Stylometry and pornographic novels: a replication study on the *Mutzenbacher* case

Simone Rebora*, Massimo Salgaro**

* Università degli Studi di Verona

simone.rebora@univr.it

** Università degli Studi di Verona

massimo.salgaro@univr.it

1. Introduction: genre and stylometry[1]

The issue of genre has been discussed extensively in the field of stylometry. As noted on multiple occasions, any authorship attribution study carried out with methods like Delta distance[2] needs to take into account all the possible biases introduced by similarities of genre. A note of caution is already proposed by two of the "founding fathers" of modern stylometry, John Burrows and Hugh Craig, when commenting on the disappointing results obtained by a stylometric analysis of the *Bussy d'Ambois* by George Chapman and the *Sejanus* by Ben Jonson, which were consistently attributed to the wrong author:

> *Bussy d'Ambois* and *Sejanus*, at all events, are sufficient reminder that the Delta procedure is genre sensitive as well as author sensitive, and that, valuable as they are, its unsupported results should not be taken either as conclusive or as purely authorial.
>
> How can it be, to pause for a moment, that the frequency patterns of the very common words reflect differences of genre? Two examples, each broad in its collateral effects, may serve. Whether in its declamatory or its reflective mode, its self-assertion or its self-scrutiny,

---

[1] Rebora wrote sections 3 and 4, while Salgaro wrote sections 1 and 2.

[2] Burrows 2002.

tragedy characteristically yields much higher frequencies of I/my than are usual in comedy, where brisk dialogic interplay brings thou/you into prominence. By the same token, the long speeches of much tragedy rely on a more complex syntax, with a corresponding increase in the incidence of the major connectives[3].

While providing only circumstantial evidence, their reasoning suggests how the key ingredient for the success of stylometric methods, namely the subconscious preferences by different authors in the usage of function words[4], cannot be entirely protected from the influence of other elements like genre.

Another, even stronger remark is made by Joseph Rudman, who states that "there is a greater stylistic difference between one author in different genres than between two authors writing in the same genre"[5]. Ros Barber confirms how such an intuition is generally shared in authorship attribution, highlighting how "it is a common practice in computational stylistics to choose comparison texts in the same sub-genre category: a history play will be tested against other history plays, a comedy against other comedies"[6].

It is not by chance that extensive projects have been dedicated to the study of genre including also a stylometric perspective. One of the most recent, entitled "Computational Literary Genre Stylistics"[7], reunited a group of literary scholars and computer scientists under the guidance of Christof Schöch, to explore the complex relationship between genre theory and stylistics. In a seminal study on the subject, Schöch tried to explore systematically such a relationship, working on "a collection of 32 plays by Pierre Corneille and Thomas Corneille, with an equal number of comedies and tragedies by each author"[8]. By calculating Burrows's Delta distance with an increasing number of most frequent words (MFW), Schöch showed how the weight of the effect of genre correlated positively with the number of MFW, becoming even more important than authorship after a certain point. His conclusion was that "collections of mixed dramatic sub-genre may be used in authorship classification tasks, provided that the wordlist used does not exceed the first 1,150 MFW, so that influence from features related to genre remains limited"[9]. While confirming the notes of warning by Rudman and Barber, such a result contradicts the intuition by Burrows and Craig, that genre could also have an impact on function words (generally corresponding to the first 100-150 MFW in a language).

Indeed, evidence in stylometry is never conclusive, but the existence of a link between genre and style is by now a shared acquisition. Also, it is commonly recognized that stylometry shows varying

---

[3] Burrows and Craig 2011, p. 36.
[4] Kestemont 2014.
[5] Rudman 2016, p. 318.
[6] Barber 2019, p. 6.
[7] CLiGS, https://cligs.hypotheses.org/sprachen/english (last access: 27 June 2025).
[8] Schöch 2013.
[9] Schöch 2013.

efficiency when working with different authors[10] and – we should add – with different genres. This paper will focus on one specific genre (pornography), which has the potential to impact also the usage of function words, thus invalidating a stylometric analysis if not treated with carefulness. It will be structured as the replication of a previous study, conducted without considering such an issue. As the recent debate on computational literary studies[11] has demonstrated, in fact, replication is one of the most powerful tools available for digital humanists to strengthen the discipline[12]. In particular, the replication of flawed setups and incorrect choices can help better understand what could go wrong when performing authorship attribution with stylometric methods.

2. The case study

2.1 The pornographic genre

In his study entitled *Penetrating language* (2008), Georg Marko proposes a "critical discourse analysis of pornography"[13]. In his work, he compares two corpora: the first is composed by 486 pornographic stories that consist of 680,000 words; the second by 68 erotic stories published in three collections between 1985 and 1995, consisting of 239,959 words. While pornographic stories are published anonymously and often in the form of letters or diaries, erotic stories are mostly written by women who also signed their works. The two genres also differ formally, as the latter also contains character descriptions and consequently includes longer stories.

The contrastive analysis of the two corpora reveals quantitative differences between the two genres, as evident from the list of the MFW in the corpus of pornographic texts[14]. Overall, pronouns seem to play a particularly important role in pornographic novels, with terms like "I", "my", "her", and "she" appearing in the very first positions.

To further distinguish the two genres, Marko proposes four hypotheses, namely that the pornographic corpus is superior in:

- fragmentation (i.e., the tendency to conceptualize humans in terms of their body and body parts);
- physicalization (i.e., the tendency to conceptualize humans in terms of their physical and in particular their visual features);

---

[10] Eder 2017.
[11] Da 2019.
[12] Schöch et al. 2020.
[13] Marko 2008.
[14] Marko 2008, p. 113.

- desubjectification (i.e., the tendency not to conceptualize humans on thinking, feeling and perceiving subjects;
- passivization (i.e., the tendency to conceptualize humans as passive recipients of events rather than as active participants in them)[15].

Especially in the case of fragmentation, pronouns also play a decisive role. The phenomenon is investigated by Marko through the occurrence of anatomical parts, finding how they are often connected with possessives (like "her bosom", "his d**k", etc.)[16]. For this reason, Marko develops his corpus search starting from the usage of such possessive expressions, finally confirming his initial hypothesis[17].

While Marko's study suggests that the pornographic genre could have a direct impact on the usage of MFW – and, consequently, on a stylometric analysis with the Delta method – one case study in authorship attribution offers the ideal context where to verify such intuition.

2.2   Josephine Mutzenbacher

Published in 1906, *Josefine Mutzenbacher oder die Geschichte einer Wienerischen Dirne von ihr selbst erzählt* (Josephine Mutzenbacher or the Story of a Viennese Whore as Told by Herself)[18] holds a distinctive place in German literary history. The novel, presented as the posthumous memoir of the Viennese prostitute Josefine Mutzenbacher (1852-1904), is introduced by an editor who claims to have minimally altered the original manuscript. The narrative, framed by the adult Josefine reflecting on her past without regret, chronicles her explicit sexual encounters from the age of five to fourteen, including incest and interactions with educators, culminating in her first day as a prostitute.

The novel has both been praised by literary critics[19] and touched by economic success[20]. At the same time, its nature of a truthful memoir has been put into question, stimulating a heated debate about its authorship[21]. Several hypotheses have been proposed over the years. Felix Salten (born Siegmund Salzmann), a prominent figure in the Jungwien movement alongside writers like Arthur Schnitzler and worldwide known for being the author of the children's book *Bambi*, is generally considered as the prime candidate. Schnitzler himself was also considered as a potential author, given his own scandalous works like *Reigen*. More recently, additional evidence has emerged pointing to Ernst

---

[15] Marko 2008, pp. 131-132.
[16] Marko 2008, p. 166.
[17] Marko 2008, p. 315.
[18] Mutzenbacher 1906.
[19] Weinzierl 1985, pp. 27-30.
[20] Weinzierl 1985, p. 25.
[21] For an overview, see Rebora and Salgaro 2022, pp. 244-250.

Klein (writer of other pornographic novels under the pseudonyms of Fedor Essée and Richard Werther) as the true author. In a study published in 2022, we collected all these hypotheses and performed a stylometric analysis, confirming Salten as the most probable author, but also identifying a style change in the last pages of the novel, probably written by another, unidentified author[22].

In 2023, Laura Untner and Murray G. Hall published the essay "Josefine Mutzenbacher. Ein pornographischer Roman von Ernst Klein?"[23] to participate in the discussion on the attribution of the novel. Their approach adopted the same stylometric methods already used in our 2022 article but differed in the construction of the corpus. In fact, while in our article we decided to exclude Klein's pornographic novels (in order not to give him any advantage in the confrontation with the other candidates, by whom no pornographic novels are available), Untner and Hall decided to include them. They emphasized that the selected texts were all published (under pseudonyms) around the same year of the *Mutzenbacher* and that they are all prose texts, concluding that such selection was "more specifically oriented towards the research question and the applied method"[24]. The results of their analyses showed strong evidence for Klein being the author of the *Mutzenbacher* novel.

One surprising result was that such evidence was strong both when performing analyses with 2,000 MFW (which would indeed include genre-related words and disregard the already-mentioned caveat by Schöch), but also with 50 MFW (where function words – and therefore subconscious authorial preferences – should play a dominant role). As already noted, however, the peculiar characteristics of the pornographic genre could explain such a result. A more detailed analysis of the texts is therefore needed to reconstruct the effects of such genre on stylometric analyses.

3. Replication of the Untner-Hall analysis

3.1 Analysis setup

All scripts and materials have been made available through a dedicated GitHub repository[25]. Digitizations of Klein's pornographic novels were kindly provided by Laura Untner herself. However, due to copyright limitations, they cannot be openly shared. All stylometric analyses were performed using the *stylo* package[26].

---

[22] Rebora and Salgaro 2022, pp. 259-261.
[23] Untner and Hall 2023.
[24] Untner and Hall 2023, p. 21.
[25] https://github.com/SimoneRebora/Josefine_Mutzenbacher_replication (last access: 27 June 2025).
[26] Eder et al. 2016.

Table 1 summarizes the structure of the corpus, which is composed by three novels each by the three candidates Salten, Schnitzler, and Klein, plus the entire text of the *Mutzenbacher* novel. To emulate the procedure implemented in our 2022 study and partially replicated by Untner and Hall, texts were normalized to solve possible typographic variations and split into 5,000-word chunks (i.e., the minimum dimension for a reliable stylometric analysis)[27]. This setup was the basis for all following analyses.

Author	Title	No. words	No. chunks
Anonymous	*Josefine Mutzenbacher oder die Geschichte einer Wienerischen Dirne von ihr selbst erzählt*	67,286	1
Ernst Klein	*Die Beichte eines Sünders. Memoiren eines Erotomanen*	24,818	4
Ernst Klein	*Das Freudenmädchen. Tagebuch einer Bordellärztin*	34,085	6
Ernst Klein	*Durchlottete Nächte, durchspielte Tage. Der Roman einer Berliner Lebedame*	33,511	6
Felix Salten	*Bambi. Eine Lebensgeschichte aus dem Walde*	39,499	7
Felix Salten	*Herr Wenzel auf Rehberg und sein Knecht Kaspar Dinckel*	15,389	3
Felix Salten	*Der Schrei der Liebe*	88,801	17
Arthur Schnitzler	*Frau Berta Garlan*	48,671	9
Arthur Schnitzler	*Traumnovelle*	25,898	5
Arthur Schnitzler	*Der Weg ins Freie*	118,013	23

Table 1. Corpus composition

3.2 MFW analyses

As the most surprising result obtained by Untner and Hall was the attribution to Klein with 50 MFW, our replication started from there. We selected the same distance measure (Cosine Delta) and performed a series of stylometric analyses with an increasing number of MFW, from two to 50, finally calculating the mean distance between each candidate author and the *Mutzenbacher*. Figure 1 shows how, while Salten is initially the closest candidate, attribution moves slowly but steadily towards Klein, with Schnitzler never reaching a prominent position.

---

[27] Eder 2013. Note that the *Mutzenbacher* novel was not split into chunks, being it the text to be attributed.
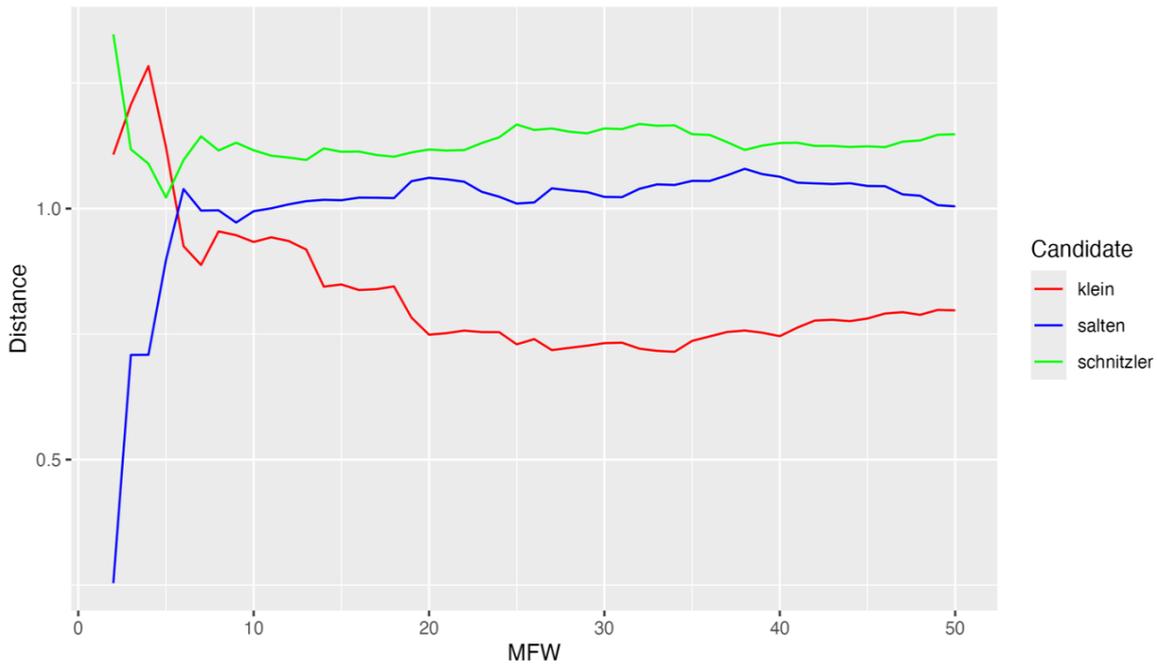
Figure 1. Evolution of distances between the *Mutzenbacher* and the three candidate authors (with Cosine Delta and varying MFW)

While Figure 1 is already informative, suggesting how there might be conditions where the non-pornographic novels by Salten are closer to the *Mutzenbacher* than the pornographic novels by Klein, it is difficult to extract meaningful information from it. In fact, distances are calculated through the cumulative contributions of all MFW in the list. Especially for higher MFW selections, the effect of the addition of a new word could be overshadowed by the effects of all the other words in the list. For this reason, a different procedure was implemented by focusing just on Salten and Klein – clearly the two strongest candidates for the attribution.

First, the mean distances between each author and the *Mutzenbacher* were calculated with 50 MFW and taken as a reference.

Second, the mean distances were re-calculated by removing one word from the list and compared with the reference distances. If the deletion caused a rapprochement of Salten to the *Mutzenbacher*, this meant that the word was favoring Klein in the attribution; if it caused a rapprochement of Klein, then it was favoring Salten. This procedure was repeated for all 50 words in the list, producing the graph in Figure 2, which re-orders the words based on how strongly they favor one or the other candidate in the attribution.
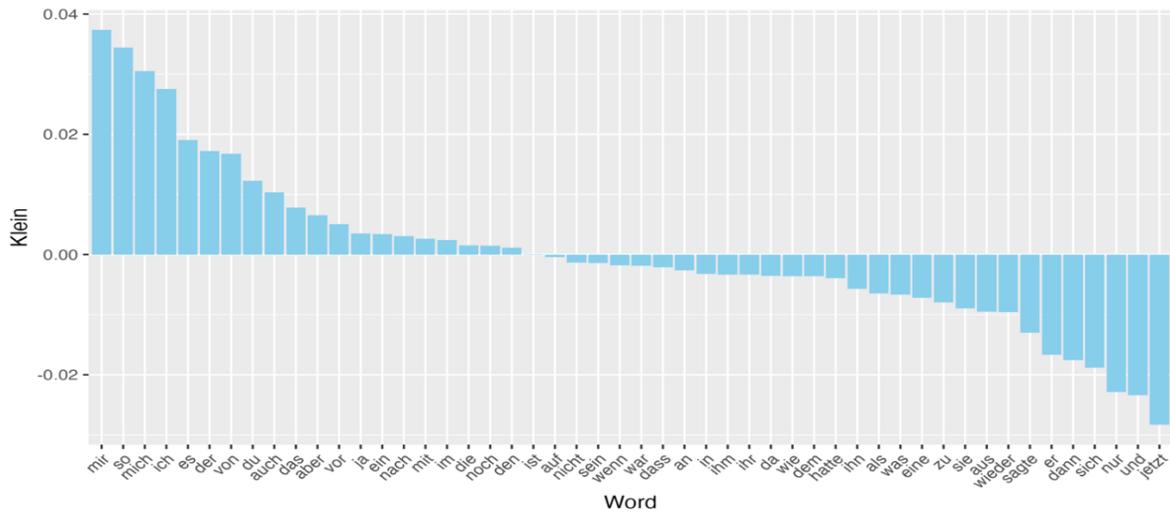
Figure 2. Words favoring Klein vs. words favoring Salten in the *Mutzenbacher* attribution with 50 MFW and Cosine Delta.

Figure 2 seems to confirm the hypothesis derived from Marko's study, as the words that are favoring Klein the most are first-person pronouns (in particular the dative "mir", the accusative "mich", and the nominative "ich"). Note also how the words favoring Salten are even more than the ones favoring Klein (29 vs. 21), but with weaker effects. A comparison of the Zeta scores of the Klein-favorable words in Figure 3 confirms how the strongest effects are caused by an over-representation of first-person pronouns. For example, "mir" is over-represented both in the *Mutzenbacher* (see the first dash corresponding to the term) and in the Klein texts (see the corresponding green box), while it is under-represented in the Salten texts (see the corresponding blue box). The same pattern can be identified for terms like "mich", "ich", and "du" (the second-person pronoun), while an opposite pattern (i.e., under-representation both in Klein and *Mutzenbacher* and over-representation in Salten) emerges for terms like "es" (neutral third-person singular pronoun).
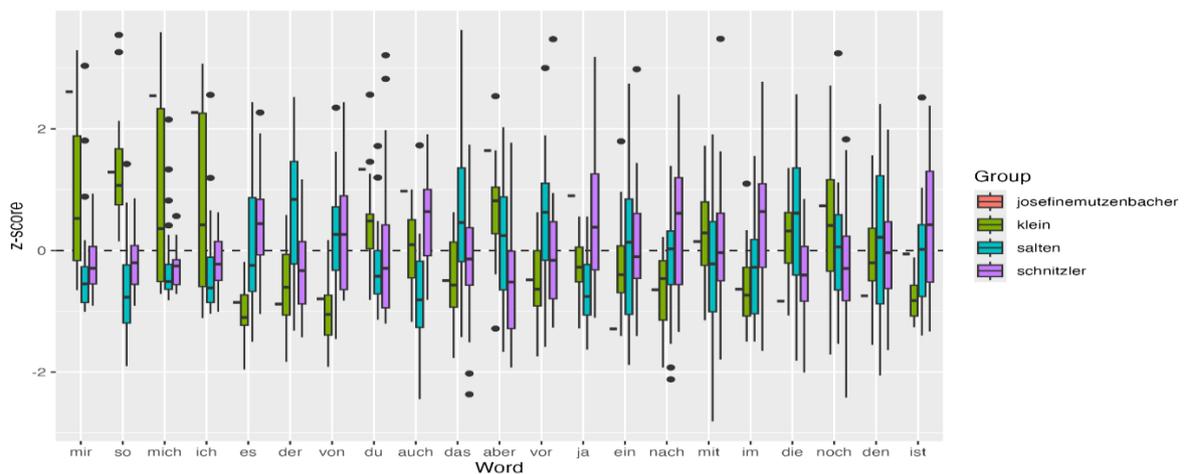

Figure 3. Zeta scores of the 21 MFW favoring Klein

These frequency patterns, while confirming the effect of the pornographic genre on function words, highlight also a deeper issue in the Untner-Hall analysis. In fact, an examination of the pornographic novels by Klein shows how two out of three are narrated in first person (resulting in 10 out of 16 text chunks), while only the shortest novel by Salten adopted this perspective (3 out of 27 chunks). The issue of narrative perspective has already been taken into consideration in stylometric research. As highlighted by Artjoms Šeļa: «The 'narrative fiction' is often considered to be a uniform register (at least silently) in attribution and evaluation practice […], but there is an obvious linguistic variation within fiction, too, for example as an effect of narrative perspective»[28]. Indeed, different narrative perspectives can skew the results of a stylometric analysis even more than genre, as they directly impact the usage of some of the most widely used words in language – that is, pronouns. When this phenomenon overlaps with the linguistic tendencies already present in a genre, the distortion of the authorial signal can become overwhelming. Other phenomena evidenced by Figures 2 and 3 could also indicate the genre influence (such as the over-representation of the intensifier "so" and of the affirmative "ja", which could be easily connected to the tendencies of the pornographic genre), however, it is first-person pronouns that drive the attribution most strongly. When the analysis finally moves to low-frequency words, the effect of genre becomes undeniable. Figure 4 replicates the procedure of Figure 2 by taking into account 2,000 MFW. For reasons of decency, we will not translate the words favoring Klein in the attribution, but the repeated references to body parts, induments, and sexually-connotated places and actions make it evident how they primarily connect to the pornographic genre.
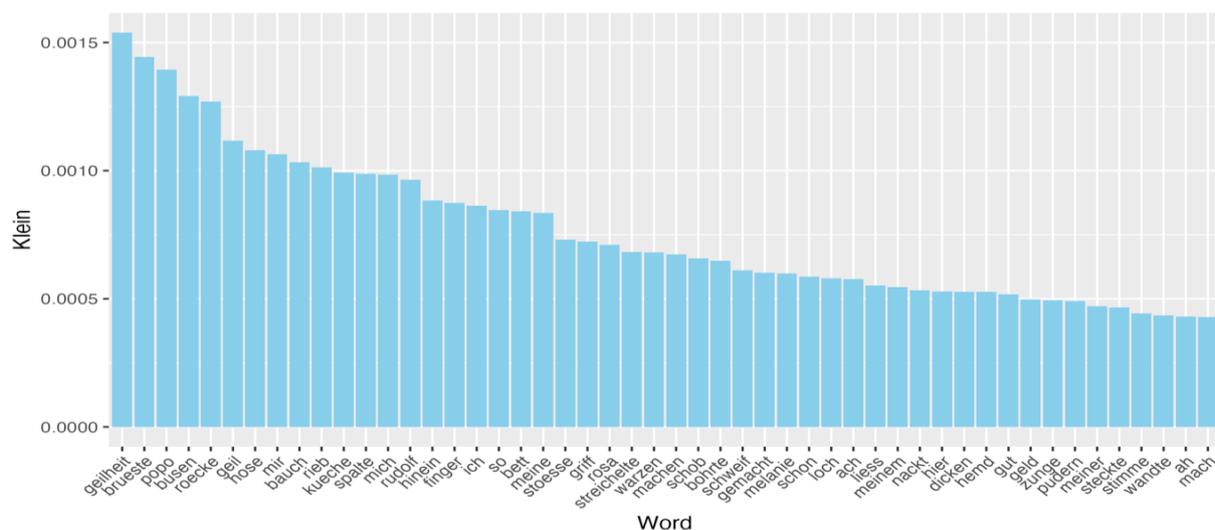

Figure 4. Words favoring Klein in the *Mutzenbacher* attribution with 50 MFW and Cosine Delta

---

[28] Šeļa 2023.

3.3   The effect of pronouns

To better understand the effect of pronouns on the stylometric analysis, a small experiment was devised that intervened on the texts by excluding all personal pronouns. The goal was that of empirically removing any direct indicator of narrative perspective by keeping the analysis as objective as possible. Note in fact how, by removing all the 12 personal pronouns, not only five words favoring Klein were disregarded, but also seven terms favoring Salten.

To strengthen the scientificity of the approach, significance tests were performed on the two series of measurements resulting from the inclusion and exclusion of pronouns, to check if the distances measured between Klein's texts and the *Mutzenbacher* were significantly lower than those between Salten's texts and the *Mutzenbacher*. As Figure 5 shows, the difference is significant in the original setup by Untner and Hall, but it becomes non-significant when removing the 12 pronouns. Results are not reversed, but the stylometric analysis becomes inconclusive, confirming how the initial attribution was mainly driven by narrative perspective.
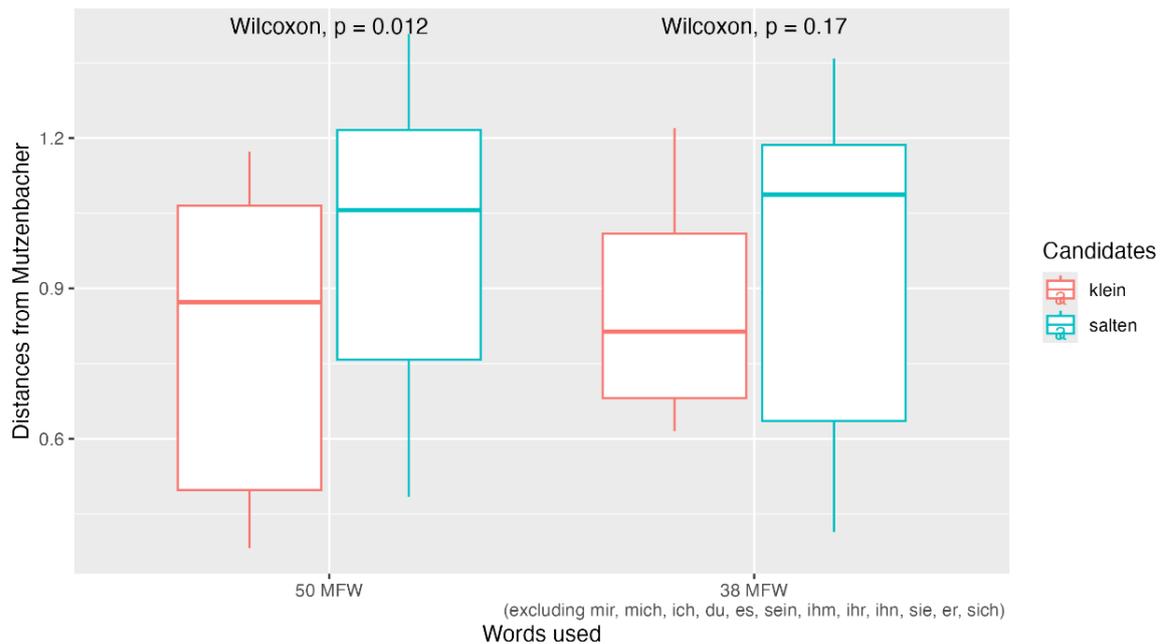

Figure 5. Comparison of stylometric analyses including and excluding pronouns

A similar result is also produced when removing first-person narratives in their entirety. Figure 6 shows the outcomes of an analysis performed without modifying the texts, but by simply limiting it to the one by Klein and the two by Salten that were not narrated in first person. Also in this case, the stylometric analysis becomes inconclusive.
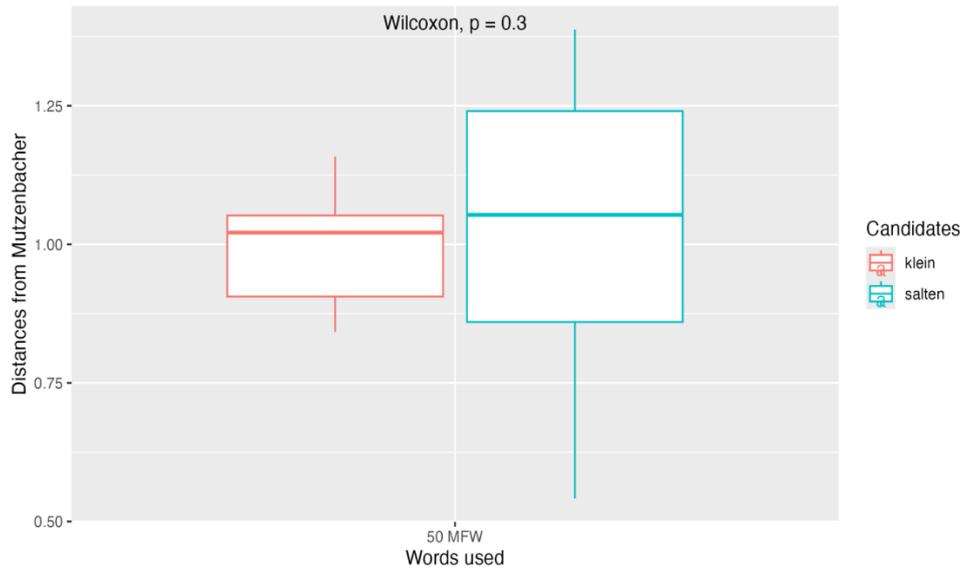
Figure 6. Results of stylometric analysis excluding first-person narratives

4. Conclusion

This study has demonstrated how the replication of previous research can allow a better understanding of fundamental issues in a discipline, such as the impact of genre and narrative perspective on stylometric analyses. At the same time, it has also confirmed the unexhausted need for guidelines and the risks that derive from adopting these very guidelines acritically. The choices made by Untner and Hall were in fact based on the available literature on the subject and, for example, they decided to give high relevance to the results produced with 2,000 MFW based on the indications provided by a fundamental methodological study in stylometry[29]. However, by simply following a general indication without interpreting it critically, they overlooked another, more fundamental issue related to the very nature of the corpus.

The results presented here showed how the Klein hypothesis is not tenable in the *Mutzenbacher* case, especially when using his pornographic novels as an object of comparison. Under the current circumstances, the Salten hypothesis remains the most probable one.

In any event, we do not claim that our research has definitively resolved the issue. Further research can and should be conducted, specifically on the impact of genre in stylometric research. One possibility would imply developing a "stylometric profile" of one genre, which would include the word frequency patterns that distinguish it. Having defined such a profile would allow "filtering it out" from the analysis, by simply subtracting the frequency scores of the profile from those of the texts under

---

[29] Evert et al. 2017. Note how studies like the one by Evert et al. usually tested their stylometric methods on artificially constructed (and genre-controlled) corpora, producing results which could be very informative in themselves, but cannot always be generalized.

examination. Such a procedure would of course imply – in the *Mutzenbacher* case, as well as in any other similar case – developing a wide and representative corpus for the genre. A complex and demanding task, which is out of the scope of the current study, but which could prove extremely useful for stylometry research in general.

Overall, we hope that our study has not just rejected one hypothesis but also opened new possibilities for further research and analysis, in the knowledge that stylometry can never provide conclusive evidence in a case of authorship attribution but just do its best to reach the highest probability for a possible answer.

Acknowledgements

References

Barber 2019
Ros Barber, *Marlowe and Overreaching: A Misuse of Stylometry*, «Digital Scholarship in the Humanities», 34.1 (2019), pp. 1–12. doi: 10.1093/llc/fqy040.

Burrows 2002
John Burrows, *`Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship*, «Literary and Linguistic Computing», 17.3 (2002), pp. 267–87, doi: 10.1093/llc/17.3.267.

Craig and Burrows 2011
Hugh Craig and John Burrows, *A Collaboration about a Collaboration: The Authorship of King Henry VI, Part Three*, in *Collaborative Research in the Digital Humanities*, edited by Marilyn Deegan and Willard McCarty, Farnham, Surrey, England ; Burlington, VT: Ashgate Pub, 2011, pp. 27–65.

Da 2019
Nan Z. Da, *The Computational Case against Computational Literary Studies*, «Critical Inquiry», 45.3 (2019), pp. 601–39, doi: 10.1086/702594.

Eder 2013
Maciej Eder, *Does Size Matter? Authorship Attribution, Small Samples, Big Problem*, «Digital Scholarship in the Humanities», 30.2 (2013), pp. 167–82, doi: 10.1093/llc/fqt066.

Eder 2017
Maciej Eder, *Short Samples in Authorship Attribution: A New Approach*, in *DH2017 Book of Abstracts*, 2017, pp. 221–24.

Eder et al. 2016

Maciej Eder – Jan Rybicki – Mike Kestemont, *Stylometry with R: A Package for Computational Text Analysis*. «The R Journal», 8.1 (2016), pp. 107–121, doi: 10.32614/RJ-2016-007.

Evert et al. 2017

Stefan Evert – Thomas Proisl – Fotis Jannidis – Isabella Reger – Steffen Pielström – Christof Schöch – Thorsten Vitt, *Understanding and Explaining Delta Measures for Authorship Attribution*, «Digital Scholarship in the Humanities», 32.suppl_2 (2017), pp. ii4–16, doi: 10.1093/llc/fqx023.

Kestemont 2014

Mike Kestemont, *Function Words in Authorship Attribution. From Black Magic to Theory?*, in *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, Gothenburg, Sweden: Association for Computational Linguistics, 2014, pp. 59–66, online: http://aclweb.org/anthology/W/W14/W14-0908.pdf.

Marko 2008

Georg Marko, *Penetrating Language: A Critical Discourse Analysis of Pornography*, Tübingen: Gunter Narr Verlag, 2008.

Mutzenbacher 1906

Josefine Mutzenbacher, *Josefine Mutzenbacher oder Die Geschichte einer Wienerischen Dirne von ihr selbst erzählt*, Vienna: Privatdruck, 1906.

Rebora and Salgaro 2022

Simone Rebora and Massimo Salgaro. *Is Felix Salten the Author of the Mutzenbacher Novel (1906)? Yes and No*, «Language and Literature: International Journal of Stylistics», 31.2 (2022), pp. 243–64, doi: 10.1177/09639470221090384.

Rudman 2016

Joseph Rudman, *Non-Traditional Authorship Attribution Studies of William Shakespeare's Canon: Some Caveats*, «Journal of Early Modern Studies», 5 (2016), pp. 307–28, doi: 10.13128/JEMS-2279-7149-18094.

Schöch 2013

Christof Schöch, *Fine-Tuning Stylometric Tools: Investigating Authorship and Genre in French Classical Theater*, in *DH2013 Book of Abstracts*, University of Nebraska–Lincoln: ADHO, 2013, online: http://dh2013.unl.edu/schedule-and-events/program/.

Schöch et al. 2020

Christof Schöch – Karina van Dalen-Oskam – Maria Antoniak – Fotis Jannidis – David Mimno, *Replication and Computational Literary Studies*, in *DH2020 Book of Abstracts*, Ottawa, Canada: Zenodo, 2020, pp. 509–14, doi: 10.5281/ZENODO.3893427.

Šeļa 2023

Artjoms Šeļa, *Corpus Building for Authorship Attribution, in Survey of Methods in Computational Literary Studies* (= D3.2: Series of Five Short Survey Papers on Methodological Issues), edited by Christof Schöch, Julia Dudar, and Evgeniia Fileva, CLS INFRA, doi: 10.5281/zenodo.7892112.

Untner and Hall 2023

Laura Untner and Murray G. Hall, *Josefine Mutzenbacher. Ein pornographischer Roman von Ernst Klein?*, «Jahrbuch Für Internationale Germanistik», 15.1 (2023), pp. 11–28.

Weinzierl 1985
Ulrich Weinzierl, *Die Wahrheit ist nackt. Über Josefine Mutzenbachers Lebensgeschichte*, 1985, in: *Josefine Mutzenbacher Oder Die Geschichte Einer Wienerischen Dirne von Ihr Selbst Erzählt*. Ungekürzter Nachdr. der Erstausg. aus dem Jahr 1906, edited by Michael Farin, Munich: Schneekluth, 1990, pp. 25–33.