Cliché Expressions in Literary and Genre Novels

Andreas van Cranenburgh
Heinrich Heine University of Düsseldorf
Universitätsstraße 1, 40225 Düsseldorf, Germany
cranenburgh@phil.hhu.de

Abstract

Should writers "avoid clichés like the plague"? Clichés are said to be a prominent characteristic of "low brow" literature, and conversely, a negative marker of "high brow" literature. Clichés may concern the storyline, the characters, or the style of writing. We focus on cliché expressions, ready-made stock phrases which can be taken as a sign of uncreative writing. We present a corpus study in which we examine to what extent cliché expressions can be attested in a corpus of various kinds of contemporary fiction, based on a large, curated lexicon of cliché expressions. The results show to what extent the negative view on clichés is supported by data: we find a significant negative correlation of -0.48 between cliché density and literary ratings of texts. We also investigate interactions with genre and characterize the language of clichés with several basic textual features. Code used for this paper is available at `https://github.com/andreasvc/litcliches/`

1 Introduction

What makes certain novels *literary*? Insofar as this is ascribed to the text itself, the text is said to exhibit the phenomenon of *literariness*: the hypothesized linguistic and formal properties that distinguish literary language from other language (Baldick, 2008). Others point to the prestige that publishers and critics confer (Bourdieu, 1996). Empirical support for literariness as a textual property is presented by van Cranenburgh and Bod (2017), who present machine learning experiments predicting literary ratings based on a wide range of textual features. This paper zooms in on the contribution of one particular feature, clichéd language.

A direct way to investigate literariness would be to define a way to measure its particular properties such as creative, original use of language. This is the aim of the Formalist tradition, which holds that poetic language distinguishes itself from standard language by the phenomena of foregrounding and defamiliarization (Mukarovsky, 1964). However, these phenomena seem difficult to operationalize computationally, at least without the collection of detailed human judgments. Text analysis can demonstrate how well a large number of textual features (such as bag of word models and syntactic features) predicts literariness, but due to the large number of features, the resulting model is hard to interpret (van Cranenburgh and Bod, 2017). By contrast, unoriginal language use can be readily detected, since it is by definition commonly attested in data. Therefore we opt to investigate clichés as a negative marker of literariness.

Clichés in literature can manifest themselves at various levels such as narrative, style, and characters. We focus on cliché expressions at the sentence level since they are the most amenable to automatic analysis using textual search.

2 Definitions & datasets

We define cliché expressions as follows:

Definition. A *cliché expression* is a fixed, conventionalized multi-word expression which has become overused to the point of losing its original meaning or effect.

Let's unpack the main terms in this definition:

FIXED: the expression in the form that is recognized cannot be changed, or only to a limited degree by filling in specified open slots.

CONVENTIONALIZED: i.e., the phrase is recognized by many speakers as a unit, instead of being put together word for word.

OVERUSED: this aspect is crucial but subjective and therefore harder to pin down. Many other multi-word expressions are accepted as a normal part of the lexicon, while cliché expressions are marked as formulaic, tired, unoriginal, etc.

Cook and Hirst (2013) state that "a cliché is a kind of ersatz novelty or creativity that is, *ipso facto*, unwelcome or deprecated by the reader." The term 'overused' might suggest that there is some range of acceptable frequency for expressions, but this limit seems hard to determine; the cliché-hood of an expression rests on a tacit, cultural judgment.

The cliché expressions we focus on are semantically compositional and syntactically regular, without non-literal meaning. However, since they are conventionalized, their use provides evidence that the author did not construct the expression word for word, but took a shortcut by employing a ready-made stock phrase. The occurrence of such expressions may therefore be taken as a negative marker for creativity and originality.

To operationalize the question of cliché-hood we use a cliché lexicon with 6,641 Dutch cliché expressions provided to us. This dataset is the source for a published collection of clichés (van Wingerden and Hendriks, 2015). In collecting this set of expressions, the focus was not on expressions that are established sayings or figures of speech, but rather formulaic commonplaces for mundane situations—language that does not necessarily stand out by itself but is recognizable as clichéd by how typical it is for a particular social situation.

We determine the frequencies of the clichés in a corpus of contemporary novels and relate them to the results of a survey investigating literary evaluations of the novels among the general public. The aim is to see whether the prevalence of cliché expressions offers insights into literary evaluations. For example, to what extent the intuition that less literary texts contain more clichés holds up.

The dataset of novels was the subject of a large online reader survey (about 14k participants), to obtain judgments of literary and general quality. This survey was conducted as part of the project The Riddle of Literary Quality,[1], investigating the textual characteristics of contemporary literature. The 401 recent Dutch novels (as well as works translated into Dutch) were the best selling and most lent books in 2007–2012. The corpus contains literary novels as well as genre novels such as thrillers and romantic novels. The participants were presented with the author and title of each novel, and for novels they had read were asked to provide ratings on a 7-point Likert scale from *definitely not* to *highly* literary. 96 % of the novels have 50 or more ratings. In the following, we use the mean of a novel's ratings as its literary evaluation.

As reference corpora we will also look at Lassy Small and CGN. Lassy Small (Van Noord, 2009) consists of written text (e.g., newswire and and Wikipedia text). CGN (van der Wouden et al., 2002) is a corpus of spoken language.

3 Matching cliché expressions

The process of searching through a corpus for a predefined lexicon of expressions is an instance of Multi-Word Expression (MWE) identification (Kulkarni and Finlayson, 2011; Constant et al., 2017).

We tokenize both the clichés and the novels to obtain a format of one cliché/sentence per line with space-separated tokens. Not all clichés consist of a fixed sequence of words; an informal notation is used allowing for optional and variable elements. In order to work with this notation, we formalize it into regular expressions. The following shows examples of the notation and its translation into regular expressions:

[1] Cf. `http://literaryquality.huygens.knaw.nl`

(1) a. Optional phrases: (. . .)
(Kijk,) dat bedoel ik nou.
(Look,) that’s what I mean.
`(Kijk , )?dat bedoel ik nou`
b. Open slots: [. . .]
Geen [bier] meer voor jou!
No more [beer] for you!
`Geen ([-\w+]* ){1,3}meer voor jou`
c. Variables: X, Y
Y, zoals X dan zou zeggen.
Y, as X would say.
`\w+ , zoals \w+ dan zou zeggen`
d. Alternatives: A/B
Daar zit een boek/artikel in!
That’s material for a book/paper!
`Daar zit een (boek|artikel) in`

The vast majority of cliché expressions in this dataset consist of full sentences (indicated by capitalization and sentence-ending punctuation). To avoid spurious partial matches, expressions with an initial capital either have to occur at the start of a sentence, or at the start of quoted speech: `(^|' )`. Similarly, expressions with sentence-ending punctuation have to end with a form of sentence- or quote-ending punctuation: `[.?!'"]`.[2] To increase recall, leading and trailing interjections are made optional. Accented characters (which can be used for emphasis) are also accepted in unaccented form. Where different forms of pronouns are possible, all alternatives are allowed (e.g., the possessive first personal pronoun *mijn* and its contraction *m’n*).

Some aspects cannot be translated precisely. When the alternatives span multiple words, the scope is not specified, so these have been edited manually. For lack of more specific criteria, and to ensure the regular expressions can be matched efficiently, we allow sequences of 1 to 3 words in open slots. Lastly, some clichés involve mini-dialogues; since we match on a per-sentence basis in the novels, these clichés will never be found.

After translating the clichés to regular expressions we remove duplicates. A handful of expressions are removed because they are too generic and generate too many matches (in these cases their cliché-hood depends on intonation or other contextual factors that cannot be automatically detected with textual matching). The resulting list of 5,771 patterns are counted across the whole corpus. We use Google’s `RE2` library to match the patterns efficiently using Deterministic Finite-State Automata.

4 Counting clichés in novels

Counting clichés in a corpus of texts results in a document-pattern matrix of occurrence counts. See Table 1 for the most frequent cliché expressions and expressions without any matches.

In order to get a picture of the overall rate of clichés, we sum the counts for all clichés in each novel, and normalize them for a fixed length (10,000 sentences) to get the cliché density of a text. This value is used to compute the correlation with the target value. See Figure 1 for the results. For both the literary ratings and quality there is a significant correlation (see the following section for how the strength of this correlation compares to other textual features). The plots show that most novels with a high number of clichés are non-literary. The highly literary novel *De Buurman* (the neighbor) by Voskuil is the strongest exception to this. This novel contains an exceptionally large proportion of dialogue, and the author is noted for his realistic depiction of arguments. In this case the use of clichés could well be a conscious stylistic choice (contrasting with the characterization of clichés as typically signalling ersatz creativity). On the other hand, novels with few clichés may or may not be literary. In other words, clichés are a

[2] A reviewer pointed out that these restrictions may bias the results toward specific authors or genres. In order to rule this out, we ran the experiments without these constraints, such that expressions only have to start and end at word boundaries. This did not have a substantial effect on the rates of clichés for any of the genres.

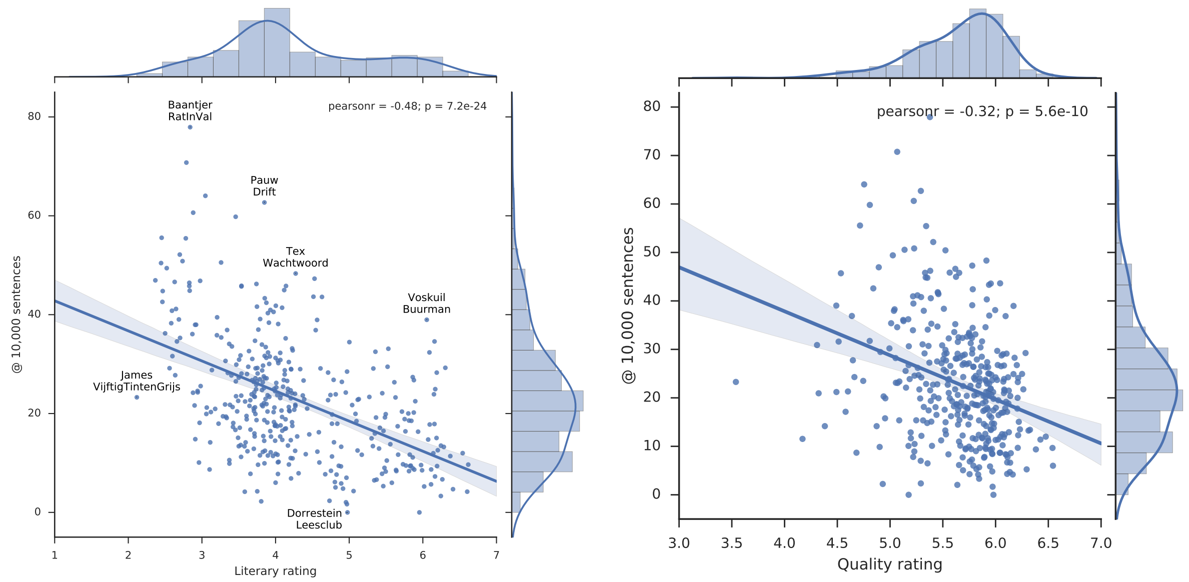

Figure 1: A simple regression of the number of clichés with literary ratings (left) and general quality ratings (right).

(2) a. Weet je het zeker ? (307)
Are you sure?
b. Is dat zo ? (245)
Is that so?
c. Waar heb je het over ? (231)
What are you talking about?
d. Laat maar . (140)
Forget it.
e. Is dat alles ? (139)
Is that all?
f. Dat meen je niet . (101)
You can't be serious.
g. Dat dacht ik al . (101)
I thought so.

(3) a. Je moet wel kunnen zien dat je op vakantie geweest bent.
It has to be clearly visible that you went on vacation.
b. Zit ik in de weg?
Am I in your way?
c. Er staat nergens dat het niet mag.
It doesn't say anywhere that it's not allowed.
d. Ik voel de bui al hangen.
I feel the storm is coming.

Table 1: Left: The cliché expressions with highest frequency.
Right: Examples of clichés without matches in any of the novels.

negative marker of literariness. For example, *50 shades of grey*, the least literary novel, has relatively few cliché expressions for novels with a similar rating, and falls below the regression line.

To compare the rate of clichés across genres and domains, Table 2 shows an overview aggregated across the main genres in the corpus and two reference corpora. The genres are derived from publisher-assigned categorizations of the novels. 'Fiction' are novels marketed as literary fiction. 'Other' is a mix of genres that did not fit in the other three and does not form a coherent category. We will therefore focus on analyzing Fiction, Suspense, and Romantic.

Especially the Romantic genre contains a larger number of clichés: twice as much as the Fiction genre. It also has more repetition of clichés than would be expected from the total number: the number of clichés that occur more than once is more than twice that of the Fiction genre. This is also confirmed by the lower type-token ratio—a ratio of 1 indicates that each type of cliché expression occurs only once; i.e., the lower the ratio, the more repetition. The violin plot in Figure 2 illustrates the genre differences and the variation within each genre. Fiction and Suspense, while having a different mean, show a similar distribution, with

	texts	sentences	clichés per 10,000 sents.	clichés per 10,000 sents., freq > 1	cliché type-token ratio
Novels	401	9,658.77	22.49	4.9	0.91
- Fiction	161	7,768.16	18.42	2.95	0.95
- Suspense	185	10,288.5	23.22	5.39	0.89
- Romantic	40	11,047.5	36.09	8.94	0.86
- Other	15	11,607.4	24.93	6.89	0.86
Reference					
- Written	1	52,157	0.38	0	1
- Spoken	1	70,277	9.25	4.41	0.63

Table 2: Overview of cliché occurrences. The rows with novels show the mean.

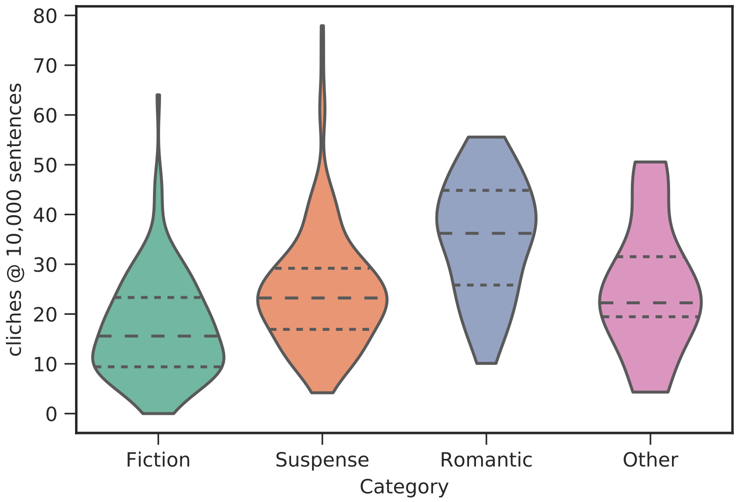

Figure 2: A violin plot of the number of clichés by genre.

outliers of novels that have more clichés (up to 65 and 80, respectively). The Romantic genre has a mean that is skewed closer to its maximum.

The reference corpora contain a much lower rate of clichés, which is probably attributable to their domain and either lack of informal dialogue (Lassy Small), or transcription of disfluencies and punctuation preventing matches (CGN).

5 Baselines features

We now consider simple baseline features, to characterize the language of cliché expressions and to see how clichés compare and relate to simpler features when predicting literary ratings. By simple we mean features that can be extracted from surface features without a trained model, as would be required for e.g., POS tagging, parsing, or named entity resolution. We consider the following four features:

MEAN SENTENCE LENGTH (number of tokens)

COMMON VOCABULARY the percentage of tokens part of the 3000 most common words in a large reference corpus. We use word counts from Sonar 500, a 500 million word corpus part of Lassy Large.

DIRECT SPEECH the percentage of sentences with direct speech punctuation.

COMPRESSION RATIO the number of bytes when the text is compressed divided by the uncompressed size. We use `bzip2` compression with the highest compression setting.

	% direct speech	Comp. ratio	% Common vocab.	mean sent. len.
Cliches	58.4	0.259	96.7	4.29
Novels	34.0	0.314	83.3	12.1
- Fiction	27.7	0.318	82.4	13.4
- Suspense	36.7	0.311	83.7	11.3
- Romantic	42.8	0.311	84.5	11.5
- Other	35.7	0.307	84.0	12.2
Reference				
- Written	14.3	0.373	73.4	9.96
- Spoken	(0.00)*	0.335	81.1	7.62

Table 3: Simple textual features compared across the clichés, novels, and reference corpora. The rows with novels show the mean. *The speech corpus does not transcribe direct speech with quotation marks.

The first two features represent a coarse measure of sentence and word complexity, respectively, similar to traditional readability measures (e.g., Flesch, 1948). If literary novels would be more difficult to read, we would expect a large correlation with these. Direct speech is relevant since most cliché expressions occur in dialogue, which is therefore a potential confounding factor. Compression ratio, by operationalizing repetitiveness, may also act as a proxy for cliché density, similar to the *n*-gram method we will look at in the next section (although it does not exploit an external reference corpus).

Table 3 compares the textual features for the clichés found in the novels, the novels themselves, and reference corpora. For each type of feature, clichés stand out as having simpler language: they consist almost exclusively of common words, are more repetitive (a lower compression ratio indicates more repetitiveness), and contain shorter sentences. The high percentage of direct speech for clichés indicates that the majority of matched clichés occurred as part of direct speech in the novels where they were found. Compared to the differences between genres, the contrast with clichés is much more dramatic.

These features can also be compared as predictors for the survey ratings; see Table 4. The degree to which a novel is seen as literary is better correlated with the textual features than general quality, as was observed before using clichés. This indicates that this difference between the predictability of literary and quality ratings is not specific to clichés. The simplest feature, sentence length, has the highest correlation, although cliché expressions have a still higher absolute correlation (-0.48 vs. 0.40).

	Ratings		Cliché
	Literary	Quality	density
Cliché density	-0.48*	-0.32*	1
Mean sentence length	0.40*	0.25*	-0.46*
Common vocabulary	-0.31*	-0.17*	-0.48*
Direct speech	-0.38*	-0.03	0.45*
Compression ratio	0.32*	0.05	-0.29*

Table 4: Correlation coefficients for simple textual features against survey ratings and cliché density. * indicates a significant result with $p \ll 0.001$.

Table 4 also shows the correlation coefficients of the number of clichés compared to the simple baseline features. All correlations are significant, and the results are in line with expectations: novels with longer sentences have less clichés, more dialogue and more simple vocabulary is associated with more clichés, and a novel that is more compressible (a low ratio) tends to have more clichés.

Since the simple features are correlated with the number of clichés, these features are not independent (i.e., they are collinear). On the other hand, the fact the clichés have a higher correlation with the literary ratings shows that clichés pick up on more than the above features. This suggests that in addition to the quantity of dialogue, the quality is also relevant, and the number of clichés appears to be a proxy for the latter. Concretely, literary and non-literary authors exhibit a different rate of clichés given the same amount of dialogue—literary authors tend to use less cliché expressions.

It could also be the case that the different genres employ different kinds of clichés. To investigate this, we inspected the top 10 most common clichés for each genre, and contrasted their frequencies. We normalize the frequencies, i.e,. the counts of each cliché are summed for each genre, and divided by the number of books, to obtain the expected frequency per novel in the genre. The clichés occurring in Fiction novels are common across all genres. A few clichés are common in both Suspense and Romantic, but not Fiction:

(4) a. Waar heb je het over?
What are you talking about?
b. Is dat zo?
Is that true?

Only the Romantic genre has several characteristic clichés that are rare in the two other genres:

(5) a. Dat meen je niet!
You can't be serious!
b. Dacht ik al.
I already thought so.
c. Doe niet zo raar.
Don't act so strange.
d. Ik red me wel.
I can take care of myself.

We conclude that, except for these outliers, the types and frequencies of clichés attested in the different genres are generally comparable.

An interesting question beyond the scope of this paper is whether certain novels may deliberately use clichés for certain parts or characters, as opposed to the clichés being part of the general style of the novel.

6 The n-gram distribution of clichés

Our approach of relying on a list of cliché expressions can be contrasted with Cook and Hirst (2013), who present an automated method of assessing whether a text is clichéd, using *n*-gram frequencies as a proxy. Their method uses *n*-gram frequencies from a large reference corpus and does not require an explicit list of cliché expressions. This implies that their method can only confirm a high density of cliché expressions, but not inspect the expressions themselves or analyze the number of types and counts for each expression. We will test their method to see if the results hold up with our cliché lexicon and corpus.

The method of Cook and Hirst (2013) is a corpus-based heuristic for cliché density based on the distribution of *n*-gram frequencies. We replicate their results on Dutch using *n*-gram counts from the 700 million word Lassy Large corpus (Van Noord, 2009); only counts of 2 and higher are included. The *n*-gram counts were extracted using Colibri (van Gompel and van den Bosch, 2016). Note that our reference corpus is significantly smaller than the one used by Cook and Hirst (2013), which consists of 1 trillion words.

We will use this reference corpus to compare three samples of text. The first two are written and spoken text from Lassy small and CGN. The third sample of text is the set of cliché expressions, i.e., a sample with 100 % cliché density. Since the cliché dataset itself consists of templates containing variable and optional elements, we extract *n*-gram counts from a list of 2457 cliché occurrences as attested in our corpus of novels.

Figure 3 shows a comparison of *n*-grams from the list of clichés versus samples from the spoken and written reference corpora. The plots are histograms, shown as line plots to facilitate comparison. Each histogram shows the absolute counts in a text sample (y-axis) of all *n*-gram tokens with a certain log count in the large reference corpus (x-axis). Only counts of 2 and up are shown (i.e., hapax legomena and

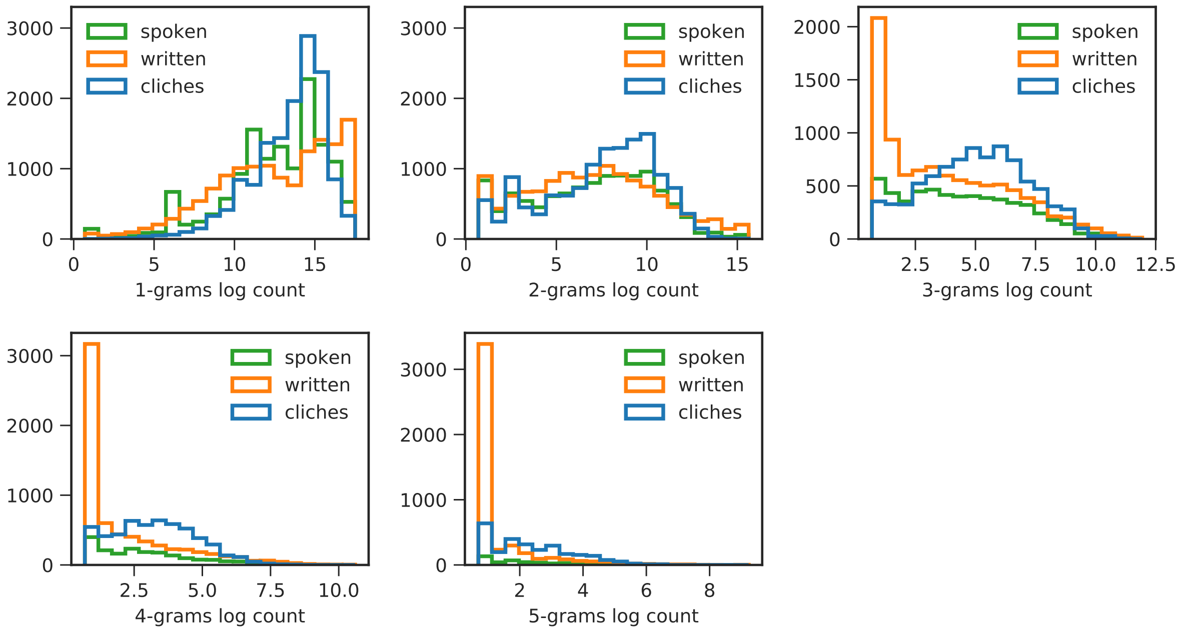

Figure 3: Histogram plots of *n*-gram distributions across several text samples. The x-axis bins the *n*-grams according to their frequency in a large reference corpus; the y-axis compares the absolute counts in several text samples.

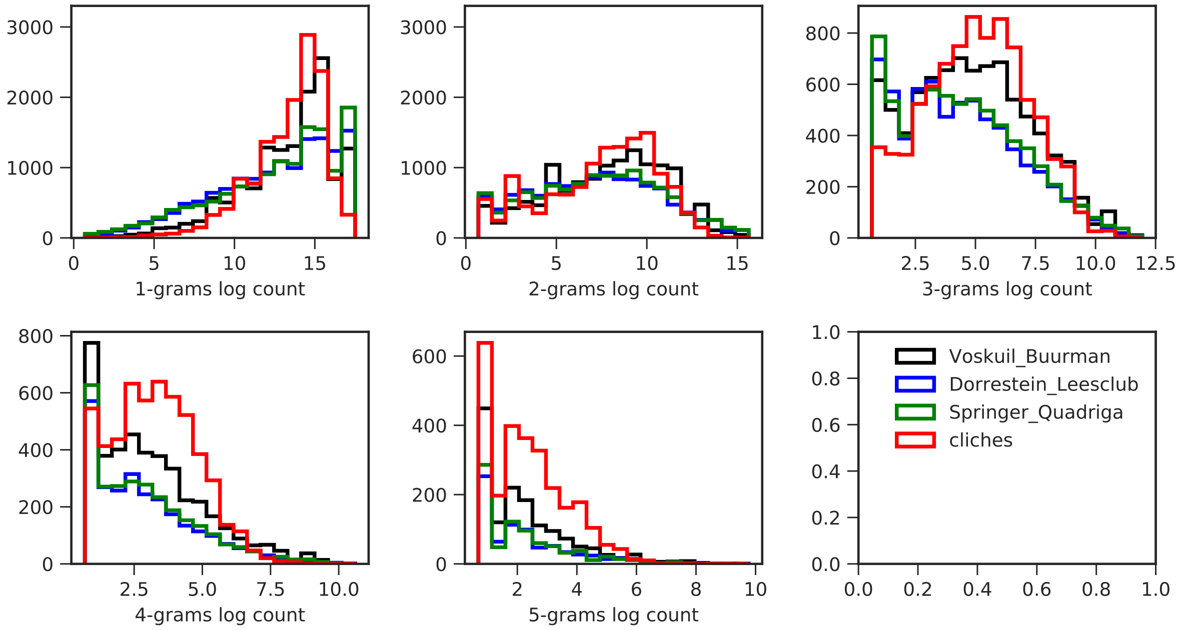

Figure 4: Histogram plots for the log count of *n*-grams in two novels without cliché matches (Dorrestein, Springer), one literary novel with many clichés (Voskuil), and the cliché lexicon.

n-grams not found in the reference corpus are not plotted). Each text sample, written, spoken, and clichés, contains approximately the same number of tokens, to ensure that the n-gram counts are comparable.

We observe characteristic peaks for the clichés in the mid to high frequency range, similar to those reported by Cook and Hirst (2013); this is most clearly visible for 2–4 grams. The distributions of the n-gram frequencies in the text samples are significantly different according to a Wilcoxon rank sum test ($p \ll 0.001$).

Differences with the graphs of Cook and Hirst (2013) are attributable to the fact that they use a larger corpus, and use only n-gram counts of 40 and up, while we use a threshold of 2 on a smaller corpus. Despite the differences clichés do appear to exhibit a readily identifiable frequency profile with n-grams s.t. $n > 1$.

Cook and Hirst (2013) argue that the n-gram heuristic is better because it is not possible to say whether the sample of 1,988 English clichés at their disposal has sufficient coverage. However, this is a precision-recall trade-off. A manually curated dataset may have limited recall, but will yield higher precision (i.e., will contain fewer false positive). Moreover, the n-gram technique cannot be used to detect whether a particular set of clichés is present in a large text, and the clichés cannot be located; the n-gram method is therefore coarse grained. Finally, while the n-gram distributions can be used as a proxy for detecting clichéd language, it is not clear whether the peaks reflect clichéd language specifically, or perhaps more generally informal and colloquial language.

In Section 4, we found that two novels do not have any matches from the cliché lexicon at all. To confirm that this is not caused by a lack of coverage in the cliché lexicon, we use the n-gram distribution method to confirm that these novels do not contain clichéd language. Cf. Figure 4 for the histograms of the two novels without cliché matches, a highly literary novel with many clichés, and the cliché dataset itself. The two novels without cliché matches indeed do not show the characteristic peaks of clichés, while the novel with many clichés has a similar shape as the cliché lexicon.

7 Conclusion

We conducted a large-scale corpus study of cliché expressions in novels and reference texts. We confirmed the intuition that novels judged as more literary tend to use less cliché expressions, and found a relatively robust effect. Clichés predominantly occur as part of dialogue and consist of short sentences with simple language. The cliché density of a text is correlated with several textual features such as sentence length, common words, and amount of dialogue, but the overlap is partial and the density of clichés is a better predictor of literariness than these other textual features. Non-literary genres such as suspense and romantic novels were found to contain more clichés.

The collection of Dutch cliché expressions exhibits a distinctive n-gram frequency profile, confirming results of previous research. Individual (unigram) word counts of clichés are unremarkable compared to reference texts, but higher order n-grams show a markedly different frequency profile.

Several interesting open questions remain. In general, what makes a particular expression a cliché expression? While we identified some basic characteristics of clichéd language, it would be interesting to try to model clichéhood directly. In the particular case of literariness, it is interesting to dive deeper into specific qualitative aspects, such as in what situations a cliché may be more or less desirable stylistically.

Acknowledgments

We are grateful to Wouter van Wingerden and Pepijn Hendriks for providing us with their dataset of clichés. Kim Jautze collaborated in this research at an earlier stage. Corina Koolen and three anonymous reviewers provided valuable comments. This work is part of The Riddle of Literary Quality, a project supported by the Royal Netherlands Academy of Arts and Sciences through the Computational Humanities Program. In addition, part of the work on this paper was funded by the German Research Foundation DFG *(Deutsche Forschungsgemeinschaft)*.

References

Chris Baldick. 2008. Literariness. In *The Oxford Dictionary of Literary Terms*. Oxford University Press, USA.

Pierre Bourdieu. 1996. *The rules of art: Genesis and structure of the literary field*. Stanford University Press.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Paul Cook and Graeme Hirst. 2013. Automatically assessing whether a text is clichéd, with applications to literary analysis. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 52–57.

Andreas van Cranenburgh and Rens Bod. 2017. A data-oriented model of literary language. In *Proceedings of EACL*, pages 1228–1238.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Maarten van Gompel and Antal van den Bosch. 2016. Efficient n-gram, skipgram and flexgram modelling with Colibri Core. *Journal of Open Research Software*, 4(1):e30.

Nidhi Kulkarni and Mark Alan Finlayson. 2011. jMWE: A java toolkit for detecting multi-word expressions. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 122–124.

Jan Mukarovsky. 1964. Standard language and poetic language. *A Prague School reader on aesthetics, literary structure, and style*, pages 17–30.

Gertjan Van Noord. 2009. Huge parsed corpora in Lassy. In *Proceedings of TLT7*, Groningen, The Netherlands. LOT.

Wouter van Wingerden and Pepijn Hendriks. 2015. *Dat Hoor Je Mij Niet Zeggen: De allerbeste taalclichés*. Thomas Rap, Amsterdam. Transl.: You didn't hear me say that: The very best linguistic clichés.

Ton van der Wouden, Heleen Hoekstra, Michael Moortgat, Bram Renmans, and Ineke Schuurman. 2002. Syntactic analysis in the spoken Dutch corpus (CGN). In *Proceedings of LREC*, pages 768–773.