I Catching

Computationally Operationalising Narrative Perspective for Stylometric Analysis

Lisanne M. van Rossum[†❖] Joris J. van Zundert[†] Karina van Dalen-Oskam[†♦]

[†]Department of Literary Studies
Huygens Institute for the History of the Netherlands

[❖]Faculty of Humanities
Utrecht University

[♦]Faculty of Humanities
University of Amsterdam

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). In this socio-textual equation narrative perspective is one factor that is fundamental to stylistic analysis, for example in the isolation of free indirect discourse by Annelen Brunner (2019), but to date has proven challenging to examine computationally. Establishing narrative perspective however, is important when we want to answer certain 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 project, we need to be able to establish narrative perspective computationally and reliably. In this paper we present and evaluate a method to do exactly this.

Method & Results

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 more narratologically theoretically informed and based on computing the ratio of combinations of pronouns.[1]

Both methods were tested and validated on 1001 selected full text fragments of Dutch fiction non-dialogue narrative that were labeled 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 and a perfect 1.00 score for the pronoun ratio approach (cf. figures 1 and 2).

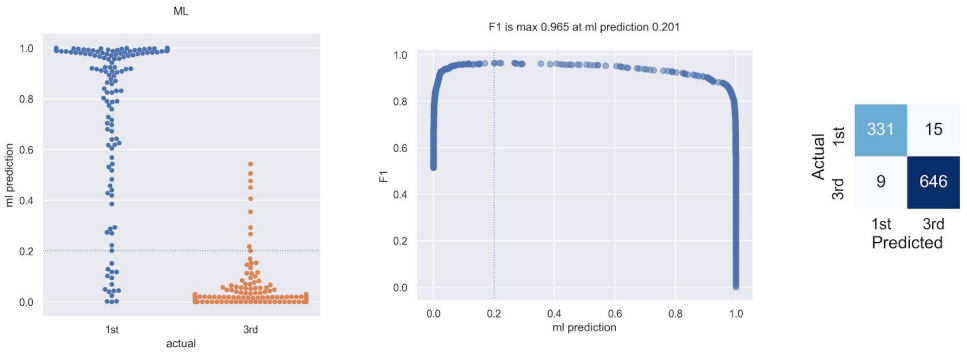

Figure 1: Swarm plot of predictions, F1 curve, and confusion matrix for the machine learning approach.

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

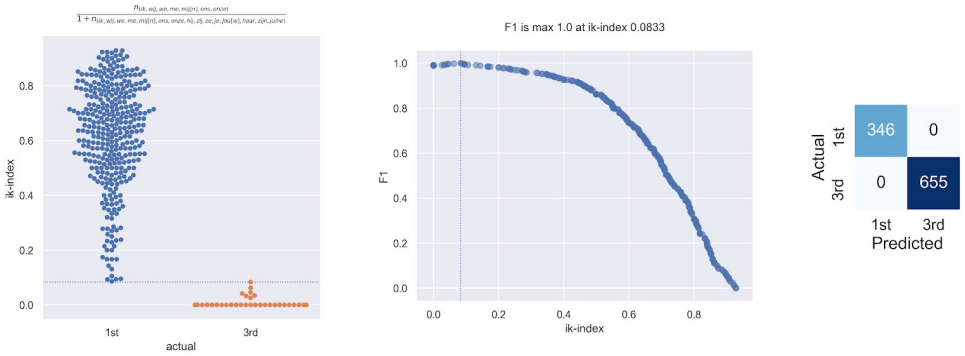

Figure 2: Swarm plot, F1 curve, and confusion matrix for pronoun index approach.

The machine learning approach used a three layer feed forward convolution neural network with a standard Keras Python implementation backed by Tensorflow. The pronoun ratio based approach we started by using equation (1).

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

Equation 1: I-index based on counts of the pronouns "ik" ("I"), "hij" ("he"), and "zij" ("she").

Equation 1 computes the ratio between first person pronoun singular "ik" (transl. "I") and third person pronouns singular "hij" and "zij" (transl. "he", "she"). This equation worked well, but evolving it we found that including more perspective revealing pronouns yielded near perfect to perfect results. The equation eventually derived is depicted as equation 2, or Van Rossum's I-index.[2]

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

Equation 2: I-index based on counts of a fuller set of pronouns including second person, plural, and possessive pronouns.

Discussion

As a first step in developing some baseline approaches to computationally describing narrative perspective present in text, these measures yield excellent and comparable results. However, we prefer the pronoun ratio based approach above the machine learning one

[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".

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 our 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 determination to establish a baseline measure. Obviously perspective in fiction is more complex than the dichotomy between first and third person view (cf. for instance Fludernik, 1995, on unfamiliar pronominal strategies such as second-person pronouns of address and impersonal pronouns in experimental fiction). In anticipation we purposefully designed this computational measure as an index to acknowledge our intention to approach literary perspective as a spectrum, rather than as a binary system. This design also opens up the possibility of the index's use for narratological shifts within the context of the full text of novels.

It must be noted however that our measures have been tested on narrative text in the sense of text without quoted or directly implied dialogue. In the case of indirect free speech, however, it is not readily clear whom the narrative perspective is most closely associated with (either speaker or story teller) and how mixed perspective in such cases could and should be measured.

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 real-world situations there is obviously no way to do this. Further experimenting needs to determine if these cut off values are both realistic and usable in unsupervised situations.

Obviously our pronoun ratio based method only works for languages that have relatively easy and unambiguously determinable pronoun identifiers, such as Dutch, German, and English. The situation for languages with verb or noun inflexion to indicate pronouns might be more difficult.

Future work

We want to expand our evaluation to examine if the found cut off rates for predictions is stable over different compositions of the corpus. This is important to establish a reliably suggested cut off in real world situations. Furthermore, we want to expand our work into the direction of automatically distinguishing dialogue and non-dialogue text to be able to work more concretely on the problem of mixed perspective story text.

References

Brunner, A. (2019) 'Speech, thought and writing representation - Towards automatic detection', *Zeitschrift für Germanistische Linguistik*, 47 (1), pp. 216-248.

Fludernik, M. (1995) 'Pronouns of address and 'odd' third person forms: The mechanics of involvement in fiction' in Green, K. (ed.) *New essays in deixis: Discourse, narrative, literature*. Amsterdam-Atlanta: Rodopi, pp. 99-129.

Koolen, C. (2018) *Reading beyond the female: The relationship between perception of author gender and literary quality.* PhD. University of Amsterdam.

Koolen, C., Van Dalen-Oskam, K., Van Cranenburgh, A., and Nagelhout, E. (2020) 'Literary quality in the eye of the Dutch reader: The National Reader Survey', POETICS, available online 15 February 2020. Available at: doi:10.1016/j.poetic.2020.101439.

Van Cranenburgh, A. (2016) *Rich statistical parsing and literary language.* PhD. University of Amsterdam.