# The Literary Pepsi Challenge: intrinsic and extrinsic factors in judging literary quality

Andreas van Cranenburgh, University of Groningen
Corina Koolen, Huygens ING

## Introduction

The project *The Riddle of Literary Quality*[1] aimed to find correlations between texts of novels and judgments of their literary quality. In other words: is the literariness of novels associated with or even explained by text-intrinsic properties? The 2013 *National Reader Survey* (NRS) collected a wealth of information on perceptions of literary quality of contemporary novels. It turns out that a machine learning model can predict the literary judgments based on the texts to a substantial extent; based on word frequencies and syntactic patterns, 61% of the variation in ratings of novels is predictable from purely textual features (van Cranenburgh & Bod, 2017; van Cranenburgh et al. 2018). This demonstrates that the text contains enough clues to distinguish literary texts from non-literary texts. However, we do not know to what extent humans rely on textual features when rating how literary a text is, since we collected judgments on whole novels by presenting the participants with the title and author of each novel. For the same reason it was not possible to identify the contribution and influence of particular aspects of the text. What we need is a blind experiment in which literariness is judged purely on the basis of text, without revealing any other information.

We therefore propose a new survey, based on fragments from the novels used in the NRS, to collect evidence that text-intrinsic characteristics play a role in ratings of literary quality, and investigate exceptions where we suspect various biases may play a role (cf. Koolen, 2018). The results will tell us more about how perceptions of literariness are formed and which particular textual aspects play a role. They will also enable a direct comparison between the performance of humans and a computer model on this task.

## Motivation

The NRS made clear that genre plays a role in judging literary quality. In the survey, Dutch respondents were asked to rate recently published novels on a scale of literary quality (1–7) and asked to motivate one of their ratings by an answer to the question "Why did you rate this book with the score for literariness as you did?" Respondents gave roughly three types of response, exemplified by Examples 1–3.

(1) "It is suspenseful, the storyline is perfect, but in a literary novel I expect a deeper layer."
(2) "It's chicklit"
(3) "Too light, simple, chicklit reads easily, but does not amount to much."

First, as expected, style and narrative structure are important (1). But in explaining why they found a novel *not* to be literary, respondents also often found it sufficient to refer to genre, without referring to textual qualities (2). It is possible that those textual qualities are implied. Some respondents did elaborate and explained low ratings in terms of both genre and style (3). However, genre exclusion may also point to bias. If a novel with a pink cover is excluded from a high rating without further explanation, what does that mean? Are we judging the text or repeating 'common sense' ideas on literary quality without questioning?

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

The first indication that extrinsic factors play a role are large gaps between the prediction of the computer model and reader judgments. The translation of *The sense of an ending*, for instance, received the highest average rating, 6.6, whereas the model predicted 5.4. This novel was awarded the Man Booker Prize the year before, which has probably influenced respondents. For *Eat, Pray, Love*, this was the other way around: the computer predicted 4.7, while readers gave it a 3.5.

A preliminary survey, conducted at a meeting of the KNAW Computational Humanities Program, showed that bias might play a role. We offered a handful of visitors five fragments (approximately one page of text), extracted from novels surveyed in the NRS. Respondents were asked: does this fragment originate from a novel with a high or low rating in the NRS? We anonymized the text by abbreviating names as initials. Remarkably, a fragment from Elizabeth Gilbert's *Eat, Pray, Love* was the only fragment that all respondents picked as a highly rated novel—which it was not.

Simkin (2013) conducted an online quiz, showing that average readers perform no better than chance at distinguishing a canonical (Dickens) from a non-canonical (Bulwer-Lytton) author. However, the fragments were short (3-4 sentences) and participants were not selected to have affinity with literature.

Given these results, it is interesting to test the influence of text and bias on literariness in a carefully designed survey.

## Survey setup

The two most important questions for the survey setup are who the participants will be, and what they will rate. We aim to select participants with literary affinity or expertise. To prevent the influence of author prestige, respondents should not see any metadata; nor do we want to cherry pick fragments. A double-blind setup with anonymized fragments will allow for this—we will set up a computer program to select equally sized fragments at fixed or random points from several novels. A trade-off needs to be made for fragment length; several sentences is too short, but more than a few pages takes too much time.

Instead of a 7-point Likert scale, as in the National Reader Survey, we will present pairs of fragments, and ask the rater which is the more literary one (pairwise ranking aggregation). This has the advantage of forcing the rater to make a concrete comparison, instead of expecting each rater to have an existing, well-calibrated scale. Rankings can be computed with the Elo rating system, the same system used to rank chess players. In addition, we can ask for a motivation.

We intend to run two experiments. The first experiment tests whether participants pass 'the challenge' and measures how humans perform at the task of recognizing literariness from unmodified text fragments. The second experiment introduces manipulations of fragments to confirm the influence of particular features, e.g., protagonist gender, sentence construction, topic. This approach is followed by Blohm et al. (2018), who present an experiment on lines from poetry rated for poeticity and grammaticality.

## References

Stefan Blohm, Valentin Wagner, Matthias Schlesewsky, Winfried Menninghaus (2018). Sentence judgments and the grammar of poetry: Linking linguistic structure and poetic effect. *Poetics*, vol. 69, pp. 41-56. https://doi.org/10.1016/j.poetic.2018.04.005

Andreas van Cranenburgh, Rens Bod (2017). A Data-Oriented Model of Literary Language. *Proceedings of EACL*, pp. 1228-1238. http://aclweb.org/anthology/E17-1115

Andreas van Cranenburgh, Karina van Dalen-Oskam, Joris van Zundert (2019). Vector space explorations of literary language. Language Resources and Evaluation. https://doi.org/10.1007/s10579-018-09442-4

Corina Koolen, *Reading beyond the female: the relationship between perception of author gender and literary quality*. Amsterdam: University of Amsterdam.

Mikhail Simkin (2013). Scientific evaluation of Charles Dickens. *Journal of Quantitative Linguistics*, volume 20, issue 1, pp 68-73. https://doi.org/10.1080/09296174.2012.754602