I Catching

Computationally Operationalising Narrative Perspective for Stylometric Analysis

Lisanne M. van Rossum[1,2], Joris J. van Zundert[1,3], and Karina van Dalen-Oskam[1,4]

[1]Huygens Institute, Royal Netherlands Academy of Arts and Sciences, Amsterdam
[2]Humanities Faculty, Utrecht University, Utrecht
[3]Università Ca' Foscari, Venice
[4]Humanities Faculty, University of Amsterdam, Amsterdam

One of the aims of literary research is to explain different interpretations and preferences of literary language. Narrative perspective is one of the many interlinked features associated with literary style. An automated way to measure a text's perspective is thus essential for the computational analysis of a large corpus of literary fiction. However, no such measure of narrative perspective exists to date. In this paper we explain a first step in developing such a measure. We take two approaches to operationalise narrative perspective as a textual predisposition towards either first or third person perspective. The first measure is based on a baseline machine learning approach, the second is based on computing the ratio of combinations of pronouns. The latter approach results in an heuristics based measure which we call the I-index. Our research demonstrates that narrative perspective can be predicted with high accuracy. We reflect on our results to broach the advantages and pitfalls of both methods.

1 Introduction

In computational literary studies, various stylometric methods are applied. To verify authorship, for instance, or to map stylistic differences between texts, authors, genres, etc. In the project The Riddle of Literary Quality, stylometry was used to investigate the concept of literary quality with the main objective to establish which linguistic features are more prominent in novels that readers rate as highly literary, as opposed to novels that are rated less high. The project demonstrated the combined importance of a novel's textual qualities for its perceived literariness, qualities such as: narrative structure and plot (Koolen et al., 2020), semantic complexity and lexical originality (van Cranenburgh, 2016), and the work's embeddedness in social structures such as genre and author gender (Koolen, 2018). One of these textual qualities is narrative perspective. The corpus used for the Riddle of Literary Quality consisted of 401 novels. In several experiments, in which sets of novels with different ratings on the scale from high to low literariness were compared, pronouns showed up at the top of the lists of preferred words (van Cranenburgh et al., 2019). In order to establish if and how these preferences relate to perceptions of literary quality, it is necessary to find out if their prevalence can be linked to a difference in narrative perspective. We would like to be able to determine if these pronoun preferences are related to prevailing perspective, which we operationalise as being narrated in either first or third person. However, clear computational measures to determine perspective do not exist. In this paper we build on the work done in The Riddle of Literary Quality and take a first step in the process to develop such a formal measure.

2 Theoretical Background

Establishing a rudimentary definition of narrative perspective is important when we want to answer questions such as whether novels with a first person narrator significantly differ in style from novels with a third person narrator, and if a text's narrative perspective correlates with its perceived literariness. Especially when dealing with large size corpora, as in the case of the Riddle of Literary Quality project, we need to be able to establish narrative perspective computationally and reliably. In this paper we present and evaluate a first methodological step to do exactly this.

Narrative perspective is a highly complex and multifaceted narratological concept (Porter-Abbott, 2008, pp.311–12). It depends on many aspects such as narrator voice, narrative time, dialogue, reliability of narrator, and so forth: "[t]o mention only a few factors: one's position with respect to the perceived object, the angle of the light, the distance, previous knowledge, psychological attitude towards the object – all these things and more affect the picture one forms and passes on to others. Storytelling is one form of such passing on," Mieke Bal writes in her *Introduction to the Theory of Narrative* (Bal, 2017, p.133). As non-binary and abstract features, these factors of storytelling currently cannot be easily computationally determined, an issue raised by Luc Herman and Bert Vervaeck in their criticism of positivistic narrative processing: "In striving for the disambiguation typical of the hard sciences, literary narratology might lose its relevance as a tool for the development of interpretations that ideally keep the complexity of a text intact. In any case, empirical narratology turns the narrative and literary dimensions of the text into quantitative data." (Herman and Vervaeck, 2005, p.105).

In contrast to the sceptical take by Herman and Vervaeck we do think it is possible to operationalise literary complexities in a way that enhances the relevance of literary hermeneutics, while remaining respectful of the literary texts themselves. We think that taking small iterative heuristic steps combined with machine learning validation on large ground truths is a feasible way forward. As we will show in the methodology section, these heuristic steps guarantee that the complexity of each literary text is directly taken into account.

In this paper we report on our work to establish if first and third person narrative voice can be reliably determined automatically. Highly accurate formal measures to determine story perspective computationally do not exist yet. Related computational narrative work has focused elsewhere: on dialogue detection (cf. Byszuk et al., 2020), narrative generation (cf. Mani and Hirst, 2013), or matters of authorial perspective (e.g.

Piper, 2015), for instance. Research so far has not yielded results that allow simple and straightforward detection of story perspective, let alone more subtle related aspects such as perspective shift within the same story, chapter or paragraphs, unreliable perspectives, and mixed perspectives. One complicating issue, for instance, is that perspective and its subtle guises are intricately tied to the distinction between dialogue and non-dialogue text that both can take various ambiguous mixed shapes. Acknowledging this complexity Brunner (2013, 2019) combines machine learning and rule based approaches to automatically detect speech, thought and writing representation. Brunner's method currently does not yield an accuracy of distinguishing dialogue from non-dialogue that enables automated precise labelling of samples. Brunner reports an F1 score of 0.87 on direct speech recognition, and scores lower than 0.5 for other forms of speech and dialogue, e.g. indirect speech, free direct speech and narrator text (Brunner, 2019, p.242). This means that automatic dialogue detection currently cannot help us in detecting narrative perspective and once more demonstrates that inferring narrative perspective while acknowledging the many complexities of literary style is not a straightforward machine learning task. Because machine learning methods that can measure perspective directly are not available we have opted to first tease out the many less complex features and their appropriate computational measures that will ultimately allow us to determine narrative perspective automatically with sufficient accuracy. As a first step we focus on pronouns as features that are related to narrative perspective.

In absence of precise machine learning methods for now we rely on manually labelling a ground truth (in the "NLP sense" of labelled baseline data for computer analysis). If we want to develop a measure of narrative perspective we are forced to start small, modest, and we should progress with small steps keeping computational and hermeneutic evaluation tightly integrated, as we reason elsewhere (van Zundert and van Dalen-Oskam, 2019). Machine learning results often agree with human intuitions. However, it remains unclear whether the underlying mechanisms of machine learning are comparable with the human cognitive process. Insufficient understanding of the heuristics of machine learning or data bias may lead to poor explanations for mechanisms behind real world phenomena (for especially egregious examples cf. Coalition for Critical Technology, 2020 and Liberman, 2010). The issue of the human comprehensibility of methodology remains a topic of much debate in the field of Digital Humanities, one that we would like to contribute to by integrating machine learning and computational methodologies as mutual checks and balances.

Our proposal for a first measure is intentionally simple and straightforward, because we want to better understand the interpretative process in the human reader. Within this first step towards making perspective computationally measurable we focus on pronoun use. Pronouns are semantically and syntactically a more clear and unambiguous category than other phenomena related to perspective, such as dialogue and emotional valence. This makes pronouns a good baseline point of entry for both machine learning and human heuristic approaches to studying perspective, as others have also recognised (e.g. Andresen and Knorr, 2020). For that reason we choose to use the analysis of pronouns as a first step in both hermeneutically and computationally understanding the measurement of perspective in literary texts.

3 Method

We applied two approaches to our development of a measure to determine perspective in fiction computationally, so as to be able to thoroughly validate our results. The first measure is a machine learning approach, where we try to let the "data speak for itself" bottom up. The second approach is narratologically informed and based on computing the ratio of combinations of pronouns.[1]

Both methods use the same set of data of 1001 fragments of novel texts. Samples were taken from the novels from The Riddle of Literary Quality project corpus of 401 bestselling and most frequently borrowed novels in the Netherlands between 2007 and 2012. The samples were taken from across the entire Riddle corpus, although some novels contributed more samples than others. Samples were not chosen randomly but manually so that none contained dialogue, and were unambiguous consecutive examples of narrative text. Multiple fragments were chosen from novels that featured more than one narrator to reflect the narrative diversity of the corpus. Token length of samples varied around 250 tokens (minimum of 229, maximum of 314, with a mean of 258.11 words and standard deviation of 10.10). All samples were converted to lowercase and all punctuation was removed. A typical example of a text would look like the sample depicted in example 1.

> verhooff daalde tijdens kantooruren af in de grijze catacomben van de dagelijkse routine waar het vergaderen ervoor zorgde dat alles nóg een tint grijzer werd vergaderen met de gemeente amsterdam vergaderen met architecten en interieurontwerpers over bijzonderheden van de nieuwbouw en renovatie van het hollands museum vergaderen met **zijn** conservatoren met projectontwikkelaars aannemers bouwplaatsmanagers opzichters subsidiegevers lobbyisten analisten fiscalisten bestuursleden commissieleden deelraadsleden verenigingsleden een enkele keer bekeek **hij** een kunstwerk er waren de vroege ochtenden en soms lange avonden in **zijn** museumvleugel **hij** sliep slecht soms maakte **hij** in **zijn** eentje een wandeling door het lege gebouw heel soms zag **hij** **zijn** zonen verhooff betrapte zich erop te verlangen naar een doodgewoon appartement met een doodgewone keuken en dito badkamer en dan waren er de verplichte reizen naar de biënnales van são paulo en venetië **zijn** aanwezigheid was vereist op openingen van tentoonstellingen in verschillende musea in europa: wining and dining in londen parijs düsseldorf madrid san francisco **hij** deed twee aankopen die de kranten haalden **hij** gaf gemiddeld drie interviews per maand aan buitenlandse dagbladen of tv zenders over het verloop van de verbouwing die geheel volgens nieuw ingestelde nederlandse traditie vertraging opliep op een zeker ogenblik moest de nieuwe vleugel tegen de vlakte de dag brak aan dat er twee verhuiswagens het bouwterrein op reden die een parkeerplek vonden tussen de containers cementwagens hijskranen en hei installaties samen met de opzichter een enthousiaste gezette vent met een oversized snor die archaïsch opkrulde bij de uiteinden liep verhooff

[1] Code for both approaches is open source (MIT license) and available in Github: `https://github.com/jorisvanzundert/riddle_ikindex`.

het zoveelste rondje langs de enorme bouwput die de laatste maanden **zijn** uitzicht had bepaald

(Source: Joost Zwagerman (2010), *Duel*. Amsterdam, Antwerpen: De Arbeiderspers.)

Example 1: Example of one of the full samples of Dutch-language literary text used, without punctuation and all lowercase. (Pronouns in bold.)

The machine learning approach used a three layer feed forward convolution neural network with a standard Keras Python implementation backed by Tensorflow. We used a take-one-out approach to training and testing. Thus, for each sample we trained the model on the entire set of samples leaving out that specific sample, after which the model made a prediction, i.e. whether the tested sample was first or third person perspective. Essentially this means that we apply a plain vanilla machine learning binary classification approach to the problem of identifying first person and third person narrative perspective.

The pronoun ratio based approach we started by using equation (1). This equation computes the ratio between first-person pronoun singular "ik" (transl. "I") and third-person pronouns singular "hij" and "zij" (transl. "he", "she").

$$I = \frac{n_{ik}}{1 + n_{hij,zij}} \tag{1}$$

This basic formula worked well but we found through iterative testing that using more pronouns would increase accuracy. Iterative evaluation also served to show that reflexive pronouns ("zichzelf", "haarzelf", "onszelf"; transl. "himself", "herself", "ourselves") should not be counted as they are essentially redundant and therefore inflate the "he", "she", and "we" count of the pronouns they refer to. Eventually we arrived at the equation depicted as equation (2), or Van Rossum's I-index.[2] Van Rossum's I-index computes the ratio of pronouns associated with first person perspective to a fuller set of pronouns found in a text, including second-person, plural, and possessive pronouns. In this form the I-index resembles an approach presented by Jockers for English (Jockers, 2014). Apart from the obvious difference in language domain the main difference with that method is in our use of all pronouns. Jockers' method uses first and third-person pronouns only. Jockers does not provide a formal validation, but for the Dutch context the use of all pronouns yielded better results than using selective pronouns (cf. section 4).

$$I = \frac{n_{(ik,wij,we,me,mij[n],ons,onze)}}{1 + n_{(ik,we,we,me,mij[n],ons,onze,hij,zij,ze,je,jou[w],haar,zijn,jullie)}} \tag{2}$$

By sorting the labelled list of fragments from highest to lowest I-index several categorisation errors came to the fore. Several of these were manual labelling mistakes

2 As a gesture of appreciation for her sound and creative contribution to developing this measure, well beyond the baseline for any expected contribution of a master student to a research project, we have chosen to name this equation "Van Rossum's I-index".

on part of the researcher where high I-indices were labelled as third person narratives. Other irregularities alerted us to the many pronominal forms that narrative perspective can take, such as Renate Dorrestein's novel *De leesclub* ("The Book Club"), which is written from a first person plural, or the "we" perspective that the titular book club implies (see example 2). We include full-length excerpts in this article because they illustrate how multifaceted narrative perspective is beyond the sentence level. The second excerpt, from *Mama Tandoori* for example, features both first-person possessive pronouns for focalization and third-person pronouns for description.

> Daar zaten **we** dan. Na drie lange dagen op het eiland was de toestand onveranderd dezelfde. De schrijver, laten **we** er geen doekjes om winden, lag nog steeds voor pampus. Het regende en het regende. Op de schaarse momenten dat het droog was, joegen er dikke mistflarden langs de grijze lucht. Het werd eb, het werd vloed. **We** leefden van zeewier, algen en een enkele kwal, door Johanna op de aangespoelde tweepitter bereid. Het was gezond en vitaminerijk voedsel, maar wel eenzijdig, al bracht Jo zoveel mogelijk variatie aan door de kelp beurtelings te koken, te bakken en te blancheren. Het gele mos, waar Tillie meteen al **haar** twijfels over had gehad, was keihard en oneetbaar gebleken. Verder groeide er op het hele eiland niets, zelfs geen distels. Gelukkig was bijna **onze** hele voorraad The Famous Grouse aangespoeld. Er waren maar een paar flessen op de rotsen gesneuveld.
>
> Ook van de Maizena waren diverse brokstukken aan land gekomen. In een afgebroken mast hadden **we** meteen op het hoogste punt van het eiland Leonies meest kleurrijke Marimekko gehesen. Met behulp van touw en zeil hadden **we** vervolgens een afdak in elkaar gezet, waaronder **we onze** kleren en onszelf konden drogen. Hadden **we** niet op een kale rots gezeten, dan waren **we** begonnen met de aanleg van een kleine moestuin, zoals bekend uit de literatuur. Robinson Crusoe, dat werk.
>
> Wat **we** het meest misten, behalve een boek, was wcpapier. En Annabel betreurde het teloorgaan van **haar** pincet.
>
> (Source: Renate Dorrestein (2010), *De leesclub*. Amsterdam, Antwerpen: Contact)
>
> (Translation)
> *There **we** were. After three long days on the island, the situation remained unchanged. The writer, let's not beat around the bush, was still out cold. It rained and it rained. On the rare occasions that it was dry, thick patches of fog swept across the gray sky. The tide ebbed and flowed. **We** lived on seaweed, algae and the occasional jellyfish, prepared by Johanna on the washed-up two-pitter. It was healthy and nutritious food, but it was monotonous, although Jo added as much variety as possible by alternately cooking, baking and blanching the kelp. The yellow moss, which Tillie had immediately had felt unsure about, had turned out to be hard and inedible. Nothing else grew on the whole island, not even thistles. Fortunately, almost our entire supply of The Famous Grouse had washed up. Only a few bottles had broken on the rocks.*
>
> *Various pieces of debris from the Maizena had also come ashore. On the highest point of the island **we** immediately hoisted Leonie's most colorful Marimekko on a broken mast. With the help of rope and tarpaulin we then fashioned a shelter under which we*

could dry our clothes and ourselves. Had we not been sitting on a bare rock, ***we*** *would have started cultivating a small vegetable garden, a known literary trope. Robinson Crusoe, that business.*

What ***we*** *missed most, other than a book, was toilet paper. And Annabel mourned the demise of her tweezers.*

(Source: Google translate with our own manual corrections.)

Example 2: Example of one of the full samples of Dutch-language literary text used, without punctuation and all lowercase. (Pronouns in bold.)

Another labelling discrepancy occurred with excerpts from Ernest van der Kwast's novel *Mama Tandoori*, a novel in which the narrator uses overwhelmingly possessive pronouns to paint a loosely autobiographical portrait of his mother (see example 3).

Mijn moeder drinkt thee met de moeder van de patiënt. Het is stil, een huis van verdriet. **Je** kunt een traan horen vallen. Dan wordt **mijn** moeder naar de patiënt geleid. **Ze** neemt **hem** op, **ze** ziet **zijn** donkere, fonkelende **haar,** de kaarsrechte scheiding. Het kapsel van een filmster. **Ze** wisselen geen woorden, **ze** zwijgen. **Mijn** moeder wast de wonden van de kapitein, verschoont het verband. Rajesh Mudgal is lijdzaam. **Hij** bijt op **zijn** kaken, **hij** denkt aan de zee. Zoute regen, golven grijs als olifanten.

Langzaam, heel langzaam, herstelt de patiënt zich, hervindt **hij** zijn krachten. **Mijn** moeder ziet het aan **zijn** ogen. Kleine rimpels in de ooghoeken, een schittering in het zwarte meer van **zijn** pupillen. De ogen van Rajesh Mudgal lachen. Op een ochtend gaat **ze** ernaar verlangen. Het is er ineens, het verlangen, en het is warm en bliksemsnel, overal en nergens. Nog nooit is **ze** zo helder wakker geworden, alsof de zon van de nieuwe dag regelrecht in **haar** geest wordt uitgestort. **Haar** lichaam is van licht, **haar** vingertoppen tintelen.

Ze vraagt zich af wat Rajesh Mudgal voelt. **Mijn** moeder probeert **zijn** gezicht te lezen. **Ze** ziet groeven. Hoe langer **ze** kijkt, hoe meer lijnen ze ontdekt, plooien van pijn. In **zijn** voorhoofd, om **zijn** mond, tussen **zijn** wenkbrauwen. **Ze** wordt overvallen door medelijden. **Mijn** moeder pakt **zijn** hand vast, de hand met vijf stompjes als vingers. **Ze** knijpt er zachtjes in en het is alsof **ze** een vonk voelt. Van schrik laat **ze** de hand los. Opeens beseft **ze** de situatie. **Zij** is een verpleegster, **hij** is afkomstig van een rijke familie. **Hij** is invalide, **zij** is een vrouw in de kracht van **haar** leven.

(Source: Ernest van der Kwast (2011), *Mama Tandoori*. Amsterdam: Nijgh & Van Ditmar)

(Translation)
My *mother drank tea with the patient's mother. It was quiet; this was a house of sorrow. You could hear a tear drop. Then* ***my*** *mother was taken to the patient. She scrutinised*

__him,__ seeing __his__ dark, gleaming hair, the perfectly straight parting: the hair of a film star. __They__ didn't exchange any words. __My__ mother cleaned the captain's wounds, changed __his__ bandages, in silence. Rajesh Mudgal endured it all. __He__ clenched __his__ jaw, thinking of the sea: salty rain, waves as grey as elephants.

Slowly, very slowly, the patient improved, recovered __his__ strength. __My__ mother could tell by __his__ eyes: new tiny wrinkles in the skin around __their__ corners, a brilliance in the black lakes of __his__ irises. Rajesh Mudgal's eyes were smiling. One morning, __she__ began to long for __them.__ Suddenly it was there, the longing, and it was warm and quick as lightning, everywhere and nowhere at once. Never before had __she__ woken with such clarity, as if the new day's sun had been poured directly into __her__ soul. __Her__ body was made of light; __her__ fingertips tingled.

__She__ wondered what Rajesh Mudgal might be feeling. __My__ mother tried to read __his__ face. __She__ saw furrows. The longer __she__ looked, the more lines, the more creases of pain, __she__ discovered: in __his__ forehead, around __his__ mouth, between __his__ eyebrows. __She__ was overcome by compassion. __My__ mother took __his__ hand, the hand with the five stumps for fingers. Squeezing it softly, __she__ seemed to feel a spark. Startled, __she__ dropped the hand. Suddenly, the situation dawned on __her.__ __She__ was a nurse; __he__ was the scion of a wealthy family. __He__ was an invalid; __she__ was a woman in the prime of __her__ life.

(Source: Van der Kwast, E. (2017) *Mama Tandoori* (transl. L. Vroomen). Melbourne; London: Scribe.)

Example 3: Excerpt from *Mama Tandoori*. Words counted as pronouns into the index in bold. This text was manually labelled as first person perspective and recognised by an earlier version of the I-index as third person perspective. After fine-tuning, the I-index rightfully recognised it as first person. Note that we did not apply linguistic parsing but mere word form matching. This results in negligible amounts of false positives, such as the first "haar" in this sample, which actually means "hair" not "her". This is a detail of parsing, it does not affect the properties of the I-index itself.

Both texts in example 2 and 3 feature a variation on a first person perspective, albeit in a different form than the initial version of the I-index accounted for which resulted in their mis-classification as third person perspective. We examined these outliers in our training material to expand the I-index into new descriptive directions, to also account for narrative perspective in its multiple personal pronouns and possessive pronoun forms. As such, we used the texts that resisted our initial method as input to refine the I-index.

4 Results

Both methods were tested and validated on 1001 selected full text fragments of Dutch fiction non-dialogue narrative that were labelled manually for first and third person perspective. Both measures were verified by leave-one-out testing, thus yielding 1001 observations each. We found that both approaches are highly effective in determining narrative perspective, yielding a F1 harmonic mean of 0.97 for the machine learning approach (figure 2) and a perfect 1.00 score for the pronoun ratio approach (figure 1).

Pictured in figure 1 are the results from the I-index's classification of our corpus of 1001 narrative text excerpts. The swarm plot shows that the majority of entries labelled as third person narrative perspective (orange dots) is located at the lower end

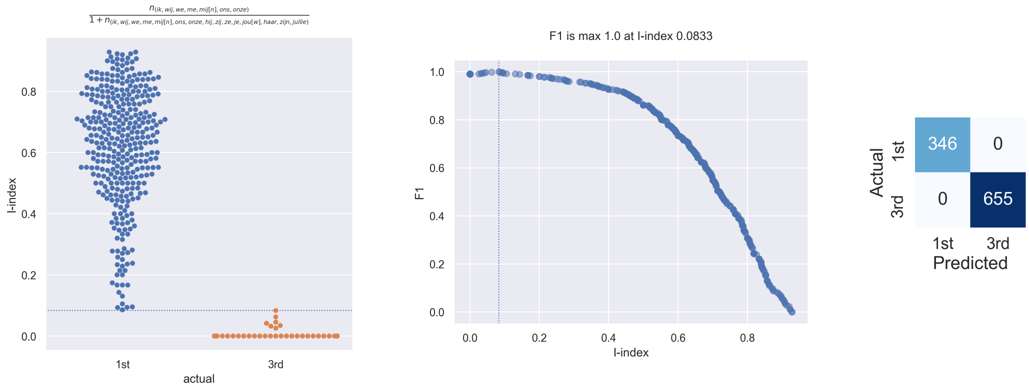

Figure 1: Swarm plot of predictions, F1 curve, and confusion matrix for the I-index classification approach.

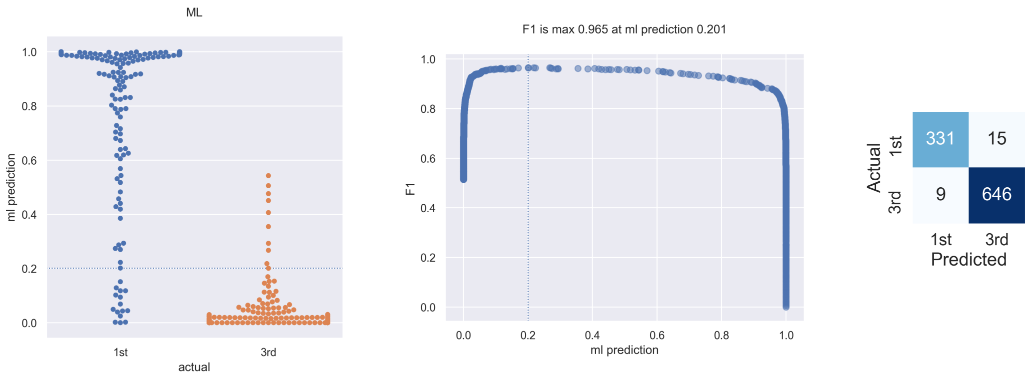

Figure 2: Swarm plot of predictions, F1 curve, and confusion matrix for the machine learning approach.

of the spectrum with a score of zero being the most frequent, which is reassuring for a measure that takes first-person pronouns as a baseline. Texts labelled as first person perspective (blue dots) are expectedly feathered out over the higher scores on the I-index scale, in the graph appearing at the moment where texts are no longer labelled as third person narratives. The threshold rate sits at an I-index of 0,0833, indicated by the dotted line. The F1 curve visualises this predictive point of optimal certainty and shows that when that threshold would be shifted upwards in the first figure, it would result in more classification errors. The confusion matrix on the right-hand side shows the model's error rate in prediction, which sits at zero for both our 346 first person and our 655 third person perspective texts, meaning that the model has made no mistakes in classifying the fragments according to our manual labelling of first and third person.

Figure 2 visualises the results from the predictive machine learning model to contrast our computational I-index method. For the machine learning model, the swarm plot points to less clear-cut distinction between texts classified as first person (blue dots) and those classified as third person perspective (orange dots). The overlap occurring at both sides of the dotted cut-off line points at the algorithm's inaccuracies in the categorisation of texts. As such, the predictive model's maximum certainty sits somewhat higher at a machine learning index of 0.2. The F1 curve indicates that the machine learning model is more consistently certain of a binary judgment than the

I-index, looking at the vertical lines sitting at both the 0,0 and 1,0 points on the x-axis, and that it is somewhat more confident than our computational measure, hovering on the higher end of the graph as opposed to the gradual downwards slope of the I-index model. The third error metric, however, confirms that despite this increased confidence, the algorithm has made more mistakes in classification; 15 excerpts that were manually labelled as first person perspective were incorrectly deemed to be third person narrative texts, and 9 excerpts that were manually labelled as third person perspective were faultily taken as first person perspective.

5 Discussion

As a first step in developing some baseline approaches to computationally describing narrative perspective present in text, the measures both yield perfect to near-perfect results. However, we prefer the pronoun ratio based approach above the machine learning one because it more clearly indicates and reveals the linguistic and narratological properties sought after and associated with the narratological inference.

Additionally, we would like to underline the versatility and flexibility of this computational measure, as it can be simply adapted to expand the range of pronouns that it accounts for, as evidenced by our own refinement process, and it can be rearranged to accommodate different narratological foci — a third person index, for instance, or an index of plural perspective. Although of minor importance, it could also be noted that the statistics based method outperforms the machine learning method as to speed.

In a sense our approach is consciously naive because of our decision to establish a baseline measure, for which we have explained our reasons in the theory section. Obviously perspective in fiction is more complex than the dichotomy between first and third person view. See, for instance, Fludernik (1995) on unfamiliar pronominal strategies such as second-person pronouns of address and impersonal pronouns in experimental fiction. We purposefully designed this computational measure as an index to reflect our intention to approach literary perspective as a spectrum, rather than as a binary system. In its current form the index discriminates between first and third person perspective. But the index is designed in such a way that it is open to adaptation for multiple and eventually mixed narrative perspectives. This open design also creates the possibility of the index's use for narratological shifts within the context of the full text of novels. This is part of our future work.

It must be noted that our measures have been tested on an unambiguous gold standard. We used narrative text in the sense of text without quoted or directly implied dialogue. In unmediated fiction narrative which features, for instance, indirect free speech, it is not readily clear whom the narrative perspective is most closely associated with (either speaker or narrator) and how mixed perspective in such cases could and should be measured. Further validation should also be applied to ensure our sampling process and choices have not introduced bias. For instance, due to pragmatic reasons many samples were taken from the first parts of novels because that is where consecutive narrative text generally occurred more often than in later parts of the texts. It is arguable that narrative parts (i.e. non-dialogue) at the start of a novel are of a different length compared to such parts in the latter half of a story, while also their style may differ. Furthermore we should expand our evaluation to examine if the found cut off rates for predictions are stable over different corpora that are not as clearly a gold standard construction as applied in our current tests.

Another conscious naivety in our approach is the establishing of the harmonic mean.

Because both observations, predictions and labels were fully known in this case, we could simply compute the cut off value for predictions that would yield the highest F1 score. In unsupervised situations there is obviously no way to do this. Further experimentation should determine if these cut off values are both realistic and usable in different situations.

6 Conclusion

In this paper we have introduced two methods to automatically categorise the narrative perspective of a text sample as a point on the line between first person or third person perspective. We have focused on the novels in modern Dutch that were part of the corpus of The Riddle of Literary Quality project, on the one hand to eventually establish a possible link between narrative mode and pronoun use, and on the other, to explore whether narrative perspective in the most basic of computational definitions is one of the features playing a role in the perception of literary quality. The first of our two measures was a machine learning approach and the second was based on computing the ratio of combinations of pronouns. We have demonstrated that both approaches can predict narrative perspective with very high accuracy. Our fine-tuned calculation of pronoun ratio - which we named Van Rossum's I-index - even yielded a perfect score. Because the I-index is transparent and easy to modify to accommodate other perspectives we prefer this approach. In its current version, however, it is only an entry point explaining the socio textual significance of narrative perspective. We are working on follow-up research in which we will develop this first step into a more comprehensive method that will be very useful in stylometric research on large and diverse corpora of literary texts in Dutch. We expect our approach may also be useful in other languages with comparable and unambiguously determinable pronoun identifiers. The situation for languages with verb or noun inflexion to indicate pronouns, however, might require further adaptation.

References

Andresen, M. and D. Knorr
2020. Exploring the Use of the Pronoun I in German Academic Texts with Machine Learning. In *Proceedings of "Workshop InfDH 2020: 'Methoden und Anwendungen der Computational Humanities'"*, Berlin, Bonn. Gesellschaft für Informatik e.V.

Bal, M.
2017. *Narratology: Introduction to the Theory of Narrative*, 4th edition, 1st edition 1985 edition. Toronto, Buffalo, London: University of Toronto Press.

Brunner, A.
2013. Automatic recognition of speech, thought, and writing representation in German narrative texts. *Literary and Linguistic Computing*, 28(4):563–575.

Brunner, A.
2019. Redewiedergabe – Schritte zur automatischen Erkennung. *Zeitschrift für germanistische Linguistik*, 47(1):216–248.

Byszuk, J., M. Woźniak, M. Kestemont, M. Leśniak, W. Łukasik, A. Šeja, and M. Eder
2020. Detecting direct speech in multilingual collection of 19th century novels. In

Proceedings of the LREC 2020 1st Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2020), R. Sprugnoli and M. Passarotti, eds., Pp. 100–104, Paris. European Language Resources Association (ELRA).

Coalition for Critical Technology
2020. Abolish the #TechToPrisonPipeline: Crime prediction technology reproduces injustices and causes real harm.

Fludernik, M.
1995. Pronouns of address and 'odd' third person forms: the mechanics of involvement in fiction. In *New Esssays in Deixis: Discourse, Narrative, Literature*, K. Green, ed., volume 103 of *Costerus New Series*, Pp. 99–129. Amsterdam, Atlanta: Rodopi.

Herman, L. and B. Vervaeck
2005. *Handbook of Narrative Analysis*, Frontiers of narrative. Lincoln, London: University of Nebraska Press.

Jockers, M. L.
2014. Simple Point of View Detection.

Koolen, C., K. van Dalen-Oskam, A. van Cranenburgh, and E. Nagelhout
2020. Literary quality in the eye of the Dutch reader: The National Reader Survey. *POETICS*, 79:101439.

Koolen, C. W.
2018. *Reading Beyond the Female: the relationship between perception of author gender and literary quality*. Phd, University of Amsterdam, Amsterdam.

Liberman, M.
2010. More on "culturomics".

Mani, I. and G. Hirst, eds.
2013. *Computational Modeling of Narrative*, number 18 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Piper, A.
2015. Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel. *New Literary History*, 46(1):63–98.

Porter-Abbott, H.
2008. *The Cambridge Introduction to Narrative*, Cambridge Introductions to Literature, 2nd edition. Cambridge, New York: Cambrdige University Press.

van Cranenburgh, A.
2016. *Rich Statistical Parsing and Literary Language*. Phd, University of Amsterdam, Amsterdam.

van Cranenburgh, A., K. van Dalen-Oskam, and J. J. van Zundert
2019. Vector space explorations of literary language. *Language Resources and Evaluation*, 53(4):625–650.

van Zundert, J. and K. van Dalen-Oskam
2019. Joris en Karina's Holistisch Letterkundig Onderzoeksbureau. *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 135(4):357–372.