Exploring the features of naturalist prose using LIWC in Nederlab

Floor Naber (Amsterdam University of Applied Sciences)

Peter Boot (Huygens ING)

Abstract: Naturalism is one of the best-studied literary movements in (Dutch) literary history. Ton Anbeek formulated in 1982 eight characteristics of Dutch naturalist fiction. Working within the digital library environment Nederlab, we test these characteristics by applying LIWC (Linguistic Inquiry and Word Count) to a corpus of naturalist fiction and a reference corpus of other fiction from the same period. We confirm some of Anbeek's claims (naturalist novels are about nervous characters experiencing a process of disenchantment), fail to confirm some (we find no evidence for the role of determinism), and cannot test some others (e.g. the naturalist author despises bourgeois society). We do find some new 'negative' characteristics of Dutch Naturalism: subjects that occur significantly less in naturalist than in non-naturalist texts. Among these are words related to work, achievement and money. These findings intuitively fit with our idea of a naturalist character but require further study.

Keywords: LIWC, literary movements / literaire bewegingen, Naturalism / naturalisme, computational literary studies / computationele literaire studies, text classification / classificatie van teksten, Nederlab

Introduction

Naturalism is a central concept in the study of 19th century and *fin de siècle* history. Although extensive theories on naturalist literature have been published by both the authors practicing it and by literary historians, there exists no consensus on the defining features of the movement and on the texts that belong to it. As practices of Naturalism varied across countries, the international character of Naturalism is a complicating factor. According to Jacqueline Bel, Naturalism can in fact be considered an 'amalgam of different movements which all aspired to produce novel and unbiased descriptions of reality'.[1]

Nevertheless, several researchers have attempted to identify the primary characteristics of this multifaceted literary movement. In 1982, Ton Anbeek studied the 'family resemblances' of Dutch naturalist novels and established a list containing eight features of Dutch Naturalism. Seven years later, Romain Debbaut published the results of a study in which he examined naturalist plays and collections of short stories as well as novels. He also included in his analysis texts written by Flemish as well as by Dutch authors. In his discussion of the characteristics of Naturalism in the Low Countries, Debbaut presented an exploration of the formal features and stylistics of the movement, in which he remarked that 'in-depth investigations of the formal features still remain to be carried out'[2]. In this article, we investigate the tenability of Anbeek's claims about the naturalistic genre, based on quantitative research into a large body of texts which have often been associated with Naturalism.

The article is the outcome of an experiment in the context of the Nederlab digital research environment, testing its suitability for literary research. In the experiment, we used the LIWC program to count frequencies of certain word groups in naturalist and non-naturalist prose.[3] As the experiment had a limited budget and time frame, we were not able to investigate individual works. Therefore, the relevance of our findings for, say, *Eline Vere* or *De Biezenstekker* must remain a subject for further study.

Distinguishing literary genres

Schools, genres, movements and periods are concepts we use in writing literary history. Over the years, there has been considerable debate about the best ways to categorize literary production. Terminological difficulties have complicated the discussion: genre, sub-genre, school, movement and period have often been used interchangeably to signify certain groupings of literary texts and authors. Some authors have deliberately labeled their work, other classifications were established by literary historians. There is a permanent tension between the conceptual definition of a genre or school, and the works that are said to belong to that genre or school. Dubrow writes: '[...] definitions of genres, like those of biological species, tend to be circular: one establishes such a definition on the basis of a few examples, and yet the choice of those examples from the multitude

[1] Translation of: 'amalgaam van verschillende stromingen die allemaal een nieuwe, vaak onbevooroordeelde beschrijving van de werkelijkheid nastreefden'. J. Bel, *Bloed en rozen. Geschiedenis van de Nederlandse literatuur 1900-1945* (Amsterdam: Bert Bakker, 2015), p. 98.

[2] Translation of: 'een grondig vormelijk onderzoek staat nog te gebeuren'. R. Debbaut, *Het naturalisme in de Nederlandse letteren* (Leuven/Amersfoort: Acco, 1989), p. 138.

[3] More about Nederlab and LIWC below.

of possible ones implies a prior decision about the characteristics of the genre.'[4] Following the challenge posed to conceptions of coherence and totality by deconstructivism and poststructuralism, the debate has concentrated on the question whether these methods of literary systemization resemble historical phenomena. David Perkins argued that literary classifications should be considered 'necessary fictions', necessary because 'we require the concept of a unified period in order to deny it, and thus make apparent the particularity, local difference, heterogeneity, fluctuation, discontinuity, and strife that are now our preferred categories for understanding any moment of the past'.[5]

However, recent computational literary research seems to establish that there is, in fact, a firm textual basis for distinguishing literary genre. In 2011, researchers participating in the Stanford Literary Lab showed how genres in nineteenth-century British literature differed in terms of most frequently used words, in terms of lexico-grammatical categories and in terms of themes and episodes, and how computer algorithms were capable of identifying literary genres based on these characteristics. Humans may recognize only the themes and episodes, but the computer recognizes genre based on lower-level elements.[6] Using a very different approach, based on collocation networks, Amancio, Oliveira jr. and Da Fontoura Costa (2012) studied 77 books published between 1590 and 1922 to detect changes in writing style. They found they could distinguish books belonging to different literary movements according to the traditional classification (Elizabethan era, Neoclassicism/Enlightenment, Gothic fiction, Realism, Naturalism and Modernism).[7] In addition, Kao and Jurafsky (2015) showed that there are several measurable features that can be used to distinguish 19th century poetry, Imagist poetry and modern professional and amateur poetry.[8]

Numerical linguistic and semantic 'evidence' can thus be used to either substantiate or counter claims made by literary historians, or to formulate new hypotheses on literary patterns. One of the strengths of computational technology is that it does not just identify features that are frequently present in a corpus, buts also those that are used less frequently than elsewhere. In terms of Leech and Short's concepts of prominence, foregrounding and deviance: [9] the investigated text corpus may deviate from a reference corpus both positively and negatively, but to a human researcher, only some of the phenomena, usually the more frequent ones, will be more prominent. In this article we will first look at the prominent features – as Anbeek and Debbaut have done, for instance – of the selected texts, and then look at other ways in which these texts deviate from a reference corpus of Dutch fiction dating from the same period of time:

[4] H. Dubrow, *Genre* (London: Methuen 1982), p. 46.

[5] D. Perkins, *Is literary history possible?* (Baltimore and London: Johns Hopkins University Press, 1992), p. 65.

[6] S. Allison, R. Heuser, M. Jockers, F. Moretti and M. Witmore, *Quantitative formalism: an experiment* (Stanford: Stanford Literary Lab, 2011), Pamphlets of the Stanford Literary Lab, Pamphlet 1, p. 8. Retrieved from https://litlab.stanford.edu/LiteraryLabPamphlet1.pdf.

[7] D.R. Amancio, O.N. Oliveira Jr. and L. da Fontoura Costa, 'Identification of literary movements using complex networks to represent texts', in *New Journal of Physics* 14 (2012), 043029.

[8] J.T. Kao and D. Jurafsky, 'A computational analysis of poetic style. Imagism and its influence on modern professional and amateur poetry', in *Linguistic Issues in Language Technology* 12.3 (2015), pp. 1-31.

[9] G. Leech and M. Short, *Style in fiction: a linguistic introduction to English fictional prose*, second edition (Harlow: Pearson Education Limited, 2007).

that is, we first look at the internal coherence of Dutch naturalist prose, checking Anbeek's theory, and then we explore other differences between the naturalist corpus and the reference corpus. What this article will not go into is to what extent these negative features are also foregrounded, i.e., relevant to the artistic effect of the text.

Nederlab and LIWC

Our project was executed within the Nederlab environment. Nederlab is a web environment which renders accessible a large collection of digitized Dutch texts, dating from approximately 800 to the present day. Among other things, it includes the National Library newspaper collection (until 1900) as well as, for our purposes more relevant, the entire DBNL collection.[10] DBNL contains mostly literary texts, especially texts which are out of copyright. In the Nederlab environment, researchers can use different computational tools – to determine word frequencies, keyword-in-context, word trends, etcetera – to examine self-composed selections of texts.

When charged with the task to investigate the suitability of Nederlab for (computational) literary research we selected Naturalism as our topic because it is a well-studied movement and we could base ourselves on existing theories. An important secondary reason for choosing Naturalism is that its works are out of copyright and therefore well represented in DBNL and Nederlab.

The tool that we decided to employ to study content and style of naturalist prose is Linguistic Inquiry Word Count (LIWC).[11] LIWC is a text analysis program that is used to count word group frequencies. It reads texts and analyses their contents based on a dictionary, divided into multiple categories. We used LIWC2007[12], which consists of seventy categories, including function words (pronouns, articles, etc.), psychological processes, which are divided into subcategories such as perception (feel, see, etc.), affect (anger, positive emotions, etc.) and social processes (family, humans, etc.), and personal concerns (home, leisure, death, etc.). Due to the hierarchical structure of LIWC, some words are included in several categories. For example, 'eyes' is both contained in 'body' and its overarching category 'bio'. LIWC output variables are expressed as the percentage of total words in a text that are included in a specific category. For instance, if the output variable for the category *anger* is 0.42, this means that 0.42 percent of the words in a specific text are included in the category *anger*. LIWC's combination of categories containing content words and categories containing function words, enabled us to study both *what* was narrated, but also *how* this was done.

Since LIWC was developed by psychology scholars, researchers who have used LIWC in analyzing literature have often done this from a psychological point of view. For example, in a landmark study Stirman and Pennebaker compared poetry written by suicidal and non-suicidal

[10] DBNL: *Digitale Bibliotheek der Nederlandse Letteren*, The Digital Library of Dutch Literature, http://dbnl.org.

[11] http://www.liwc.net/.

[12] P. Boot, H. Zijlstra and R. Geenen, 'The Dutch translation of the Linguistic Inquiry and Word Count (LIWC) 2007 dictionary', in *Dutch Journal of Applied Linguistics* 6.1 (2017). We chose not to use the recent LIWC2015 version, product of an automatic translation into Dutch. To take into account the effects of changes in spelling we experimented with a dictionary automatically enriched with older spelling forms. Our impression was that the disadvantage of the extra noise created by the automatic enrichment outweighed the advantage of allowing for older spellings.

poets using LIWC.[13] Ryan L. Boyd writes '[t]he analysis of language from this perspective allows us to understand the individual *behind* a given text – their motivations, preoccupations, emotional states, and other facets of their mental universe'.[14] However, LIWC has also been used to study literature outside of a psychological context. For instance, Katherine Blackburn in her thesis uses LIWC to study the narrative arc in novels, short stories and students' narratives.[15] Another example is Andrew Piper's study using LIWC on a collection of approximately twenty-eight thousand texts, dating from the late eighteenth century to the early twenty-first, to determine which features distinguish fiction from non-fiction.[16]

LIWC is certainly not a perfect tool for studying literature. It counts words without regard for polysemy, it cannot detect context or symbolic meaning and fails to recognize irony and figurative language. As Piper also notes, the results of such analyses should interpreted cautiously. It is not self-evident that 'all of the words in the "insight" dictionary [are] really indicative of moments of cognitive insight in novels'[17]. However, as Piper also notes, an advantage of LIWC over topic modeling is that it allows us to test beliefs independently from the collections themselves and to build upon prior assumptions about linguistic categories.

Presently, LIWC is not available as a standard tool in Nederlab. Nederlab staff created special-purpose scripts for us that allowed us to compute the necessary frequencies for the texts that we wanted to investigate. Therefore, the results that we report here, while using the LIWC dictionary, do not use the LIWC program for the computations. Without doubt, there will be differences between the results as we report them and those that would have been produced by the LIWC program. We trust these differences will be small and of no consequence for the conclusions of the article.

Method

Text corpus

Although naturalist characteristics can be traced back several centuries, Dutch Naturalism reached its peak in the *fin de siècle*. We choose to place the beginning of the Dutch version of the literary movement in 1885, the year in which Arij Prins published his *Uit het leven*, which is often considered the first Dutch naturalist collection of stories.[18] When Lodewijk van Deyssel announced the 'death' of Naturalism in 1891, Naturalism had just started to emerge in Flanders.

[13] S.W. Stirman and J.W. Pennebaker, 'Word use in the poetry of suicidal and nonsuicidal poets', in *Psychosomatic Medicine* 63.4 (2001), pp. 517-522.

[14] R.L. Boyd, 'Psychological Text Analysis in the Digital Humanities', in *Data Analytics in Digital Humanities* (2017), 161-189 (p. 163).

[15] K.G. Blackburn, *The narrative arc: exploring the linguistic structure of narrative* (Austin: University of Texas at Austin, 2015).

[16] A. Piper, 'Fictionality', in *CA: Journal of Cultural Analytics*, 20-12-2016. Retrieved from http://culturalanalytics.org/2016/12/fictionality/.

[17] Piper, 'Fictionality'.

[18] Debbaut, *Het naturalisme in de Nederlandse letteren*, p. 153; J. Bel, *Nederlandse literatuur in het fin de siècle: een receptie-historisch overzicht van het proza tussen 1885 en 1900* (Amsterdam: Amsterdam University Press, 1993), p. 282.

Hints of Naturalism can be found in prose published well into the twentieth century. It is impossible to identify a clear-cut boundary, because Naturalism gradually passed into different forms of milder and socially engaged Naturalism-inspired literature. That being said, we decided to place the ending of Dutch Naturalism in 1910, because around that time other literary opinions and movements started to predominate,[19] such as neo-romanticism.[20]

While there is no agreement on books that can be said to belong of Dutch Naturalism, some texts have frequently been associated with the movement. Because we were interested in naturalist patterns outlined in previous research, we decided to compile a data set containing those exemplary texts. To be included in our naturalist corpus, a literary text had to be associated with Naturalism in at least two of the following sources: 1) *De naturalistische roman in Nederland* by Ton Anbeek; 2) *Het naturalisme in de Nederlandse letteren* by Romain Debbaut; 3) *Bloed en rozen. Geschiedenis van de Nederlandse literatuur 1900-1945* by Jacqueline Bel and/or *Alles is taal geworden. Geschiedenis van de Nederlandse literatuur 1800-1900* by Willem van den Berg and Piet Couttenier[21]. If the text was also mentioned in relation to another movement more than two times, it was excluded from the preliminary list. To ensure that formal features did not problematize the comparison with naturalist prose, plays and poetry were excluded. This resulted in list containing a total of 40 naturalist works.

Unfortunately, only 32 of those works were digitized in the DBNL Subsequently, we excluded all novels that contained more than ten percent text written by individuals other than the author (e.g. editors). The final corpus consisted of 29 works, which are listed in table 1. Where possible, we opted for the first edition.

Author	Country	Title	Year
Aletrino	Netherlands	*Zuster Bertha*	1891
Baekelmans	Flanders	*Uit grauwe nevels*	1901
Buysse	Flanders	*De biezenstekker*	1890
Buysse	Flanders	*Het recht van de sterkste*	1893
Coenen	Netherlands	*Verveling*	1892
Coenen	Netherlands	*Een zwakke*	1896
Coenen	Netherlands	*In duisternis*	1903
Coenen	Netherlands	*Zondagsrust*	1902
Couperus	Netherlands	*Van oude menschen*	1906
Couperus	Netherlands	*Eline Vere*	1889
Couperus	Netherlands	*Langs lijnen van geleidelijkheid*	1900

[19] G.J. van Bork, 'Naturalisme', in *Van romantiek tot postmodernisme: opvattingen over Nederlandse literatuur*, ed. by G.J. van Bork and N. Laan (Bussum: Coutinho, 2010), pp. 3-161 (p. 133).

[20] R.A.J. Kraayeveld, 'Naturalistisch proza in Nederland', in *Ons Erfdeel* 27 (1984), pp. 92-101 (p. 99).

[21] W. van den Berg and P. Couttenier, *Alles is taal geworden. Geschiedenis van de Nederlandse literatuur 1800-1900*, second edition (Amsterdam: Bert Bakker, 2016).

Couperus	Netherlands	*Noodlot*	1890/1891
Deyssel, van	Netherlands	*Een liefde*	1887
Emants	Netherlands	*Een nagelaten bekentenis*	1894
Emants	Netherlands	*Inwijding*	1901
Emants	Netherlands	*Liefdeleven*	1916/22
Emants	Netherlands	*Op zee*	1899
Emants	Netherlands	*Waan*	1905
Groeningen, van	Netherlands	*Martha de Bruin*	1889/1890
Heijermans	Netherlands	*Diamantstad*	1904
Heijermans	Netherlands	*Trinette*	1893
Meester, de	Netherlands	*Zeven vertellingen*	1899
Netscher	Netherlands	*Studie's naar het naakt model*	1886
Prins	Netherlands	*Uit het leven*	1885
Querido	Netherlands	*De Jordaan. Amsterdamsch epos I*	1912
Querido	Netherlands	*Menschenwee*	1903
Stijns	Flanders	*Hard labeur*	1904
Streuvels	Flanders	*Lenteleven*	1899
Vermeersch	Flanders	*De last*	1904

Table 1: Corpus 1, Naturalist Texts

It is important to understand that selecting the works that the theorists have identified as naturalist creates a tricky problem. As we have no independent criterion to determine whether a work is in fact a naturalist work, we have to test the scholars' theories on the basis of the works that they considered paradigmatic for the genre. They might have selected other works as the basis for their theories, and their theories presumably would look different. This interdependence between a genre concept and its paradigmatic exemplars is to some extent unavoidable as Dubrow noted. We have sought to mitigate it by relying not just on Anbeek's favourite examples but considering also the works mentioned by other theorists of Naturalism (though these have in all likelihood been influenced by Anbeek).

The authors Coenen, Couperus and Emants dominate the final list, whereas other writers are underrepresented because of selective digitization. For instance, the Flemish writer Gustaaf D'Hondt featured in our initial list, but we couldn't include his collection of short stories *Novellen en schetsen* as it has not been digitized by DBNL. This aggravates to some extent the problem of the interrelatedness of the genre definition and its prototypical examples. The canonical status of works increases the likelihood of their being digitized. Our analysis, which limits itself to the digitally available, and therefore the more canonical naturalist works, is therefore more likely to confirm the theories that we investigate than if the entire naturalist corpus, however one would

[22] *Liefdeleven* was published after 1910, but we decided to include it because it is often associated with Naturalism.

define it, would be digitally available.[23] Another distortion of the corpus, though less important, is caused by some works only being available in multiple parts. For example, Emants' *Liefdeleven* ('Love life') accounts for as much as 14 percent of the body of texts, because it was published as a six-part serial in the Dutch literary periodical *De Gids* ('The Guide'). In total, our corpus consisted of 44 separate texts.

How to create a reference corpus that could be used as a norm for comparison for naturalist prose was not self-evident. Because there is no digitized representative sample of Dutch-language literature from the selected time period – even if a representative sample were theoretically possible, in practice, as noted above, digitization is applied to canonical works first – we based our sample on Nederlab's supply of Dutch literature dating from the same time period. Therefore, we selected all historical novels, novellas, novels, fairy tales and tales from the Nederlab DBNL collection which had a publication date ranging from 1885 to 1910. From this selection of 210 texts, we removed those that: were classified by Nederlab as children's literature; or did not appear in print for the first time between 1885 and 1910; or were written in foreign languages; or were translations; or were associated with Naturalism at least once in the four sources mentioned before.[24] This selection procedure yielded a list consisting of 46 non-naturalist works of fiction, which are listed in table 2. Because *Jeanne Collette* consists of two volumes, we ended up with a corpus containing 47 text files:

Author	Title	Year
Alberdingk Thijm	*Een koninklijke misdaad*	1887
Booven, van	*Tropenwee*	1904
Brink, ten	*Madame de Fontenay*	1897
Busken Huet	*Robert Bruce's leerjaren*	1898
Busken Huet	*Jozefine*	1898
Eeden, van	*De nachtbruid*	1909
Eeden, van	*De kleine Johannes. Deel 1*	1887
Eeden, van	*De kleine Johannes. Deel 2*	1905
Eeden, van	*De kleine Johannes. Deel 3*	1906
Erens	*Korte verhalen*	1906
Huygens	*Barthold Meryan*	1897
Melati van Java	*De ring der grootvorstin*	1889
Jong van Beek en Donk, de	*Hilda van Suylenburg*	1897

[23] Bergenmar: 'There is a risk that the existence of large corpuses of digitised literary texts (…) in combination with effective text mining methods invites a certain kind of approach to literary history, using the archives with available texts and metadata instead of finding out what is lacking in these. (…) The digital literary canon reproduces the canonization already existing in print'. J. Bergenmar, 'Reception history across languages – a challenge for the digital humanities'. Paper presented at *Digital humanities in the nordic countries*, 2016.

[24] Literary texts of which the author was associated with Naturalism and which could not be ruled out as being part of his or her naturalist works, were eliminated from the list as well.

Kollewijn	*Verweghe en zijn vrouw*	1901
Lapidoth	*Goëtia*	1893
Liefde, de	*Uit drie landen*	1900
Loveling	*Sophie*	1885
Maas	*Landelijke eenvoud en andere novellen*	1910
Maclaine Pont	*De poorterszoon van Hoorn*	1895
Maurik, van	*Krates, een levensbeeld*	1885
Maurik, van	*Verspreide novellen*	1885
Maurik, van	*Uit één pen*	1886
Maurik, van	*Oude kennissen*	1895
Maurik, van	*Stille menschen*	1890
Maurik, van	*Amsterdam bij dag en nacht*	1890
Maurik, van	*Toen ik nog jong was*	ca. 1887
Maurik, van	*Papieren kinderen*	1888
Oever, van den	*Kempische vertelsels*	1905
Oordt, van	*Mooi Annie of de schipbreukelinge*	1898
Paap	*Jeanne Collette*	1896
Paap	*Vincent Haman*	1898
Paap	*De doodsklok van het Damrak*	1908
Perk	*De wees van Averilo*	1888
Reynvaan	*Zuster Clara*	1892
Rikken	*Codjo, de brandstichter*	1904
Schendel, van	*Een zwerver verliefd*	1904
Seipgens	*Langs Maas en Geul*	1890
Snijder van Wissenkerke	*Kitty*	1896
Soer	*Catharina: Roman uit den patriottentijd*	1909
Suchtelen, van	*Quia absurdum*	1906
Toussaint van Boelaere	*Landelijk minnespel*	1910
Vosmaer	*Inwijding*	1888
Waals, van der	*Noortje Velt*	1907
Wit, de	*Orpheus in de dessa*	1903
Wit, de	*De godin die wacht*	1903
Woude, van	*Een Hollandsch binnenhuisje*	1888

Table 2: Corpus 2, non-naturalist texts

Text analysis

We based our hypotheses on the eight characteristics of Dutch Naturalism identified by Ton Anbeek in his study *De naturalistische roman in Nederland* ('The naturalist novel in the Netherlands'), based on his analysis of several naturalist novels.[25] They can be summarized as follows:

1. At the heart of the story is a character of a nervous disposition.
2. The story line presents an account of disenchantment.
3. Characters' lives are largely determined by heredity and social environment.
4. The naturalist author despises (bourgeois) society.
5. Naturalist fiction often deals with topics related to sexuality, including taboo aspects such as masturbation, brothels and homosexuality.
6. Naturalist language is characterized by realistic dialogue and *écriture artiste*.
7. There often is a third-person perspective, the events are experienced from the perspective of the character and there is no guiding narrator.
8. Characters are approached objectively.

Other researchers have commented on this list, criticizing the choice of literary texts and the fact that the corpus is limited to novels.[26] In addition, they raised the question whether all characteristics are of equal value and if not, which ones are the most important.[27] The controversy surrounding these characteristics makes the list an interesting test case for the potential of LIWC, and other computational tools, to contribute to scholarly discussion about literature. In the following, we compare Anbeek's features with the results of the LIWC analysis. First, we explain how we translated these features into LIWC categories.[28]

> **Nervous characters (1).** Since Anbeek states that characters of an oversensitive temper are central to Naturalism, we expected that naturalist prose would score higher on the categories *feel*, *anx(iety)*, *body* and *health*. We expected the opposite to be true for the category *posemo* (positive emotions).
>
> **Disenchantment (2)**. In naturalist prose, characters struggle with the discrepancy between their strong ideals and the harsh reality. The outcome of the conflict is either the disillusioned protagonist settling for an unhappy life or suicide. Therefore, we expected that naturalist prose would achieve higher frequency levels on *negemo* (negative emotions), *sad* and *death*.

[25] T. Anbeek, *De naturalistische roman in Nederland* (Amsterdam: De Arbeiderspers/Wetenschappelijke Uitgeverij, 1982), pp. 49-71.

[26] Kraayeveld, 'Naturalistisch proza in Nederland', p. 93.

[27] Kraayeveld, 'Naturalistisch proza in Nederland', p. 100; Van Bork, 'Naturalisme', pp. 156-157.

[28] We did not base hypotheses on main categories such as affect or bio, because their output scores represent the sum of the percentages of subcategories.

Determinism (3). The 'deterministic conditions' that make up the third characteristic can be split into two groups: heredity ('race') and social circumstances ('milieu'). Based on this, we predicted that naturalist texts would score higher on both *cause* and *family*.

Social criticism (4). According to Anbeek, naturalist prose unmasks and condemns the double standards, materialism and class consciousness of bourgeois morality. This could lead to the hypothesis that naturalist prose would score significantly higher on *money* and *achieve*. However, the opposite could also be true, considering that other prose dating from the same period of time could itself be subject to this bourgeois morality. Therefore, we decided to omit this feature from our list of hypotheses.

Sexuality (5). In naturalist prose, certain former taboos related to sexuality are ignored. Anbeek states that these topics are often associated with feelings of guilt and anxiety. Therefore, we expected that naturalist prose would score higher on the categories *sexual* and *anx*.

Colloquialisms and artistic language (6). Naturalist authors attempt to give a realistic impression of spoken language, by using colloquialisms, while at the same time they want to creatively engage with language. Although this could be reflected in a lower amount of words being found in the dictionary, too many other variables could be responsible for differences in this respect, so we decided not to include this feature in our analysis.

Point of view (7). Anbeek draws attention to the naturalist preference for the third-person narrative mode. Therefore, we expected that naturalist texts would score significantly higher on the categories *shehe*, *they* and *shehethey*[29] and relatively low on the categories *i* and *we*. Another change in narration that Anbeek identifies concerns the increase of passages in which the reader encounters the narrated events through the eyes of a character. This statement allowed us to predict that naturalist prose would score higher on the category *see*. Finally, Anbeek draws attention to the disappearance of the omniscient narrator, who introduces the characters and indicates his opinion on them. It wasn't clear to us how to translate this characteristic into LIWC terms.

Objectivity (8). While an unbiased depiction of society and characters is one of the most important aims of naturalist prose, this feature is also very challenging in terms of quantification. Moreover, this principle seems hard to reconcile with naturalist criticism of bourgeois morality. 'Thus the naturalist narrative', David Baguley explains, 'invariably functions as a mediator between the idealistic discourse of characters, society, tradition, and the ironic discourse of the seemingly absent narrator, for the narrative purports to be a neutral form of discourse, limited to an account of facts and events, yet operates as a vehicle for the frustration of idealistic expectations and the justification of tacit ironic responses'.[30] This function is probably too abstract for LIWC to capture. Still, we predicted

[29] This category is a Dutch addition to the original LIWC dictionary.

[30] D. Baguley, *Naturalist Fiction: The Entropic Vision* (Cambridge: Cambridge University Press, 1990), p. 145.

that decrease in moral commentary could be reflected in a lower score on *relig(ion)*, which contains several evaluative words relating to morality.

Finally, moving beyond Anbeek's characteristics, since previous research addressed the alcohol content of naturalist prose,[31] we predicted that naturalist prose would score higher on *ingest*. Because this category contains words related to all kinds of food, drinks and drugs, we also created a new *alcohol* dictionary category. The contents of this category were based on a combination of the alcohol-related terms from LIWC's category *ingest*[32] and the results of a text search into three naturalist and three non-naturalist texts. In summary, we expected that naturalist prose would score higher on the categories *alcohol*, *anx*, *body*, *cause*, *death*, *family*, *feel*, *health*, *ingest*, *negemo*, *sad*, *sexual*, *see*, *shehe*, *shehethey* and *they*, but lower on the categories *i*, *posemo*, *relig* and *we*. Examples of the words included in some of the different categories can be found in table 3:

LIWC category	Dutch examples	English examples
i	ik, mezelf, mijn	I, myself, my
anx	zenuwachtig, eng	nervous, scary
cause	dus, scheppen, verband	therefore, create, connection
see	afbeelding, oog, toekijken	image, eye, watch
sexual	bloot, lust, zoen	naked, lust, kiss
death	crematorium, doden, urn	crematorium, kill, urn
shehethey	hemzelf, hij, zich	himself, he, itself

Table 3. LIWC categories and word examples.

As noticed previously, text analyses with LIWC only generate category-level frequencies. To be able to interpret these percentages, we also analyzed word differences between the two corpora. To this end, we calculated for all texts the 5000 most frequently used words. We then computed which words occurred significantly more in the naturalist corpus than the reference corpus or vice versa,[33] and grouped them according to the LIWC categories. Where relevant, the results of this analysis will be presented alongside the discussion of our LIWC results.

Results and discussion

Firstly, we performed unpaired t-tests to examine which of our twenty language dimensions differed significantly between the two corpora. Table 4 displays the mean category scores in percentages and the category standard deviations for both naturalist and non-naturalist prose, as well as the significance and the effect size (Cohen's *d*)[34] of the difference. Sixteen of our

[31] J. Bel, 'Het alcoholpromillage van de Nederlandse naturalistische roman', *Rozenberg Quarterly*. Retrieved from http://rozenbergquarterly.com/het-alcoholpromillage-van-de-nederlandse-naturalistische-roman/ [04-11-2017].

[32] We not only included words from this category in the LIWC2007 dictionary, but also the dictionary of LIWC2015, because it contained a few extra alcohol-related words.

[33] Using a chisquare test with significance level $p < 0.01$.

[34] To calculate Cohen's *d*, we divided the mean differences by the pooled standard deviations.

hypotheses were supported by the results ($p < 0.05$). The effect sizes range from just below medium to high:

LIWC category group	LIWC category	Anbeek claim	Nat. (N=44)		Non-nat. (N=47)		Hypothesis confirmed	p-value	effect size (*d*)
			av	*sd*	*av*	*Sd*			
Function words	I	7	1.22	1.33	1.73	1.26	yes	0.03	0.40
	We	7	0.18	0.13	0.37	0.29	yes	0.00	0.86
	Shehe	7	5.37	1.86	4.35	1.43	yes	0.00	0.61
	They	7	2.89	0.99	2.12	0.80	yes	0.00	0.86
Social	Family	3	0.54	0.33	0.73	0.33	no	0.00	0.56
Affect	Posemo	1	2.40	0.76	2.71	0.59	yes	0.02	0.44
	Negemo	2	2.38	0.53	1.98	0.48	yes	0.00	0.80
	Anx	1, 5	0.52	0.18	0.42	0.14	yes	0.00	0.64
	Sad	2	0.82	0.26	0.70	0.20	yes	0.01	0.50
Cognitive	Cause	3	0.63	0.25	0.72	0.21	no	0.02	0.42
Perceptual	See	7	1.60	0.42	1.41	0.72	no	0.06	0.33
	Feel	1	1.14	0.28	0.86	0.30	yes	0.00	0.94
Biological	Body	1	1.29	0.44	1.01	0.31	yes	0.00	0.75
	Health	1	0.49	0.20	0.38	0.24	yes	0.01	0.51
	Sexual	5	0.17	0.08	0.12	0.07	yes	0.00	0.61
	Ingest		0.45	0.22	0.37	0.18	yes	0.03	0.40
Personal concerns	Relig	8	0.23	0.18	0.35	0.22	yes	0.00	0.57
	Death	2	0.16	0.18	0.14	0.09	no	0.32	0.10
Dutch LIWC addition	Shehethey	7	6.70	1.71	5.22	1.52	yes	0.00	0.92
Custom category	Alcohol		0.26	0.16	0.18	0.10	yes	0.00	0.59

Table 4. Value per category (LIWC 2007). Values are in mean percentages.

Nervous characters. As expected, naturalist prose contained significantly more words that were classified as *feel*, *anxiety*, *body* and *health*, whereas non-naturalist prose contained significantly more words related to *positive emotions*. Looking at specific words, for instance in the *body* category, shows that, while there are a few words that occur more frequently in the reference corpus ('Eye'), nearly all body parts score higher in the naturalist corpus ('eyes', 'nose', 'nostril', 'mouth', 'lips', 'teeth', 'jaws', 'tongue', etc. etc). An orientation towards the body seems a very strong characteristic of naturalism. For the positive emotions, many of the significantly non-naturalist words are abstract and timeless, such as 'beauty', 'wisdom', 'truth', 'divine'; at the naturalistic side, positive emotion words include quite different words such as 'kiss(ed)', 'lust', 'passion', 'pleasure' (*genot*).[35] Zooming in on word differences in the category *feel*, we found

[35] LIWC's emotion categories do not just contain the emotions themselves but also things that are supposed to cause positive or negative emotion, such as 'murder' in the negemo category and 'kiss' in *posemo*.

many sensation words that occurred significantly more often in naturalist prose than in the reference body of texts[36], for instance physical adjectives such as *zacht* ('soft'), *breed* ('wide/broad'), *strak* ('tight'), *vochtige* ('moist') and *slap* ('weak').

Interestingly, this trend towards concrete language was also found in a computational analysis of almost 3000 nineteenth-century British novels by Ryan Heuser and Long Le-Khac (2012). They found a chronological transformation from evaluative, abstract language into non-evaluative, concrete language over the nineteenth century.[37] The researchers related this trend to a shift from telling to showing and remarked that this suggested that 'the modes of evaluation and characterization changed, moving from explicit to implicit narration, from conspicuous commentary to the dramatization of abstractions, qualities, and values through physical detail'.[38] Applying this statement to the results found regarding the difference between naturalist texts and the reference corpus, this could suggest that naturalist prose belongs to a new category of literature, while other literature published in the same time period reflects older types of literary description.

Disenchantment. As expected, naturalist prose scored significantly higher on *negative emotions* and *sadness*. However, there were no significant differences in the occurrence of *death* words. For the negative emotions, while naturalist prose contains more anger-related words, such as *woedend* ('furious'), *driftig* ('heated') and *slaan* ('to hit'), non-naturalist prose contained significantly more words related to war, such as *vijanden* ('enemies'), *strijd* ('battle') and *aanval* ('attack'). This suggests that while naturalist prose revolves around individuals, the reference corpus is about groups. Besides, whereas naturalist prose contained significantly more words related to indifference, such as *onverschillig* ('indifferent'), *leegte* ('emptiness') and *verveling* ('boredom'), non-naturalist prose contained significantly more morality related words, such as *zonden* ('sins'), *straffen* ('to punish', 'punishments') and *schuldig* ('guilty'). Clearly, for the reference corpus, earlier social and religious values retain the validity that they have lost for the characters in naturalist prose.

Determinism. Contrary to what we expected, naturalist prose scored significantly lower on the category *family*. The naturalist view of the family as a cause of hereditary illness perhaps loses out to the family as a traditional factor binding the individual to society. Moreover, the word frequency of *causal* words is even significantly higher for non-naturalist prose. Naturalism may explicitly view certain events as determined by the past, but apparently causality is important to more traditional literary schools as well. On reflection, it is probably to be expected that in a traditional world view events are more tightly related than in the modern view espoused by Naturalism. It is also possible that the (objective) naturalist narrator wants his readers to draw their own conclusions about causal relations, rather than indicating them in the text. That could

[36] All words mentioned in our results scored significantly higher in one of our data sets compared to the other in a Chi2 test ($p < 0.01$).

[37] R. Heuser and L. Le-Khac, *A quantitative literary history of 2,958 nineteenth-century British novels: The semantic cohort method* (Stanford: Stanford Literary Lab, 2012), Pamphlets of the Stanford Literary Lab, Pamphlet 4. Retrieved from https://litlab.stanford.edu/LiteraryLabPamphlet4.pdf.

[38] Heuser and Le-Khac, *A quantitative literary history of 2,958 nineteenth-century British novels*, pp. 45-46.

explain why the most significant *cause* word for the naturalist corpus is *waarom* ('why'), while the most significant word for the reference corpus is *daarom* ('therefore', 'that's why').

Sexuality. As expected, naturalist texts scored significantly higher on *sexuality*. An examination of sexuality related words that occur significantly more in naturalist prose than in the reference body of texts indicates several explicit sensation verbs, such as various inflections of 'kissing' and 'embracing' as well as 'lust', 'naked' and *begeerte* ('(sexual) desire'). There is no indication among the significantly different words for a special naturalist interest in taboo aspects of sexuality.

Point of view. In our analysis of personal pronouns, we found that non-naturalist prose scored significantly higher on the categories *i* and *we*, and significantly lower on *shehe*, *they* and *shehethey*, as expected. This suggests that Anbeek was right in stating that the third-person narrative predominates naturalist fiction. However, naturalist fiction did not score significantly higher on the category *see*. Interestingly, a number of words from the *see* category that were used significantly more in the naturalist texts correspond to Heuser and Le-Khac's theory of 'hard seed' fields, such as several color-related words (e.g. *zwarte* ('black'), *grijze* ('grey') and *rood* ('red')).

Objectivity. Although we have discussed the difficulty of counting frequencies of words related to the ideological objectivity of a literary text, we predicted that naturalist prose would score significantly lower on *religion* than our reference text collection. This expectation turned out to be fulfilled. An examination of significant religion related word differences showed that words related to morality, such as *zedelijk* ('morally'), *zonden* ('sins') and *moreel* ('morally'), appeared significantly less frequently in naturalist prose. This finding can be supplemented by observations of positive emotions. The reference data set showed significantly larger amounts of positive emotion words related to moral evaluation, such as *verheven* ('sublime'), *edele* ('noble'), *onschuld* ('innocence'), *fatsoenlijk* ('decent') and waardigheid ('dignity'). Conversely, naturalist prose contained significantly more occurrences of *oprecht* ('sincere') and *oprechtheid* ('sincerity'), indicating an interest in truthfulness (although 'truth' itself is more used in the reference corpus).

It is hard to say which of the characteristics is the strongest indicator of a naturalist text. As noted above, some of the characteristics couldn't really be translated into LIWC terms. For all of them, it would be naïve to assume that LIWC can capture their full meaning. Nevertheless, the effect sizes give us an idea of which factors are the most important. The effect sizes for the categories *feel* (d = 0.94), *shehethey* (d = 0.92), *we* (d = 0.86), *they* (d = 0.86) and *negemo* (d = 0.80) are large (d ≥ 0.80). This would suggest that nervous characters, disenchantment and point of view are the most distinguishing features of Dutch Naturalism.

Other language dimensions

As well as testing our hypotheses, we decided to compare the naturalist and the reference body of text on the other LIWC-categories. We included all linguistic, psychological and personal language dimensions, but excluded the spoken categories *assent*, *nonfluencies* and *fillers*. We had no preconceptions about the direction or the strength of the differences, but simply present

the results here as suggestions for further research in Naturalism. Table 5 shows the results of an unpaired, two-tailed t-test, listing the top 25 LIWC categories that differ significantly, ranked by effect size:

Category	Nat mean	Nat sdev	Non-nat mean	Non-nat sdev	p-value	Effect size (*d*)
work	0.48	0.20	0.77	0.30	0.00	1.14
feel	1.14	0.28	0.86	0.30	0.00	0.94
shehethey	6.70	1.71	5.22	1.52	0.00	0.92
friend	0.08	0.05	0.12	0.05	0.00	0.89
we	0.18	0.13	0.37	0.29	0.00	0.86
they	2.89	0.99	2.12	0.80	0.00	0.86
negemo	2.38	0.53	1.98	0.48	0.00	0.80
article	7.04	1.67	8.26	1.44	0.00	0.79
body	1.29	0.44	1.01	0.31	0.00	0.75
money	0.30	0.14	0.43	0.25	0.00	0.65
future	0.61	0.27	0.78	0.25	0.00	0.65
achieve	0.82	0.29	1.02	0.32	0.00	0.65
anx	0.52	0.18	0.42	0.14	0.00	0.64
anger	0.69	0.23	0.56	0.17	0.00	0.63
shehe	5.37	1.86	4.35	1.43	0.00	0.61
home	0.70	0.33	0.54	0.17	0.01	0.61
sexual	0.17	0.08	0.12	0.07	0.00	0.61
alcohol	0.26	0.16	0.18	0.10	0.01	0.59
relig	0.23	0.18	0.35	0.22	0.01	0.57
family	0.54	0.33	0.73	0.33	0.01	0.56
health	0.49	0.20	0.38	0.24	0.02	0.51
sad	0.82	0.26	0.70	0.20	0.02	0.50
swear	0.03	0.03	0.02	0.01	0.03	0.49
adverb	4.19	1.37	3.66	0.72	0.03	0.48

Table 5. The top 25 LIWC categories with the largest effect size (Cohen's d).

Although some of the categories with the largest effect size correspond to the naturalist characteristics formulated by earlier scholars, the list in table 5 also contains some new categories. At the top of the list is *work* (d = 1.14), followed by new categories such as *friend* (d = 0.89), *article* (d = 0.79), *money* (d = 0.65), *future* (d = 0.65), *achieve* (d = 0.65), *anger* (d = 0.63), *home* (d = 0.61), *swear* (d = 0.49) and *adverb* (d = 0.48). Of these categories, *anger*, *home*,

swear and *adverb* occur predominantly in naturalist prose; *work*, *friend*, *article*, *money*, *future* and *achieve* occur especially in non-naturalist prose. This is certainly suggestive: naturalist fiction is apparently about lonely and isolated individuals (low on *friend*, as well as on *we*, as noted above), who are not involved in society (low on *work*). The naturalist character is oriented toward the past (low on *future*) and doesn't care about status and success (low on *money* and *achievement*). He is angry about his life (high on *anger*) and lives his life mostly at home (high on *home*). Some of the words in the *home* category support these interpretations. Specifically, naturalist words there are '(bed)room', 'bed' and several words related to bedclothes, as well as 'window' and related words: together they suggest an individual hiding in his room, spending his days in bed, looking out towards a world he doesn't participate in. On the non-naturalist side, frequently occurring *home* words include '(nuclear) family', 'inhabitants' and 'servants': here, home is a place where people live together.

Furthermore, the predominance of the category *adverb* for naturalist prose and *article* for non-naturalist prose is interesting in light of research on the use of function words from genre to genre. The LIWC2015 psychometrics manual shows that articles are higher for more formal contexts (e.g. newspaper articles), whereas adverbs are generally higher in more 'natural' or 'realistic' contexts (as in conversation or in writing about one's feelings in an Expressive Writing paradigm).[39] The patterns shown in table 5 regarding articles and adverbs suggest that naturalist texts offer a more natural and less formal type of language. [40]

The fact that words concerning *work*, *friend*, *article*, *future*, *achieve* and *money* all appear more in non-naturalist prose than in naturalist prose highlights an interesting omission in the current characterization of naturalist prose. Anbeeks identification of 'family resemblances' only allows him to formulate similarities between the novels he analyzes, and does not enable him to detect topics or stylistic elements that are absent from or do not often occur in these novels, compared to other works of fiction. Our study shows that large quantitative comparisons of textual data allow researchers to detect both prominent, overrepresented, as well as underrepresented features. However, as we will discuss later, a cautious approach is required when interpreting the frequencies.

Conclusion

Understanding of Naturalism. In this article we looked at Anbeek's ideas about Dutch Naturalism and tried to test these using a corpus approach. We translated the claims into hypotheses about the relative frequencies of certain LIWC categories in a naturalist and a non-naturalist corpus. We found evidence for a number of the claims: the main character has a nervous disposition, the story is an account of disenchantment, sexuality plays an important role, there is a third-person perspective and characters are approached objectively. We did not find evidence of the characters' lives being determined by heredity and social environment nor of an interest in taboo aspects of sexuality. We did not test Anbeek's claims about the naturalistic

[39] J.W. Pennebaker, R.L. Boyd, K. Jordan and K. Blackburn, *The development and psychometric properties of LIWC2015* (Austin: University of Texas at Austin, 2015), p. 10.

[40] Within the adverb category, we note an interesting preference of Naturalism for modern word variants: for 'now', 'very' and 'already', modern forms are used in naturalist prose, old-fashioned forms in non-naturalist prose (*nu/nou* vs. *thans*, *heel* vs. *zeer* and *al* vs. *reeds*). This might be related to a naturalist preference for colloquial words.

author despising society or about naturalistic style. The evidence for the disenchantment and point of view claims was mixed, as the differences for *death* and *see* were not significant.

An examination of words that were used significantly more or less in naturalist prose compared to a reference corpus, generated results which corresponded to Heuser and Le-Khac's notion of the nineteenth century shift in the novel's style and narration: a transition from abstract, evaluative language through concrete, non-evaluative language. Naturalist prose showed a remarkable number of 'concrete sensation' word occurrences, ranging from colors to body parts and physical descriptions. On the other hand, it did contain relatively low amounts of words related to moral evaluation. This seems to confirm the hypothesis that naturalist prose belongs to a new category of (realist?) literature, whereas other prose published in the same time period uses older types of literary description. Further research could elaborate on these findings, and use it as a starting point to re-evaluate the relationship between naturalist, romantic and realist literature.

Furthermore, our analysis showed major differences between the two corpora for LIWC categories that did not correspond to Anbeek's claims, such as *work*, *home* and *achievement*. Based on these categories we could draw an intuitively plausible picture of the typical naturalist character. This shows an advantage of quantitative analysis of large bodies of texts over traditional close reading strategies: it enables researchers to assess elements that are absent in certain textual data sets by comparing these texts to reference data sets.

While we believe these results are interesting, there are important limitations to what we have done. Due to time restraints, many of our outcomes are suggestive rather than conclusive. The high-level findings about word usage should be complemented by study of individual works where our findings should be related to individual text passages. It should also not be forgotten that for most of the claims, the translation into LIWC terms was problematic and, in some cases, impossible. This brings us to the topic of LIWC's suitability for this type of research.

Suitability of LIWC for literary purposes. LIWC as a tool for literary study has no doubt important limitations. While an advantage is the fact that it allows researchers to represent large bodies in general, independent categories, this reduction of complexity is also its main drawback. It does nothing but counting words in certain categories, without taking into account polysemy, metaphor or syntactic relations. It is clear that many aspects of literature are beyond the reach of LIWC. One specific aspect of literature is that literary texts are multilayered structures, in which perspectives of characters, narrators and (implied) authors converge in intricate ways. When LIWC counts words, should they be considered as characteristic for the author, for the genre or for the main character? And is frequency of occurrence equivalent to literary importance? Furthermore, although LIWC's dictionary might seem objective, its contents are subject to human choices that are often to some extent arbitrary.

However, as we have shown above, LIWC can in fact be used to detect genre-level patterns which can support or refine existing claims about literature, as well as suggest new ones. Apparently, vocabulary choices by themselves to some extent reveal genre properties. It should be added that we consider the analysis of significantly used individual words within the categories, which goes beyond what LIWC can do, as an essential part of the analysis, as it helps understand the effects at the category level.

When moving from the genre to the individual book, LIWC analyses could still be used as a starting point, followed by contextual examinations of uncovered striking features. One

important step for literary scholars would be to add categories to LIWC, depending on the demands of the texts under study, the way we did for alcohol. However, there will remain literary features whose detection is beyond the scope of word counting tools, such as objectivity, metaphorical language and sarcasm. Here other approaches might become necessary, such as techniques that utilize word sense disambiguation, techniques that consider context, such as collocation analyses, and techniques employing part-of-speech tagging. But it is also possible that the attempt to translate a theoretical term into computationally tractable categories shows that the theoretical term is not as clear as we thought it was. That is, in fact, what we believe is the case for the notion of objectivity.

It should be noted that in this article, we only looked at book-level LIWC scores, as Nederlab does not facilitate access at lower levels (e.g. chapters). This would be an interesting extension of the sort of research that we have done. There exists by now a body of computational research into the 'emotional arc' of fictional works.[41] To characterise e.g. a narrative as one of disenchantment implies more than a negative overall output, as we have assumed here: it implies a hopeful beginning and a disillusioned ending. Our present measurements cannot give us that information. Another extension would be some way of distinguishing words from the various character perspectives from those from the narrator. That would help us understand something of the complex interaction of perspectives within the novel. However, this would require types of analysis that are far beyond the capabilities of LIWC by itself. What we have presented here shows only the first steps of a computational analysis of Naturalism.

Bibliography

Allison, S., R. Heuser, M. Jockers, F. Moretti and M. Witmore, *Quantitative Formalism: an Experiment* (Stanford: Stanford Literary Lab, 2011), Pamphlets of the Stanford Literary Lab, Pamphlet 1. Retrieved from https://litlab.stanford.edu/LiteraryLabPamphlet1.pdf.

Amancio, D.R., O.N. Oliveira Jr. and L. da Fontoura Costa, 'Identification of literary movements using complex networks to represent texts', in *New Journal of Physics* 14 (2012), 043029.

Anbeek, T., *De naturalistische roman in Nederland* (Amsterdam: De Arbeiderspers/Wetenschappelijke Uitgeverij, 1982).

Archer, J. and M.L. Jockers, *The Bestseller Code: Anatomy of the Blockbuster Novel* (New York: St. Martin's Press, 2016).

Baguley, D., *Naturalist fiction: The Entropic Vision* (Cambridge: Cambridge University Press, 1990).

Bel, J., 'Het alcoholpromillage van de Nederlandse naturalistische roman', *Rozenberg Quarterly*. Retrieved from http://rozenbergquarterly.com/het-alcoholpromillage-van-de-nederlandse-naturalistische-roman/ [10-08-2017].

---, *Nederlandse literatuur in het fin de siècle: een receptie-historisch overzicht van het proza tussen 1885 en 1900* (Amsterdam: Amsterdam University Press, 1993).

---, *Bloed en rozen. Geschiedenis van de Nederlandse literatuur 1900-1945* (Amsterdam: Bert Bakker, 2015).

[41] See e.g. J. Archer and M.L. Jockers, *The Bestseller Code: Anatomy of the Blockbuster Novel* (New York: St. Martin's Press, 2016), pp 73-111.

Berg, W. van den, and P. Couttenier, *Alles is taal geworden. Geschiedenis van de Nederlandse literatuur 1800-1900*, second edition (Amsterdam: Bert Bakker, 2016).

Bergenmar, J., 'Reception History Across Languages – a Challenge for the Digital Humanities'. Paper presented at *Digital Humanities in the Nordic Countries*, 2016. Retrieved from http://www.hf.uio.no/iln/english/research/networks/digital-humanities/news-and-events/events/2016/pdf/bofab.pdf#page=51.

Blackburn, K.G., *The Narrative Arc: Exploring the Linguistic Structure of Narrative* (Austin: University of Texas at Austin, 2015).

Boot, P., H. Zijlstra and R. Geenen, 'The Dutch Translation of the Linguistic Inquiry and Word Count (LIWC) 2007 Dictionary', in *Dutch Journal of Applied Linguistics* 6.1 (2017), pp. 65-76.

Bork, G.J. van, 'Naturalisme', in *Van romantiek tot postmodernisme: opvattingen over Nederlandse literatuur*, ed. by G.J. van Bork and N. Laan (Bussum: Coutinho, 2010), pp. 3-161.

Boyd, R.L., 'Psychological Text Analysis in the Digital Humanities', in *Data Analytics in Digital Humanities* (2017), 161-189.

Debbaut, R., *Het naturalisme in de Nederlandse letteren* (Leuven/Amersfoort: Acco, 1989).

Heuser, R. and L. Le-Khac, *A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method* (Stanford: Stanford Literary Lab, 2012), Pamphlets of the Stanford Literary Lab, Pamphlet 4. Retrieved from https://litlab.stanford.edu/LiteraryLabPamphlet4.pdf.

Kao, J.T. and D. Jurafsky, 'A Computational Analysis of Poetic Style. Imagism and its Influence on Modern Professional and Amateur Poetry', in *Linguistic Issues in Language Technology* 12.3 (2015), pp. 1-31.

Kraayeveld, R.A.J., 'Naturalistisch proza in Nederland', in *Ons Erfdeel* 27 (1984), pp. 92-101.

Leech, G. and M. Short, *Style in fiction: a linguistic introduction to English fictional prose*, second edition (Harlow: Pearson Education Limited, 2007).

Pennebaker, J.W., Boyd, R.L., Jordan, K. and Blackburn, K., The Development and Psychometric Properties of LIWC2015 (Austin: University of Texas at Austin, 2015). Retrieved from https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf

Perkins, D., *Is Literary History Possible?* (Baltimore and London: Johns Hopkins University Press, 1992).

Piper, A., 'Fictionality', in *CA: Journal of Cultural Analytics*, 20-12-2016. Retrieved from http://culturalanalytics.org/2016/12/fictionality/.

Stirman, S.W. and J.W. Pennebaker, 'Word Use in the Poetry of Suicidal and Nonsuicidal Poets', in *Psychosomatic Medicine* 63.4 (2001), pp. 517-522.

About the Authors

About the authors

Floor Naber graduated in Dutch literary studies at Utrecht University. Her master thesis focused on the representation of social mobility in literature concerning the Dutch East Indies. She worked as a junior researcher on the project 'LIWC in Nederlab' (Huygens ING - KNAW), in which she explored the possibilities of LIWC for literary studies within the Nederlab environment. She currently teaches language and communication skills at the Amsterdam University of Applied Sciences.

Peter Boot is a senior researcher at the Huygens Institute for the History of the Netherlands. He graduated in mathematics (Leiden) and Dutch literature and culture (Utrecht). In between he worked as a programmer and software consultant. He wrote his PhD thesis about annotation in digital editions and its potential implications for research in the humanities (*Mesotext. Digitised Emblems, Modelled Annotations and Humanities Scholarship*, 2009). He works as a consultant on digital edition projects (Van Gogh, Mondrian, medieval miscellanies). With others, he created the Dutch translations of LIWC that were used in the present article. His current research focuses on online response to books (on e.g. dedicated review sites or booksellers' sites) and what it can teach us about readers and reading.