Dutch authors that circulates on the Web.

Data, method, and limitations

The sample of authors used in this study comprises all Wikipedia entries featuring the category "Dutch writer" (*Nederlandse schrijver*). These authors are regarded as nodes in a network of related Google searches. For each of these nodes, metadata such as name, gender and birth year were extracted semi-automatically from their entries by using Wikipedia's (Dutch) API and through manual correction.[46] Subsequently, edges were drawn between these nodes by entering all author names in the Google search engine. For each node, all "related searches" returned by the search engine were semi-automatically scraped and stored.[47] When a given person X occurred in the list of related searches of author Y then a relation was defined between X and Y. The number of related searches ranges from 0 to a maximum of 25, which is imposed by Google. The maximum number of edges per node in the network is thus 25. The position of X in the related searches list of Y determined the weight of that relation. For instance, when X occurred at the first position in Y's related searches, a heavier weight was ascribed

[42] Lenny Vos, *Uitzondering op de regel. De positie van vrouwelijke auteurs in het naoorlogse Nederlandse literaire veld* (Dissertation University of Groningen, 2008).

[43] Underwood et al. "The Transformation of Gender in English-Language Fiction."

[44] Meijer, *De lust tot lezen*, 295-296.

[45] Note the high number of references to Meijer's work in Table 2.

[46] https://nl.wikipedia.org/w/api.php

[47] The task could not be automated entirely, mainly because Google no longer maintains its API for automatically retrieving related searches based on a query. "Semi-automatically" here means that we used an interface that extracted the names of the related persons from the html-block of the carousel, which had to be copied into the interface manually.

to edge X-Y than when X occurred at the last position. The network thus constructed enabled further analysis.

Subsequently, edges were drawn between these nodes by entering all author names in the Google search engine. For each node, all "related searches" returned by the search engine were semi-automatically scraped and stored.[48] When a given person X occurred in the list of related searches of author Y then a relation was defined between X and Y. The number of related searches ranges from 0 to a maximum of 25, which is imposed by Google. The maximum number of edges per node in the network is thus 25. The position of X in the related searches list of Y determined the weight of that relation. For instance, when X occurred at the first position in Y's related searches, a heavier weight was ascribed to edge X-Y than when X occurred at the last position. The network thus constructed enabled further analysis.

There are at least three biases that affect the validity and representativeness of the sample. First, the assignment of the label "Dutch writer" was made by various Wikipedia editors and their criteria for qualifying an individual as such is neither transparent nor consistent. As a result, some individuals who are in fact published Dutch writers are not recognized in the dataset and vice versa. Furthermore, the label is not reserved for authors of literary prose or poetry only: it also applies to journalists, historians, essayists and so on. Secondly, in this approach, only the related searches of the 2,287 authors were stored. The creation of a threshold was unavoidable, since the number of nodes would increase rapidly if the related searches of people who occurred as related searches were also included. However, this creates a blind spot in the network: it is perfectly possible that an author was not categorized as Dutch writer on Wikipedia but nevertheless would be central according to Google's logic. The selected approach does not account for these cases. Thirdly, the sample shows an unequal gender distribution and a strong preference for modern and contemporary writers (see Figure 1). These biases presumably reflect the overall emphasis in Wikipedia entries in terms of gender and periodization[49] but the hypothetical deviation from that trend in this sample cannot be assessed easily.

[48] The task could not be automated entirely, mainly because Google no longer maintains its API for automatically retrieving related searches based on a query. "Semi-automatically" here means that we used an interface that extracted the names of the related persons from the html-block of the carousel, which had to be copied into the interface manually.

[49] There have been attempts to balance out the gender imbalance in Wikipedia entries. Atria, a Dutch institute for emancipation and women's history, organizes events to write entries for famous women who are absent from Wikipedia. See: https://www.atria.nl/nl/agenda/schrijf-mee-voor-wikipedia (in Dutch), accessed 9 August, 2018.

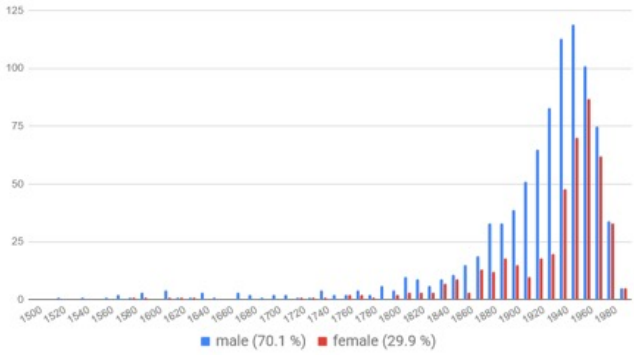

Figure 2. Number of authors (N = 2,287) by birth decennium

Aside from the biases in the sample, the data obtained in this method are unstable because Google's search results are variable over time. Van den Bosch et al. 2016 demonstrated that the search results returned by Google varied significantly when a fixed set of search terms were queried over a period of nine years.[50] To estimate the bias introduced by this variability, we repeated the initial data collection nine months later using the same sample of 2,287 Dutch writers.[51] A comparison between both measurements (M1 and M2) confirms the variability of search results observed by Van den Bosch et al. 2016. In only 46.1% of the search queries, Google returned an identical list of related searches in M1 and M2. A first explanation for that variability is that Google acquired more information on search behavior during the nine months between both measurements: M1 resulted in 5,863 unique names whereas 6,300 names were found in M2. This result implies that the average number of related searches per query increased (from 2.6 to 2.8). However, it is not the case that Google simply has *more* information at its disposal. That information has also changed: 20.4% of the names observed in M1 did not recur in the search results from M2 and in 15.4% of the queries, the list of related searches was longer in M1 than in M2. Nevertheless, the variability of the search results is limited: in the majority of the cases many (if not all) of the names from M1 were also returned in the list of related searches obtained in M2,

[50] Antal van den Bosch, Toine Bogers, & Maurice de Kunder, "Estimating Search Engine Index Size Variability: A 9-year Longitudinal Study," *Scientometrics* 107 (2016) 2: 839-856. Furthermore, Google's search functionalities pose various limitations to linguists who use the search engine as an entry point to large volumes of language data (cf. Adam Kilgariff, "Googleology is Bad Science," *Computational Linguistics* 33 (2007) 1: 147-151.). These constraints compromise the usability of the data for (longitudinal) comparative approaches or for using Google as a source for monitoring (linguistic) trends over time.

[51] The first measurement took place on 28 & 29 October and 4 & 5 November 2017, and the second on 14, 15, 20, 21, 22, 23, 24, 27 and 30 August 2018.

with an average overlap per query of 64.0%. Therefore, it is unlikely that future measurements would result in a fundamentally different image, even though the bias introduced by the variability is considerable and complicates generalizations based on the data reported in this study.

A final potential problem for the usability of Google's search results is caused by Google's personalized search. To reduce the possible influence of personalized search results, all cookies and cache were removed from the browser used to query the author's names. Furthermore, the potential impact of personalization on the data was evaluated in a small experiment in which a sample of 100 authors were simultaneously queried and all related searches that Google returned were stored on different machines at different locations (all cookies and cache were removed in advance). The inter-annotator agreement of that experiment turned out to be 100.0%. In other words, personalized search did not affect the specific results reported in this study. This does not mean, of course, that personalized search (including the moment and location of the query) would not have any effect on the specific related searches that a given user would get. This experiment simply provides a guarantee that the patterns observed in our data should be attributed to the information available in the Knowledge Graph at the moment of query, rather than to the specific (search settings of the) user who happened to collect the data.

Results I: network centrality

The resulting directed, weighted network consists of 5,863 nodes that represent people labeled as "Dutch writer" on Wikipedia and 6,242 edges that represent their related searches on Google (see Graph 1), with an average degree of 1.065. A first observation is that the network is poorly connected. Python's NetworkX package reports a network density of 0.0002, which is probably due to the high number of disconnected components. This also explains the relatively low average degree of 1.065: 1,357 people only appear once in the observed related searches, which is 23.1% of the total number of nodes. In Figure 2, graph 1 (left), these nodes comprise the majority of the nodes in the peripheral layers of the network.

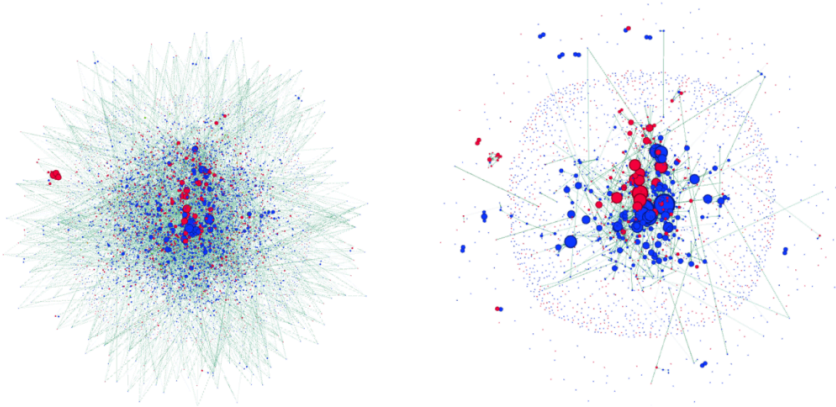

Figure 3. Network visualizations. Graph 1 (left): complete network (N = 5,863). Graph 2 (right): filtered network with Dutch writers only (n = 2,287). Node size indicates PageRank value, node color indicates gender (red = female, blue = male, grey = unknown)

The large number of nodes not labeled with the Wikipedia category "Dutch writer" covers a substantial part of the total network (60.7%, 3,557 nodes). When those nodes are filtered out, a clearer image arises of the Google network's center and periphery (see Figure 3, Graph 2, right).

What stands out is the large number of isolates in the filtered right graph (58.9%, 1,346 nodes), which means that Google does not connect the majority of the Dutch writers to any of the other Dutch writers. This indicates that Google's Knowledge Graph tends to relate people labeled as Dutch writer to people who are not labeled as such but who are, instead, writers from other countries: actors, politicians, TV celebrities, athletes, et cetera. Google's representation of the Dutch literary field is thus populated by people who are generally marginal if not absent in books on literary history, which is illustrated by a cursory examination of the *GNT* registers.

Which authors are most central in the network? The centrality measure that is most suitable to represent a node's importance in this context is PageRank,[52] the same algorithm Google's search engine uses, among other algorithms, to rank Web pages by relevance. PageRank is based on the seemingly circular assumption that a node in a network becomes more important when it is connected to other important nodes. The importance of a node is not dependent on the num-

[52] Sergey Brin, and Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Computer Networks and ISDN Systems* 30 (1998): 107-117.

ber of links but on the relative importance of incoming links as a node gets upvoted with the PageRank value of an incoming link.[53] For the present network of Dutch writers, these upvotes are a suitable indication of importance in the overall network as important Dutch writers are likely to be featured in the top related searches of other important people in the network.

Degree centrality is a less insightful metric for the operationalization of canonicity.[54] Because of the low variance in the number of ingoing relationships (in-degree) and outgoing relationships (out-degree) and the large group of people with the maximum of 25 relationships that Google returns, degree centrality is not able to discriminate sufficiently between important and less-important authors. Closeness centrality is also not applicable to this network due to the high number of disconnected components in the graph.[55] The PageRank algorithm is more appropriate because it is able to handle a high number of disconnected components and because it is not dependent on the degree of individual nodes only. We used Python's NetworkX package in combination with Gephi to produce a ranking of the top 20 nodes with the highest PageRank value (see Table 3).

Rank	Name	Gender	Community	PageRank
1	Harry Mulisch	male	Literary authors	0.000697
2	Remco Campert	male	Literary authors	0.000673
3	Arnon Grunberg	male	Literary authors	0.000617
4	Saskia Noort	female	Popular authors	0.000563
5	Arend van Dam	male	Children' s authors	0.000539
6	Gerda van Wageningen	female	Class 1323	0.000538
7	Leon de Winter	male	Literary authors	0.000504
8	Tommy Wieringa	male	Literary authors	0.000499
9	Connie Palmen	female	Literary authors	0.000486
10	Maarten 't Hart	male	Literary authors	0.000484
11	Esther Verhoef	female	Popular authors	0.000482
12	Herman Koch	male	Literary authors	0.000473
13	Julia Burgers-Drost	female	Class 1323	0.000469
14	Susan Smit	female	Popular authors	0.000464
15	Ronald Giphart	male	Literary authors	0.000461
16	Gerard Reve	male	Literary authors	0.000458
17	Marion Pauw	female	Popular authors	0.000455
18	Godfried Bomans	male	Literary authors	0.000448
19	Mark Janssen (illustrator)	male	Class 135	0.000443
20	Henny Thijssing-Boer	female	Class 83	0.000442

Table 3. Top 20 of nodes in the Google networks with highest PageRank value. Com-

[53] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. "The PageRank Citation Ranking: Bringing Order to the Web," Technical Report. Stanford InfoLab (1998).

[54] Degree centrality is based on the assumption that a node gets more important when it has more relations to other nodes in the network. See Linton Freeman, "Centrality in Social Networks: Conceptual Clarification," *Social Networks* (1978) 1: 215-239, for a discussion of the pros and cons of basic centrality measures for network analysis, and Tore Opsahl, Filip Agneessens, and John Skvoretz, "Node Centrality in Weighted Networks: Generalizing Degree and Shortest Paths," *Social Networks* 32 (2010): 245-251 for a similar discussion on weighted graphs.

[55] Closeness centrality is based on the assumption that a node gets more important when it is nearer to all other nodes in the network.

munity categories are based on the clustering by Gephi's modularity functionality.

This top 20 ranking contains various authors whose canonical status in the Dutch literary reception is quite uncontroversial. Two authors of the so-called "Great Three" of postwar Dutch literature are present: Harry Mulisch (position 1) and Gerard Reve (position 16).[56] The second position is occupied by Remco Campert (1929-), the last surviving author of an iconic, postwar poetic movement called *de Vijftigers* ("Those of the Fifties"). Additionally, Arnon Grunberg (1971-: position 3) can be considered as one of the highest acclaimed living authors. It might not be surprising that these authors occupy key positions in the network as Google's Knowledge Graph is partly based on what is available in existing online databases. There are numerous Web articles written about the Great Three, Remco Campert, and Arnon Grunberg, whose publications and public performances attract wide attention from the media. Furthermore, Remco Campert might be on the top of the list simply because of the length of his writing career. For several decades, Campert produced a number of weekly columns in different daily and weekly papers. The same applies to the extremely productive Grunberg: besides his novel production he regularly publishes in several periodicals and newspapers. Most of these articles are also published online, which subsequently feeds back into the Knowledge Graph.

Mulisch, Reve, Campert, and Grunberg are commonly considered to be highly esteemed, canonical authors, which is also reflected in the *GNT* (they are discussed on 50, 46, 26 and 8 pages respectively). Their central place in the ranking suggests that Google reproduces certain canonical mechanisms: authors who are appreciated by the traditional literary institutions (literary supplements, prize juries, publishers etc.) have a good chance of ending up as important nodes in this network. That mechanism, however, does not explain the high positions of authors in this top 20 without an acclaimed position in the literary field. This applies to authors like Saskia Noort (position 4), Esther Verhoef (position 11), Susan Smit (position 14) and Marion Pauw (position 17). Saskia Noort, Esther Verhoef and Marion Pauw are all bestselling authors of literary thrillers. Susan Smit, Saskia Noort and Marion Pauw have been part of the Writers on Heels initiative, launched in 2005, which advocated an intermediate position between high, "heavy" literature and chick lit.[57] This self-proclaimed new movement was met with resistance by Dutch critics.[58] As none of them are included in the *GNT*

[56] "The Great Three" of Dutch postwar literature refer to Harry Mulisch (1927-2010), Gerard Reve (1923-2006) and Willem Frederik Hermans (1921-1995).

[57] https://www.youtube.com/watch?v=nXzTveBESIE, accessed 5 April, 2018.

[58] E.g. Vrouwkje Tuinman, "Writers on Heels," *NRC Handelsblad*, September 30, 2005, accessed 4 July, 2019; Herman Franke, "Hooggehakte schrijvers op oorlogspad," *De Volkskrant*, October 14, 2005, accessed 4 July, 2019.

registers, it is particularly interesting that these authors *do* show up among the most central individuals in the Google network. Their position in the ranking based on PageRank possibly indicates the Internet's alleged democratic potential: unhindered by preconceptions of what good literature should be, the Knowledge Graph connects authors with one another in case they end up co-occurring in the search behavior of individuals. Despite their high sales figures and large reading audiences, authors such as Noort, Verhoef, Pauw, and Smit hardly ever show up on lists of the Dutch literary canon composed by experts in the field, nor are they discussed in an academic literary history like the *GNT*. Google, however, follows different rules. The Knowledge Graph operates bottom-up and is data-driven, taking into account everything available in terms of what is on the Web and how people search through this.

Do we observe a gender bias in this bottom-up view of canonicity? In order to find an answer to that question, we computed a Pearson correlation coefficient to assess the relationship between the gender of the authors and their PageRank score. There appeared to be no correlation between these two variables, r=0.010, n=5720 , p=0.456. Also, a linear regression was calculated to predict authors' PageRank scores based on their gender. No significant regression equation was found (F(1, 5718) = 0.555, p=0.456), with an R^2 of 0.000. Gender is thus not a significant predictor of PageRank value. That means that gender does not determine the centrality or "canonicity" in the complete network, which again indicates that the Knowledge Graph creates hierarchies that are different from the traditional, gendered literary canon.

These findings can be put in perspective when compared to the gender ratio in the top-down view of canonicity as expressed by the *GNT*. We computed a Pearson correlation coefficient to assess the relationship between the gender of these authors and their frequency of occurrence in the *GNT*. A significant negative correlation does exist between these variables, r= - 0,063 , n=3455 , p<0.001. A linear regression was conducted to predict the frequency of occurrence scores of authors mentioned in the *GNT* based on their gender. A significant regression equation was found (F(1, 226707.927) = 13.586 , p<0.000), with an R^2 of 0.004.[59] An author's predicted frequency of occurrence is equal to a B value of 4.319 - 0.031 (gender) with male coded as 0 and female coded as 1. This means that female authors in the *GNT* score 0.031 lower on frequency of occurrence than male authors. As opposed to the Google-related searches network, gender thus is a significant predictor of these authors' place in the rankings of a tradi-

[59] The low R^2 value can be explained by the fact that most authors occur only once in the *GNT*; the correlation between frequency of occurrence and gender is probably due to a small portion of frequently mentioned, male authors.

tional canon as reflected in the *GNT*.

Finally, some authors in this top 20 seem to show up completely out of the blue. A striking example is Gerda van Wageningen (position 6). She is an author of romantic and historical fiction, who has published over a hundred books and sold over 2.5 million copies (according to Wikipedia).[60] She is, nevertheless, not a critically acclaimed author and is rarely discussed in the media or during university seminars; her name is also absent in the *GNT*. In Figure 3, Graph 1 (left), Van Wageningen features at the left side of the network, represented by the biggest red node in a relatively isolated subcomponent. Her high PageRank value is probably due to several connections between nodes in that subcomponent with nodes in the giant component. These connections effectively boost this nodes' PageRank value and, subsequently, the PageRank value of all others nodes in that subcomponent are also upvoted. Van Wageningen is not directly related to any of the nodes in the giant component where all the highest PageRanked nodes reside, but indirectly through connections with people that do have edges with authors in the center of the network. This also holds for the relatively high PageRank value of Julia Burgers-Drost (position 13) and Henny Thijssing-Boer (position 20) who are both authors of Christian genre fiction and all are part of that same island in the network. Another reason for the high PageRank value of Van Wageningen, Burgers-Drost and Thijssing-Boer is that they form a connected subcomponent in which each node "upvotes" other nodes in the same subcomponent. In other words: these authors end up high in the rankings not because they are directly related to the center of the network but because they are closely connected to one another.

Comparing the network centrality of authors in the Google canon with the prominence of authors in the *GNT* registers thus reveals a difference in gender bias: gender does not correlate with PageRank value in the Google canon but does correlate with frequency of occurrence in the *GNT*. That difference brings us to the possible relationship between the two different notions of canonicity. In order to assess whether an author's PageRank value in the Google canon is related to his/her frequency of occurrence in the *GNT*, we computed a Pearson correlation coefficient to assess the relationship between frequency of occurrence and PageRank score (see Figure 4). There appeared to be a weak but significant positive correlation between the two variables (r=0.181, n=5857 , p<0.000), which suggests that these two different operationalizations of canonicity are at least not entirely independent from each other. While both metrics signify different notions of literary importance, it is clear that some information from existing hierarchies—as reflected in the *GNT*—feeds back into the information economies

[60] https://nl.wikipedia.org/wiki/Gerda_van_Wageningen, accessed 4 July 2019.

mediated by Google.

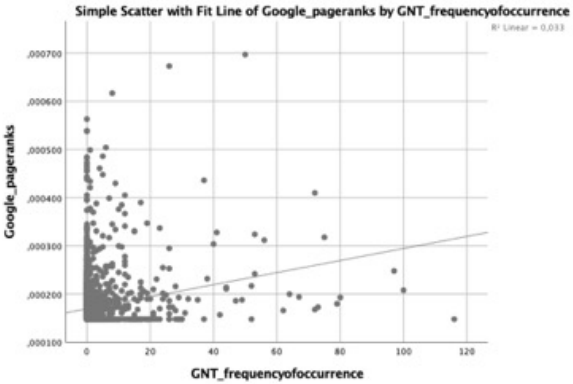

Figure 4. Scatterplot of PageRank value in the Google network against frequency of occurrence in the GNT.

Results II: literary communities

Can we discern groups of authors that cluster together in the network? In order to answer this question, we used Gephi's modularity function to detect communities.[61] A first observation is that there is a large number of communities.[62] That result underscores the fragmented nature of the network in terms of overall connectivity, as the low density and low average degree have already suggested. After filtering out the major part of the smallest communities, three communities appeared to be largest and most dense (see Figure 4). These communities are highlighted by green, red, and blue in Figure 4 and comprise respectively 5.6%, 2.2% and 4.0% of the total network. Betweenness centrality was used for node size in order to discriminate between nodes that perform bridging functions between different communities.[63]

[61] Modularity in Gephi "measures how well a network decomposes into modular communities": https://github.com/gephi/gephi/wiki/Modularity, accessed 5 April, 2018.

[62] In Gephi, the resolution of the modularity function can be adjusted to influence the number of communities. The higher the resolution is set, the fewer (but more strongly connected) communities arise. With the default resolution of 1.0, over 1500 communities were detected. Experimenting with different resolutions demonstrated that the lowest possible number of resolutions arise with a resolution from 10.00 onwards—setting the resolution higher than 10.00 does not yield a lower number of communities.

[63] Betweenness centrality is based on the assumption that a node gets more important when it connects more (disconnected) parts of the network.

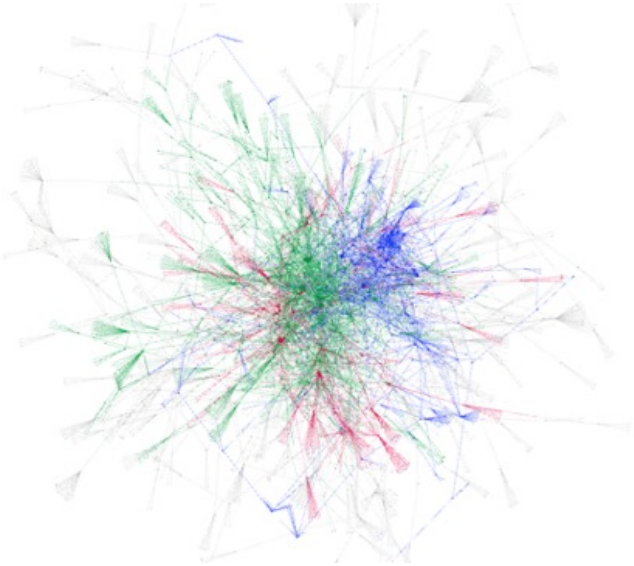

Figure 5. Communities in the giant component of the network. Node and edge color indicate community, node size indicates betweenness centrality.

To understand the possible distinctions between these communities, we labeled the green community as "literary authors," the red community as "popular authors," the blue community as "children's authors": these labels are based on the individuals associated with each community. Note that the labels were not generated by the modularity algorithm but were assigned by us after communities had been identified.

The first community (see Figure 5) features authors that are commonly associated with more literary and, therefore, more canonical forms of fiction. It includes authors such as Remco Campert, Harry Mulisch, Cees Nooteboom, Jan Wolkers, Gerard Reve, Louis Couperus, who have all been vital to the image of important Dutch literary developments (hence their high number of occurrences in the *GNT* registers: 26; 50; 23; 15; 46; 75 times respectively). In terms of gender, male authors are even more overrepresented in this community (77.1% male; 22.9% female) than in the total network (70.1% male; 29.9% female). We performed a Chi-Square Goodness of Fit test to see if this gender distribution significantly deviates from the hypothesized gender distribution of 70.1 / 29.9 in the total network, which appeared to be above the threshold of significance ($\chi 2$ (1) = 7,743, p = 0.05). However, it should be taken into account that the overall gender distribution in the total network is already skewed towards male authors.

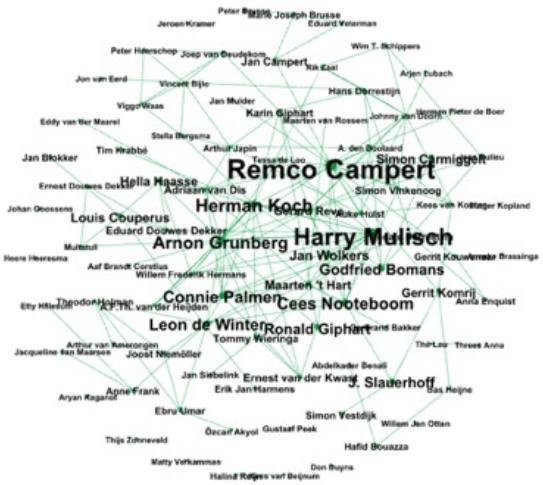

Figure 6. Literary authors community. Node and edge color indicate community, node and label size indicate betweenness centrality. For the sake of readability only nodes with a minimum degree of 10 are shown.

The second community (see Figure 6) features authors who are commonly associated with popular genre fiction, such as chick lit and (literary) thrillers: Saskia Noort, Heleen van Royen, Susan Smit, Judith Visser. Strikingly, none of them are mentioned in the *GNT* registers. Compared to the literary authors community, this community is made up of considerably more female authors (54.9% male; 45.1% female). We performed a Chi-Square Goodness of Fit test to see if this gender distribution significantly deviates from the hypothesized distribution of 70.1/ 29.9, which appeared to be the case ($\chi 2$ (1) = 9,972, $p < .005$). Although there is still an overrepresentation of male authors, it visibly deviates from the gender divide in the total network (70.1% male; 29.1% female) in favor of female authors. This can be interpreted as a repetition of the stereotype that more popular, though less literary forms of fiction are more likely to be written by women.[64]

[64] This stereotypical relation between genre and perception of literary quality was studied in depth in Koolen, *Reading Beyond the Female*.

Figure 7. Popular authors community. Node and edge color indicate community, node and label size indicate betweenness centrality. For the sake of readability only nodes with a minimum degree of 7 are shown.

The gender divide in the third community (see Figure 7, 68.5% male; 31.5% female) is closely similar to the gender divide in the total network (70.1% male; 29.1% female). We performed a Chi-Square Goodness of Fit test to see whether this gender distribution significantly deviates from the hypothesized distribution of 70.1/ 29.9, which indeed was not the case ($\chi 2$ (1) = 0,154, $p >$ 0.05). It features some characteristic Dutch authors of children's books such as Annie M.G. Schmidt, Joke van Leeuwen, Jan Terlouw, Francine Oomen, Ted van Lieshout, and Carry Slee. (Note that not every author makes sense in this community with regards to the label assigned to it. Bert Schierbeek is clearly not a writer of children's literature and, as a winner of the highly prestigious Constantijn Huygens Prize, he would fit better in the picture of the literary authors community.)

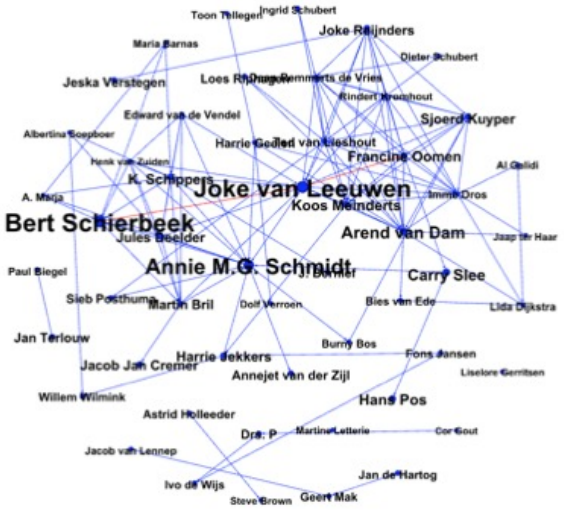

Figure 8. Children's literature community. Node and edge color indicate community, node and label size indicate betweenness centrality. For the sake of readability only nodes with a minimum degree of 10 are shown.

Genre is what mainly defines the three communities described above. In the composition of two of the three communities, gender appears to play a role. The literary canon according to Google thus reproduces the well-known connection between genre and gender: more literary forms of fiction are commonly associated with male authors, and less literary forms with female authors. That the Google canon of Dutch literature consists of these gendered communities can be interpreted as a reproduction of the same power mechanisms that were criticized in the Canon Wars. The predominantly male writers belonging to the community of "literary authors" are the ones that appear in canon lists composed by experts of institutions like the Royal Academy of Dutch Language and Literature. Apparently, existing preconceptions and biases about what constitutes good literature are recurring in the connections that Google's Knowledge Graph makes between authors.

Conclusions

The canon of Dutch literature has been dominated by male authors for centuries. While Google's representation of literary canonicity partly depends on that tradition, we can now establish that the Web also enables new notions of literary importance. The association between an author's gender and literary quality, once again proven by recent quantitative analysis in Koolen 2018, echoes on the Web and therefore informs the Knowledge Graph. Google does not operate in a virtual vacuum unaffected by discursive traditions that shape the symbolic capital of Dutch authors in the world behind indexes and algorithms. The expectations about the Web's potential to enforce a new, democratic "knowledge politics" should therefore be modest, at least with regard to the specific algorithmic mediation of the kind of information analyzed in this contribution. However, it is arguable that Google also enables an understanding of the "canon" that is different from the usual function of that phenomenon in the literary field. The hierarchy between authors constructed by its Knowledge Graph relies on the availability of information about Dutch authors on the Web and search behavior of Google users who search for this information. The preferences of that heterogenous group often reflect but, in some cases, deviate from those of the traditional literary institutions (e.g. academies, prize juries, publishers, literary supplements).

The results of the network analysis allow two opposing conclusions. The first conclusion would be that traditional notions of literary quality determine the importance of authors in Google's logic only to a very limited degree. The ranking of the nodes in the network produced a list of authors who can be regarded as the most central in terms of their PageRank value. On the one hand, this list ranks the usual suspects first: authors whose canonical status seems secure, such as Harry Mulisch, Remco Campert, and Arnon Grunberg. On the other hand, it also features popular authors without canonical status in the literary field, such as Saskia Noort, Esther Verhoef, and Susan Smit. Moreover, we found no correlation between gender and centrality in the total network (as measured by PageRank value), which means that the canon according to Google is not determined by gender. Arguably, the Knowledge Graph thus opens up possibilities for emancipation and diversity among the authors who are available, findable, and prioritized on the Web. Furthermore, our analysis foregrounded authors like Gerda van Wageningen and Julia Burgers-Drost, who generally remain under the radar of literary scholars or critics. These results suggest that, to a certain extent, Google's Knowledge Graph justifies the technological optimism about the alleged potential of the Internet to create a new politics of knowledge.

The second conclusion contradicts the first just mentioned and concerns the conservative element in the data Google returned. The dissemination of the network

in three main communities highlights Google's confirmation of existing norms of literary quality. The authors who cluster together reproduce the association between genre and literary quality already present in traditional literary historiography. More importantly, the observed communities suffer from the same gender bias that critics have underlined during the Canon Wars. The old stereotype remains intact: popular literature, romantic novels, chick lit or thrillers are—relatively speaking—more often associated with female authors and are segregated from mostly male authors of literary quality. Needless to say that it requires more than just a smart search engine to challenge a knowledge politics and a literary canon that has been shaped by literary institutions for centuries. Before web technology can enable a new, bottom-up politics of knowledge or literary canon, one needs a crowd of users who are both able and willing to actively evade that history.

Leaving the specific consequences of the Knowledge Graph aside, we hope to have demonstrated that this Web technology negotiates both familiar and new hierarchies of canonicity. As a result of that function and because of its central position in the behavior of Western Internet users, Google seems to have entered the (Dutch) literary field. In this article we have come to regard Google as a media institution that interacts with the dynamics of the literary field. The search engine claims a unique and dynamic position within the interaction between authors, media, and readers. We therefore propose to add Google to the list of media that have been studied for decades within the tradition of sociological approaches to literature. Admittedly, such research needs to account for the variability of Google's search results. After all, parts of the specific network constructed on the basis of our sample remain only a snapshot of the ever-changing architecture of the search engine. Yet, we also showed that Google's variability over time is not drastic and that different measurements are unlikely to result in a fundamentally different picture. Moreover, as the object of cultural analytics, Google's behavior is similarly dynamic as other literary institutions and media, which likewise change continuously on the ever-changing media landscape. Yet, unlike traditional media, Google's power in the contemporary politics of knowledge continues to grow, which makes it even more important to study its role as an information broker on the cultural field.