Author gender and text characteristics in contemporary Swedish fiction

Mats Dahllöf
Uppsala University, Sweden

Abstract
The present study addresses the question of to what extent and how authors’ gender is reflected in the textual properties of bestselling fiction in Swedish during the period 2015–2020. The empirical material was a corpus of 235 female-authored books and 214 male-authored works. The analysis of the texts departed from text property measures targeting grammatical and lexical aspects of language use. Differences between the genders were analysed using the probability of superiority measure in combination with a threshold criterion. The results suggest that authors of bestselling fiction in the Swedish book market to a high degree engage in forms of gender performance when they compose their texts. The differences could in most cases be interpreted as conforming to patterns that have previously been reported for other languages and categories of language use. The gender performance to a large extent agreed with traditional stereotypes about the interests of women and men. There were also differences in grammar-related stylistic preferences. Among the female themes, positive emotion and social interaction were prominent. The male examples include weapons and animosity, as well as numerical quantification. A more grammar-related tendency is that male authors tend to package a larger fraction of their text into noun phrases.

Keywords
Bestseller, corpus linguistics, distant reading, fiction, gender, style, stylometry

Corresponding author:
Mats Dahllöf, Department of Linguistics and Philology, Uppsala University, P.O. Box 635, 751 26 Uppsala, Sweden.
Email: mats.dahllof@lingfil.uu.se
I. Introduction

‘In order to succeed [y]ou have to write like a man!’, a female writer of crime fiction claimed, when she was interviewed by Alacovska (2017: 389). Archer and Jockers (2016: 137) drew an almost opposite conclusion: “when it came to mastering the style that is most typical to the bestseller list, women were the clear leaders.” Observations like these share the assumption that there are differences between how women write and how men write (even if people might sometimes choose to write like the opposite gender). The present study is about these differences. This is one among several related questions about how gender is performed in the production, consumption, and critical reception of literature.

1.1. Female and male authors in popular fiction

The differences between female and male language use have stimulated considerable interest in various academic disciplines. When it comes to scholarship on literature and gender, political, theoretical, and empirical modes of analysis have often gone hand in hand. The methodological approach taken here is text-oriented and relies on the quantitative tools of corpus linguistics.

The present study concerns recent commercially successful fiction in the Swedish book market, including works in both printed and digital form. The latter are of two kinds, e-books (electronic text) and audiobooks (recordings by narrators), both of which are mainly distributed through online platforms. By focusing on recent popular fiction the current study gauges a high-impact section of literature and an important branch of contemporary Swedish culture.

Historically, literature and literary criticism used to be areas completely dominated by male authors and scholars. The emergence of feminist literary criticism some 50 years ago led to an increased interest in female authors and the significance of gender. It is a common view that this scholarship was driven by activism: ‘[F]eminist literary studies and feminist literary theories are […] the product of feminist politics’, as Rooney (2006: 77) puts it.

For the ‘first generation of feminist literary critics’ it was a ‘foundational premise of feminist literary theory that the gender differences appearing in fiction arose out of biological differences, a division of labor that accommodates those differences, or a bias that limits the possibilities of one gender for the benefit of the other’ (Armstrong, 2006: 103). There are also scholars who reject this kind of ‘essentialism’. Judith Butler, for instance, holds that ‘gender is performative’ (Butler, 1999: xv) and that ‘the very notions of an essential sex and a true or abiding masculinity or femininity are […] part of the strategy that conceals gender’s performative character’ (Butler, 1999: 180).

Some researchers have expressed an ethical wariness when discussing quantitative research into female-male differences in language use. Bamman et al. (2014: 153), connecting their reasoning to Butler, argue that ‘[c]omputational and quantitative models have often treated gender as a stable binary opposition, and in so doing, have perpetuated a discourse that treasures differences over similarities, and reinforces the ideology of the status quo’. Koolen (2018: 129) makes a similar claim: ‘The ethical problem with this type of research […] is that stressing difference […] comes with the tendency of enlarging the
perceived gap between female and male authors; especially when results are interpreted using gender stereotypes.’ She adds, however (on the same page): ‘The reason we do not propose to abandon gender analysis in [Natural Language Processing] altogether is that female-male differences are quite striking when it comes to cultural production.’

Whereas ‘linguists tend to view sex differences in language use as caused by social stereotypes, researcher bias […], or gender roles’ (Luoto, 2021: 2), and thereby regard gender from the ‘nurture’ position, there are also scholars, including, as we saw above, ‘the first generation of feminist literary critics’ (Armstrong, 2006: 103), who approach gender and language from the ‘nature’ position. One source of inspiration for research aiming at biology-based models is that ‘[a]dvances in cognitive neuroscience and evolutionary science have increased our knowledge of mammalian sexual differentiation of the brain and how this process creates sex differences […] in various psychobehavioural traits in humans […], but the way in which such differences may be reflected in language use is not well known’ (Luoto, 2021: 1). A study by Mascaro et al. (2018: 1) provided ‘preliminary evidence’ that higher testosterone levels in men ‘may influence social behavior by increasing frequency of words related to aggression, sexuality, and status, and that it may alter the quality of interactions with an intimate partner by amplifying emotions via swearing’. A related line of thought is that differences in language use between the genders correspond to underlying features of their psychology and cognitive styles: ‘Differences in the use of function words reflect differences in the ways that individuals think about and relate to the world. […] An examination of gender differences in function word use might shed new light on the psychology of men and women’ (Newman et al., 2008: 216). Argamon et al. (2007) proposes that ‘a single underlying distinction between inner- and outer-oriented communication […] may explain both gender-linked and age-linked variation in language use’.

A new ‘paradigm’ of research into the differences between female and male language use was established when the problem came to be addressed by the methods of large-scale corpus linguistics and machine learning two decades ago. Koppel et al. (2003), the classical reference in this context, approached author gender prediction as a challenging problem in the creation of text classification models and used such models to analyse gender differences. Work in this vein is not only relevant as an academic exercise. Machine learning models able to predict various facts about individuals based on what they write and read also lie at the heart of the commercially important technologies of collecting ‘user profile information’ (Zuboff, 2019: 74–82).

1.2. Research questions

The research questions behind the present study can be stated as follows:

- The primary question is to what extent the gender of authors is reflected in text properties when it comes to contemporary bestselling fiction in Swedish.
- If we find gender-related differences in language use, which are these? To what extent does contemporary bestselling fiction in Swedish exhibit the same kinds of gendered differentiation that have been reported for other languages and for other categories of language use?
Finally, I will explore the wider implications of my findings concerning the differences between female and male writing. What are the implications of the gendered aspects of contemporary bestselling fiction in Swedish for readers of literature and for the fiction market?

The empirical approach scrutinizes a corpus of several hundred works by analysing how thousands of text property measures correlate with author gender. To use the words of Allison (2018: 16), discussing a methodology of this kind, our study ‘emphasis[es] specific kinds of evidence [by] subordinating textual complexity to the point one wants to make about it’. In doing so, we ‘focus on units that are much smaller or much larger than the text’ (Moretti, 2000: 57). I will not claim – and will not be able to claim – that gender-related correlations align with particular kinds of causal mechanisms. As Newman et al. (2008: 233) point out: ‘[I]t is important to note that our analyses merely identify how men and women communicate differently, without addressing the issue of why these differences exist. Gender differences in language use likely reflect a complex combination of social goals, situational demands, and socialization – just to name a few – but these data do not identify these origins.’

2. Previous studies

If we approach gender as a factor that plays a role in the contemporary literature market and writing practices, as I do in the present article, we can note that the examination of text properties is one of several possible entry points. Other important factors, which are likely to interact with the more closely text-related ones, are to what degree female and male readers have different reading habits, how the gender of characters in narratives is connected to how they act and are portrayed, how the gender of authors influences peoples’ opinions about their books, and how the gendered personas of authors play a role in book marketing.

2.1. Aspects of literature and gender

The differences between female and male literature consumption habits are a fundamental aspect of how gender plays a role in the wider context of the literature market. Dutch data suggest that ‘male readers barely read […] female authors […], while] female readers read both genders’ (Koolen, 2018: 140). A study by Summers (2013) reported results in a largely similar vein: The male reader group strongly prefers male authors. Female readers have similar preferences, but to a lesser degree. The vast majority of the male cohort also prefer books with male protagonists. It is most likely that publishers and streaming services take such gender-related preference patterns into account when marketing literary works. Berglund (2015: 30), having studied recent Swedish crime fiction, concluded that ‘that male and female authors of crime fiction in general have been marketed recognizably different in almost all possible ways’. It also seems that knowledge about an author’s gender influences how critics assess the value of literary texts (Alacovska, 2017; Koolen, 2018: 96; Touileb et al., 2020).
Another gender-related effect on reading habits is that women seem to read considerably more fiction than men do. This ‘gender gap’ has been known for almost a century, at least from an American perspective (Tepper, 2000). Recent Swedish data (Nordicom, 2022: 80–81) also show this gap: Girls clearly read more than boys in childhood (9–14 years of age), and the difference widens for the older cohorts in the survey, see Figure 1.

Another gender-related issue, often highlighted in feminist literary criticism, is how female and male characters are depicted in fictional narratives. Alacovska, discussing the professional identities and conditions under which authors work, noted that ‘[c]rime fiction has been traditionally considered an androcentric – masculine and conservative – genre featuring a male detective’ (Alacovska, 2017: 385) in a way that disadvantages female authors.

A study conducted within the present project (Dahllöf and Berglund, 2019) showed, using pronoun evidence and the statistical tool of topic modelling (cf. Section 2.2), that ‘Appearance’, ‘Family circle’, and ‘Eating and drinking’ were topics more strongly connected to female characters in modern bestselling fiction in Swedish. Male characters, by contrast, gravitated towards ‘Fighting, violence’, ‘Police work’, and ‘War and nations’. The same kind of procedure was also applied to a comparable, but older and more highbrow corpus, ‘Swedish classics (1821–1941)’. In that case ‘Dance, music’, ‘Family relations’, and ‘Mental states’ came out as the most ‘female’ topics, while ‘Money, work’, ‘God, religion’, and ‘Authority’ were themes more strongly associated with male characters.

![Figure 1. Swedish data (Nordicom, 2022: 80–81) on the fraction of individuals in different gender and age groups who on any day read some fiction (Swedish 'skönlitteratur') in 2021. (The plot was created for the present article.)](image)
2.2. Previous studies on differences between female and male language use

A quantitative comparison between two categories of text, such as male and female, has to be based on measures which quantify various text properties. Such measures characterize a text in roughly the same way as, e.g., age and height correspond to two numbers describing a person. A common and simple approach is to compute the frequencies of graphic words, i.e., words defined as letter sequences, relative to the length of each text. This requires no language-specific analysis. By using language-specific tools for analysing grammar, measures involving morphological and syntactic categories can be extracted. Johannsen et al. (2015: 104), for instance, extract features based on part of speech tags and dependency subtrees ‘to obtain comparable syntactic analyses across several different languages and demographics’. Their analysis was based on the Universal Dependencies (UD) framework, whose aim is to achieve cross-linguistically consistent annotation of grammatical categories and functions. (UD was also used in the present study).

Measures can also be based on lexical semantic categories. Many studies on gender differences in language (e.g. Newman et al., 2008; Schler et al., 2006; Schwartz et al., 2013), including in fiction (e.g. Koolen, 2018: 96; Luoto 2021), and thousands in other fields, have used the text analysis software Linguistic Inquiry and Word Count (LIWC) to ‘create to capture people’s social and psychological states’ (LIWC, n.d). LIWC is consequently the most used ‘package’ of manually compiled dictionaries. In particular, it involves dictionaries ‘of words, word stems, emoticons, and other specific verbal constructions that have been identified to reflect a psychological category of interest.’ (LIWC, n.d).

An alternative to employing manually created dictionaries for lexical semantic analysis is to use statistical models to discern content themes (constituted by concepts related as witnessed by regular co-occurrence). Argamon et al. (2007), for instance, applied ‘automated factor analysis […] to find groups of related words that tend to occur in similar documents’. One of the factors found (labelled ‘school’) involved the words school, teacher, class, study, test, finish, english, students, period, paper, and pass. Schwartz et al. (2013) used a related statistical approach, so-called topic modelling, to reveal conceptual categories in the form of ‘cluster[s] of semantically related words’ (Schwartz et al., 2013: 9). Topic modelling has also been used in studies of female-male differences by Park et al. (2016) and by Koolen (2018).

Several researchers, such as Argamon et al. (2009), Koppel et al. (2003), and Dahllöf (2012) have studied the engineering of text classifiers for predicting whether texts are female- or male-authored. This is a method of investigating how gender relates to various quantifiable aspects of language use. Archer and Jockers (2016) used classifiers to measure how female/male the styles of various books are. However, they argued that such results must be interpreted with caution, as there might be other factors than gender at work.

Research on the differences between female and male language use have departed from data concerning a variety of contexts and genres. As can be expected, English is by far the most studied language. Several projects have compiled multi-genre corpora: Koppel et al. (2003) studied a genre-controlled subset of the British National Corpus. Sabin et al. (2008)
created a multi-genre dataset of ‘correlated corpora of writing and speech samples from a single population of subjects [with] opinion-related […] topics prescribed and moderated’. Newman et al. (2008) collected data from a large number of research projects in psychology, both original writing and transcribed speech, along with samples of fiction. One of their observations is that ‘[t]aken together, the general pattern of variation across contexts suggests that gender differences are larger on tasks that place fewer constraints on language use’ (Newman et al., 2008: 229).

Other studies on gender differences have departed from more specialized corpora representing discourse from specific kinds of context, such as telephone conversations (Boulis and Ostendorf, 2005), blog entries (Argamon et al., 2007), movie reviews (Otterbacher, 2013), Facebook posts (Park et al., 2016; Schwartz et al., 2013), Twitter posts (Bamman et al., 2014), Swedish parliament speeches (Dahllöf, 2012), research articles (Thelwall et al., 2019), and crowdsourced texts about COVID-19 experiences (Van der Vegt and Kleinberg, 2020). Johannsen et al. (2015) investigated cross-lingual syntactic variation over age and gender drawing their data from a corpus of user reviews from 24 countries, covering 13 different European languages. They found several grammatical features which are indicative of either gender for more than half of the languages. Scholars have also investigated gendered language use in fiction-only corpora in several languages, e.g., French (Argamon et al., 2009), English (Luoto, 2021), and Dutch (Koolen, 2018).

Some studies explored the ways in which women and men exhibit differences in different ways in different genres and circumstances. Sabin et al. (2008) compared female and male subjects from whom e-mail, essay, interview, blog, chat, and discussion data had been elicited. They saw that gender differences actually seem to depend on interactions between genre and gender. Schler et al. (2006) analysed a corpus of entries from female and male bloggers of different age groups. They found that ‘[r]egardless of gender, writing style grows increasingly ‘male’ with age [age brackets: 13–17, 23–27, and 33–42 years]: pronouns and assent/negation become scarcer, while prepositions and determiners become more frequent.’ Argamon et al. (2007) made a similar observation: ‘with few exceptions, the factors and parts-of-speech that are used significantly more by younger (older) bloggers are also used significantly more by female (male) bloggers’. Dahllöf (2012) found, using a classifier-based approach, that gender differentiation was more pronounced for the older and for the right-wing members of the Swedish parliament, than for the younger and for the left-wing MPs. The gender of addressees also seem to matter: Bamman et al. (2014: 136) found, studying Twitter exchanges, that there is ‘a significant correlation between the use of mainstream gendered language – as captured by classifier confidence – and the extent to which an individual’s social network is made up of same-gender individuals’. Sexual orientation has been suggested as a potentially relevant parameter. A study comparing English-language novels by heterosexual and homosexual authors, found that homosexual male authors produced ‘female-typical psycholinguistic outputs[, i.e. that] they use language in a way that resembles heterosexual women’s language use’ (Luoto, 2021: 10).

Research on the differences between female and male language use has reached a number of conclusions which have repeatedly been replicated. An overview is given
below. We should note that there is considerable variation in the details of how the studies have been conducted.

2.3. Features observed to be more frequent in texts with female authors

An often reported difference between female and male language use is the higher fraction of (personal) pronouns in the former. This has been seen several times for English data (Argamon et al., 2007; Bamman et al., 2014; Luoto, 2021; Newman et al., 2008; Otterbacher, 2013; Schler et al., 2006; Schwartz et al., 2013; Van der Vegt and Kleinberg, 2020). This tendency has also attested for 11 out of 13 European languages (Johannsen, et al., 2015) and for literature in French (Argamon et al., 2009) and in Dutch (Koolen, 2018). In particular, this seems to be true for first person singular pronoun forms (Newman et al., 2008; Otterbacher, 2013; Sabin et al., 2008; Schwartz et al., 2013; Van der Vegt and Kleinberg, 2020) and third person singular feminine (forms of she) (Otterbacher, 2013).

The graphic word elle (she) was the strongest single feature for predicting female authorship in French literature according to the classifier-based study of Argamon et al. (2009).

Many studies have also seen that female-authored text tends to contain a higher fraction of verbs than ‘male’ text (Van der Vegt and Kleinberg, 2020). Johannsen et al. (2015) found support for this in user reviews for ten languages out of the 13 they examined. Newman et al. (2008) saw that this held for verb forms in both present and past tense, but not significantly for future tense.

Intensive adverbs, e.g., very and really, is another category of words that women use more frequently according to several studies (e.g. Newman et al., 2008; Park et al., 2016). Johannsen et al. (2015) saw this tendency in five languages.

Among conceptual themes which more frequently surface in female discourse, ‘family’ seems to be one that has emerged strongly in many studies, e.g., in the work by Schler et al. (2006) and Newman et al. (2008). Both studies relied on the LIWC dictionary in which the ‘Family’ category was given beforehand as a subcategory of ‘Social processes’. Argamon et al. (2007) used factor analysis to analyse gender differences finding that a factor given the ‘suggestive heading […] Family’ had a higher mean frequency in female-authored than in male-authored blog posts. When Park et al. (2016) applied topic modelling to Facebook messages, three of the ‘[t]op 20 female-linked language topics’ contained the word family. Results ranking the family theme as more common in female-authored texts have also been reported by Zhang et al. (2011), Schwartz et al. (2013), and Van der Vegt and Kleinberg (2020). Several studies (Argamon et al., 2009; Boulis and Ostendorf, 2005; Dahllöf, 2012; Sabin et al., 2008) found that words relating to the concept of child have been particularly prominent in female discourse. Thelwall et al. (2019) saw that ‘Mothers’, ‘Babies’, and ‘Children’ were the three top-ranking themes as regards being more common in female-authored research articles. The conspicuousness of the family theme in female-authored works has also been attested for fiction in French (Argamon et al., 2009) and in Dutch (Koolen, 2018). Sabin et al. (2008), however, report an opposite tendency for male blogging. The category of ‘Friends’ (LIWC) is also one to which female authors refer more often than male authors do. This has been noted for English (Newman et al., 2008; Schler et al., 2006; Schwartz et al., 2013),
and for Dutch literature (Koolen, 2018). In Luoto’s (2021) study of novels in English, the largest ‘sex difference effect size’ in favour of female authors was observed for the ‘Social’ (LIWC) category.

A factor labelled ‘AtHome’ was extracted as more common in female language use by Argamon et al. (2007). Similar tendencies have been noted using LIWC (Newman et al., 2008; Schwartz et al., 2013), including for Dutch fiction (Koolen, 2018). An opposite tendency for blogs was observed by Sabin et al. (2008).

According to Park et al. (2016: 7) ‘the most strongly female-linked topics included words describing positive emotions’. Other studies have seen the same tendency using LIWC, for emotions generally (Newman et al., 2008), for both positive and negative emotions (Bamman et al., 2014; Schler et al., 2006), for positive emotions only (Schwartz et al., 2013), or for positive emotions more strongly (Luoto, 2021). An emotional inclination in female authors has also been attested for French (Argamon et al., 2009) and Dutch literature (Koolen, 2018), in the latter case for positive emotions.

Another theme that has been associated with female authors is food and eating. This has been noted for blogs (Schler et al., 2006). The factor analysis of Argamon et al. (2007) found a component which they labelled ‘Food/Clothes’ which was present to a higher degree in female-authored blogs. Thelwall et al. (2019) saw that ‘Food health’ was one of the themes more common in female-authored research articles. The category Body (LIWC) scored highly for female writing in Dutch fiction (Koolen, 2018), but the opposite tendency has been observed for French literature (Argamon et al., 2009), and for English (Schwartz et al., 2013).

2.4. Features observed to be more frequent in texts with male authors

Many studies have found that determiners (articles)⁵ are more common in male language use (Argamon et al., 2007; Luoto, 2021; Mehl and Pennebaker, 2003; Newman et al., 2008; Otterbacher, 2013; Schler et al., 2006; Schwartz et al., 2013; Van der Vegt and Kleinberg, 2020). This has also been attested for literature in French (definite feminine la being an exception) (Argamon et al., 2009) and in Dutch (Koolen, 2018). Determiners (by definition) syntactically belong to nouns, but measures based on nouns are better for cross-lingual comparisons, as the grammar of determiner use varies among languages. Accordingly, Johannsen et al. (2015) found that male language use is generally denser in nouns. This difference was significant for 11 out of the 13 European languages their data covered. Newman et al. (2008: 229) summarize and interpret their observations saying that ‘when given the freedom to talk about any topic, men (but not women) elected to talk about concrete objects, which require nouns and, of course, articles’.

It has also often been seen that numerals, another noun-associated category, are used significantly more densely in male-authored texts (Bamman et al., 2014; Luoto, 2021; Newman et al., 2008; Sabin et al., 2008; Schwartz et al., 2013). This has also been attested for literature in French (Argamon et al., 2009) and Dutch (Koolen, 2018), as well as for online user-reviews in 11 out of 13 European languages (Johannsen et al., 2015). Thelwall et al. (2019) found a thematic correlate of this tendency in their study on research articles: ‘Pure math’, ‘Measurement’, and ‘Computing’ were among the top-ranking themes for male-authored works.
Preposition is another part of speech which occurs more frequently when a male author is responsible for the text according to several studies (Argamon et al., 2007; Newman et al., 2008; Otterbacher, 2013; Sabin et al., 2008; Schler et al., 2006; Schwartz et al., 2013). Data on literature in French (Argamon et al., 2009) and Dutch (Koolen, 2018) agree with this pattern.

As we saw above, words referring to emotions, and in particular to positive emotions, has been linked to female discourse. For texts with male authors it has been reported that concepts covering anger (Luoto, 2021; Mehl and Pennebaker, 2003; Schwartz et al., 2013) and negative emotion (Schwartz et al., 2013) are more frequent. Otterbacher (2013), working with movie review data, saw that ‘[t]he themes discussed more often by males as compared with females include […] violence’. This tendency was confirmed by Park et al. (2016: 10) whose ‘[t]op 20 male-linked language topics’ include three which can be labelled ‘fighting’, ‘guns’, and ‘killing’ (my selection of central nouns), respectively. Several studies have shown that swear words, which tend to express anger and other negative emotion, appear more often in male discourse (Bamman et al., 2014; Boulis and Ostendorf, 2005; Luoto, 2021; Mehl and Pennebaker, 2003; Newman et al., 2008; Park et al., 2016; Schwartz et al., 2013). However, Argamon et al. (2007) report an opposite tendency for blogs.

Death-related words are more common in male-authored discourse according to several studies (Luoto, 2021; Park et al., 2016; Sabin et al., 2008; Schwartz et al., 2013). A reasonable conjecture is that this is due to the prominence of violence-related themes in male-authored writing. This is supported by the findings of Park et al. (2016), where the keywords ‘kill, kills, killed, murder, killing, die, swear, dead’ co-occur in one of the top-scoring male-linked topics.

Nouns standing for anatomical concepts were found to be ‘[e]nduring [m]ale [t]erms’ in a corpus of French literature (Argamon et al., 2009), but as we saw above, there is evidence that the opposite holds for Dutch literature.

Several studies have seen that male authors refer more often to political matters (Argamon et al., 2007; Park et al., 2016; Van der Vegt and Kleinberg, 2020). Argamon et al. (2009) found that several ‘enduring male terms’ belonged to an ‘Authority’ thematic group in French literature. The factor analysis of Argamon et al. (2007) uncovered a ‘Games’ factor which was significantly more present in the blogs from male authors. In the study by Park et al. (2016: 7), ‘sports and competition’ appeared as a ‘strongly male-linked topic’. Similar tendencies have been noted by Schler et al. (2006) and Newman et al. (2008), and, concerning ‘Achievement’, by Schwartz et al. (2013).

Transactions involving money is a key mechanism in politics and other competitive fields. So, it is not surprising that words relating to money appear more frequently in male-authored discourse. This has been seen in several studies (Newman et al., 2008; Schler et al., 2006; Schwartz et al., 2013; Van der Vegt and Kleinberg, 2020), including for a corpus of Dutch literature (Koolen, 2018) Park et al., (2016) found a male-linked topic clustering ‘government, economy, tax, budget, pay, taxes, country, income, benefits, obama’, adding support for the connection between money and politics. Sabin et al. (2008) report a conflicting finding, where the category ‘Money & finance’ (LIWC) was positively correlated with female-authored blog posts.
Religion is another area where gender differences have been observed. Argamon et al. (2007) found that male bloggers referred to religion more often than female bloggers. It was also a more common research theme for male scholars in the study by Thelwall et al. (2019). A similar situation has been attested for French literature, where ‘male authors favor religious terminology rooted within the church, whereas female authors spend more time discussing spirituality in a personal, more secular language’ (Argamon et al., 2009). Zhang et al. (2011) found, analysing data from an Islamic web forum, that female and male posters talked about religious matters preferring different words and emphasizing different themes.

To recapitulate, scholars have investigated the differences between male and female discourse in a substantial number of corpus-based studies and they have invariably found pronounced gender-based differences both as regards content-related features and more formal aspects of language. Even if different semantic categories have been used to interpret and summarize the conclusions, many textual properties have consistently, over several studies, across different genres, and to some degree across different languages, been identified as being correlated with either female or male authorship.

3. Method and materials

Before turning to details, let me summarize the steps involved in the research procedure: First, I curated a corpus based on criteria defining the segment of recent popular literature to be explored. Secondly, the texts were submitted to automatic grammatical analysis. Based on the output, I computed relative frequencies for a range of linguistic features for both single books and authors (who in many cases were represented by several works). The last step of the quantitative analysis was to extract those features for which pronounced gender-related differences could be seen according to a threshold criterion. The analysis of the texts departed from text property measures which are theoretically straightforward and surface-oriented. It was carried out by means of Python programs written for the current project. The programs made use of existing open-source natural language processing resources for Swedish.

3.1. Corpus

The corpus behind the present study was collected with the aim to cover a substantial and representative selection of adult fiction in Swedish which has been popular in recent years. In order to steer free from dataset and interpretation bias (cf. Koolen, 2018: 133), I defined the corpus by means of a set of criteria as objectively operationalized as possible. It thus includes all books, including translations, which have earned the status of bestsellers and/or beststreamers in the Swedish book market during the period 2015–2020, along with a batch of original ‘born-audio’ works in Swedish aimed at the audiobook segment. As the collection is complete given the criteria defining it, the study is not based on a random selection process.

The restriction to adult fiction means that the categories history, biographies, teens and young adult, and children’s books were excluded from the corpus. The bestsellers had earned that status according to the Swedish Publishers’ Association (SvF), either in
hardback or paperback. As beststreamers (cf. Berglund, 2021a), I count those novels, not found among the bestsellers, which have been among the twenty most streamed audiobooks each year in the leading Swedish online streaming platform Storytel. In addition, the current corpus includes all adult fiction in Swedish published by the platform in their ‘Storytel Original’ series from its inception in 2016 until May 2021. Storytel Original works are thus ‘born-audio’ in the sense that they have been written for online audio streaming. (They are however at the same time available as e-books in the platform.) Most of these works are available in ‘seasons’ comprising 10–24 episodes. The length of each season is roughly that of a novel, and I concatenated all the episodes of each season into a single novel-like work in this study.

Three labels were used for author gender: female, male, and ‘F/M team’, the last one for teams of authors with both female and male members. ‘Female’ and ‘male’ covered both single authors and one-gender teams. This information was mainly derived from gendered pronouns, pictures, and first names in the presentations of the authors in the publishers’ marketing materials or on their own homepages. The authors in the corpus are in many cases well-known media personalities. I am not aware of any difficulty in the binary gender labelling of the individual authors. Still, there might be discrepancies between the public personas of authors and the facts of their private lives which are unknown to me. As Koolen (2018: 26) notes, a crucial aspect is whether authors are taken to be ‘female or male by their readers’. (The gender labelling is included in the listing of the corpus in the Supplementary Materials.)

I assigned the books in the corpus to three categories which are highly relevant both from a reader and from a book-trade perspective. The largest, most popular, and most easily defined one is crime fiction. I also use ‘prestige’ as a category in the description of the corpus. This category is operationalized by the condition of having been awarded or nominated for a major literary prize. As it happens, no crime novel in the corpus has earned that distinction. This is not surprising in view of the ‘genre fiction’ status of the crime category (Berglund, 2021b). A third category is formed by the books which are neither in the crime nor the prestige class. See Berglund and Dahllöf (2021) for details on this category assignment.

Table 1 gives an overview of the composition of the corpus. A complete listing of the corpus can be found in the Supplementary Materials. Unfortunately, the corpus cannot be shared for copyright reasons.

Table 1. The composition of the corpus.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Bestseller (313)</th>
<th>Beststreamer (56)</th>
<th>Storytel original (98)</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Crime</td>
<td>Prestige</td>
<td>NCNP</td>
<td>Crime</td>
</tr>
<tr>
<td>Female</td>
<td>86</td>
<td>15</td>
<td>56</td>
<td>28</td>
</tr>
<tr>
<td>Male</td>
<td>90</td>
<td>10</td>
<td>47</td>
<td>22</td>
</tr>
<tr>
<td>F/M team</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Σ</td>
<td>185</td>
<td>25</td>
<td>103</td>
<td>53</td>
</tr>
</tbody>
</table>

NCNP: Non-crime non-prestige. There were no prestige novels among the beststreamers and Storytel originals.
The 467 works, totalling 48.5 million words, range in length between 14,000 and 378,000 words. The median length is 96,000 words. The subcorpora authored by female and male authors comprise 23.1 and 23.4 million words (over 235 and 214 books), respectively. Another 2 million words are from two-gender teams. One author (Jan Guillou) is represented by both the largest number of books, 14, and words, 1.7 million, but most of the 197 authors (single or teams), viz. 117, are represented by one book only. There are 104 female authors, 84 male authors, and 9 two-gender teams. The study involved comparisons both at the level of individual works and at the level of authors (see below). Crime is the dominating genre to which 60% (281/467) of the titles belong.

3.2. The statistical analysis of gender differences

I analysed gender differences both for single books and for authors. Books are interesting as the units in which literature comes packaged to readers. When books are the objects of study, the issue is how books written by women are different from books written by men. Studying individual authors is also important, as their books can be seen as expressions of their individual ‘authorial styles’. In order to extract text property measures for the individual authors, I collected all text for each author. The author-oriented part of the study served to eliminate the effects of the uneven representation of individual authors. For authors of several books in the corpus this step also neutralized variation among these books.

I choose to compare female-authored and male-authored texts primarily by means of the notion of probability of superiority, formalized as $P(X > Y)$ (Grissom, 1994; McGraw and Wong, 1992; Wolfe and Hogg, 1971). $P(X > Y)$ is the probability that some feature score ($X$) sampled at random from one population (e.g. a collection of female-authored texts) is larger than the corresponding score ($Y$) sampled from another population (e.g. a collection of male-authored texts). This is a comparatively intuitive and simple measure, which deserves to be used more often in corpus studies. It is a so-called effect size measure, i.e., it quantifies how large a difference is. An example of a score used in the current study is the relative frequency of nouns. As the current study departs from a finite corpus, the $P(X > Y)$ value can be computed by comparing all pairs of one text from the first category and one from the second category. If we consider a hypothetical example with three texts in each category, associated with the values \{20\%, 21\%, 22\%\} and \{18\%, 19\%, 30\%\}, respectively, the first category would ‘win’ in six cases and ‘lose’ in three, so $P(X > Y) = 6/(6 + 3) \approx 0.67$.

Probabilities of superiority close to 0.5 correspond to independence, whereas higher values indicate a positive correlation between the feature and the first category (and, inversely, that values smaller than 0.5 capture a positive correlation between the feature and the second category). It should be easy to see that the relationship $P(Y > X) = 1 - P(X \geq Y)$ holds generally, as $X \geq Y$ holds whenever $Y > X$ does not obtain.

The probability of superiority only quantifies effect sizes in terms of to what extent entities outscore each other or not as regards a text property. It otherwise ignores the magnitudes of the numerical differences between the two scores. In this way it does not rely on an assumption that variables exhibit normal (bell curve) distributions. This is
important, as many linguistic features tend to be absent in many texts and exceptionally common in a few instances.

In previous research, Cohen’s $d$ appears to be the effect size statistic which has most commonly been used to compare the distributions of various measures concerning female and male language use (Koolen, 2018; Luoto, 2021; Newman et al., 2008; Park et al., 2016; Van der Vegt and Kleinberg, 2020). However, the $d$ statistic is based on the assumption “that the populations sampled are normally distributed and that they are of homogeneous (i.e. equal) variance” (Cohen, 1988: 20). This assumption has generally been disregarded when female and male language use have been compared. When it comes to relative frequencies of linguistic elements in text corpora, distributions are, except in cases of very frequent linguistic elements, far from being normally distributed.

The probability of superiority values reported here were based on comparing every text in one subpopulation (in the present case female-authored texts) with every text in the other subpopulation (male-authored ones) for a given text property measure. This means that the present study was a total population survey. Consequently it did not involve any random sampling and the problem of sampling errors does not enter the picture. Significance testing was, for that reason, not relevant in the way it is in sampling-based studies (cf. Hirschauer et al., 2020).

Significance testing can however be applied in a specialized way in corpus exploration. A $p$ value is then interpreted as ‘the probability that [a] keyness is accidental’ (Biber et al., 2007: 138). So, if we find that the probability of superiority value for a certain linguistic property measure deviates from 0.5, we can compute a $p$ value telling us how likely that has come about as a result of random variation in language use. The significance test appropriate for probability of superiority is the Mann-Whitney test (Grissom, 1994).

In order to extract those text property measures which in a strongly attested way are associated with gender differences, I operationalized a threshold criterion defining that condition. This criterion considers both book and author level differences, and it requires that a difference is attested both for the crime books and for the non-crime subset of books. This is motivated by the fact that the crime category comprises 60% of the books, and that crime might be a genre which is gendered in its own way (c.f. Alacovska, 2017). So, a gender difference is strongly attested if and only if either gender gives a probability of superiority larger than 0.6 and an associated $p$ value smaller than 0.01 for each of the following four text cohort comparisons:

1. All author text collections (104 female vs 84 male authors),
2. All books (235 female-authored vs 214 male-authored books),
3. All crime books (132 female-authored vs 134 male-authored books), and
4. All non-crime books (103 female-authored vs 80 male-authored books).

Note that (3) and (4) give two $p$ values smaller than 0.01 for disjoint subsets of the data. As this amounts to a very restrictive threshold, what we will see in the discussion of results is, as it were, the tip of the iceberg. Other less strongly attested and potentially interesting gender-related differences can most likely be found in the data.
3.3. Grammar-based text property measures

The corpus comprised raw text extracted from e-books. I derived a set of measures characterizing texts defined as relative frequencies of parts of speech (such as verb and adjective), inflection categories (such as plural and past tense), and grammatical functions (such as subject and modifier). So, for instance, the fraction of verbs is in this sense a measure characterizing a text, which for the current corpus is in the range from 16 to 26%.

To be more specific about the technical details, the analysis (tagging) of the texts was made by means of the open-source Stanza (2020) system and its ‘talbanken 1.0.0’ model for Swedish trained on the ‘UD Swedish Talbanken’ corpus (Nivre and Smith, 2015), which is based on the Universal Dependencies (UD) framework. UD provides part of speech and dependency labels, based on traditional grammar, but adapted to achieve systematic and cross-linguistically consistent annotation. The dependency labels formalize grammatical functions as relations between words. So, for instance, the subject relation typically holds between a verb and a noun or a pronoun.

For part of speech labels, dependency relations, and punctuation marks I computed their frequencies relative to the total number of words (formed by letters and/or numbers). For morphosyntactic features, the corresponding text property measures were computed in relation to the total number of instances of the associated part of speech. So, for instance, the value for active voice was the fraction between the number of verb tokens in active voice and the total number of verb tokens. In this way the active voice measure would reflect how verbs are used. The frequency of active voice forms relative to the total number of words, which I did not consider, would most likely rank the texts in a way very similar to that of the plain verb measure (the frequency of verbs relative to the total number of words).

This procedure, together with the threshold criterion (of Section 3.2), allowed me to generate toplists of morphosyntactic measures, one for features distinctive of texts from female authors and one for the male authors.

3.4. Text property measures based on graphic words

I also explored relative frequencies of graphic words (defined as sequences of letters as they appear in the texts). These were lowercased to achieve case insensitivity. This is a standard procedure to neutralize use of capitalization, due, for instance, to sentence-initial position. I only considered those graphic words which occurred in at least half of the books in the corpus (i.e. in at least 234 books) in order to remove rarer words. This constraint left me with 4754 graphic words.

Ranking the most distinctive graphic words by means of probability of superiority gave me a batch of top-listed words for the cohort of female authors compared to the male authors, and another one for the comparison made in the opposite direction. These words lists were then analysed through the lens of lexical semantics in order to determine whether some linguistic or conceptual pattern could be found.
4. Results

Tables giving a more complete presentation of the results, including the original graphic words in Swedish and their translations, effect size scores (probability of superiority, Cohen’s $d$), and $p$ values, are included in the Supplementary Materials.

4.1. Gender and grammar-based text property measures

My set-up defined 230 grammar-based text property measures. Of these, 13 met the threshold criterion for being clearly higher in female-authored fiction (Table 2), and 17 of them were in the corresponding manner associated with male writing (Table 3).

In line with the criterion for strongly attested (stated in Section 3.2), the four probabilities of superiority on each row in Tables 2 and 3 are to be read as a presentation of evidence. The scores for the crime and non-crime books concern different (categories of) books and are thus independent. The all book (‘Books’) score derives from all female-and male-authored works being compared to each other, and consequently includes the crime only and non-crime only comparisons, along with all ‘cross-genre’ comparisons (female crime/male non-crime and female non-crime/male crime). The author score is based on comparisons involving the collected works in the corpus for each author, and that probability of superiority consequently depends on data reflected in the three other scores. A situation like the one for verbs in Table 2, where the probability of superiority for authors (0.66) is smaller than that for books (0.75), must be due to the fact that books by the productive authors to a higher degree than those by the one-book authors confirm the dominant pattern. This can also be seen in Figure 3.

Table 2. Top twelve grammar-based text property measures whose value tend to be higher in female-authored text.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Group</th>
<th>Authors</th>
<th>Books</th>
<th>Crime</th>
<th>Non-crime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indefinite</td>
<td>Adjective</td>
<td>Gram. number</td>
<td>0.67</td>
<td>0.72</td>
<td>0.70</td>
</tr>
<tr>
<td>Singular</td>
<td>Adjective</td>
<td>Gram. number</td>
<td>0.67</td>
<td>0.69</td>
<td>0.68</td>
</tr>
<tr>
<td>Uter (common)</td>
<td>Wh-determiner</td>
<td>Gram. number</td>
<td>0.66</td>
<td>0.65</td>
<td>0.62</td>
</tr>
<tr>
<td>Verb</td>
<td>Verb density</td>
<td>0.66</td>
<td>0.75</td>
<td>0.77</td>
<td>0.73</td>
</tr>
<tr>
<td>Uter (common)</td>
<td>Adjective</td>
<td>Gram. number</td>
<td>0.66</td>
<td>0.68</td>
<td>0.62</td>
</tr>
<tr>
<td>Auxiliary (UD)</td>
<td>Verb density</td>
<td>0.65</td>
<td>0.69</td>
<td>0.70</td>
<td>0.71</td>
</tr>
<tr>
<td>Pronoun</td>
<td>Verb density</td>
<td>0.65</td>
<td>0.71</td>
<td>0.71</td>
<td>0.72</td>
</tr>
<tr>
<td>Neuter</td>
<td>Adjective</td>
<td>Gram. number</td>
<td>0.64</td>
<td>0.68</td>
<td>0.71</td>
</tr>
<tr>
<td>Nominal subject (UD)</td>
<td>Verb density</td>
<td>0.63</td>
<td>0.70</td>
<td>0.72</td>
<td>0.69</td>
</tr>
<tr>
<td>Positive</td>
<td>Adjective</td>
<td>degree</td>
<td>0.63</td>
<td>0.65</td>
<td>0.62</td>
</tr>
<tr>
<td>Nominative</td>
<td>Noun</td>
<td>Verb density</td>
<td>0.62</td>
<td>0.68</td>
<td>0.69</td>
</tr>
<tr>
<td>Subject form</td>
<td>Pronoun</td>
<td>Verb density</td>
<td>0.62</td>
<td>0.65</td>
<td>0.62</td>
</tr>
<tr>
<td>Clausal complement (UD)</td>
<td>Verb density</td>
<td>0.62</td>
<td>0.63</td>
<td>0.63</td>
<td>0.62</td>
</tr>
</tbody>
</table>

The scores are probabilities of superiority based on authors’ texts compared as units, on all books, on all crime books, and on all other (non-crime) books, respectively. The ‘Group’ is according to my interpretive reading of the result. Rows ordered by the first score.
The relative frequency of determiner dependencies was the strongest indicator of gender on the author level (probability of superiority 0.75) of all measures based on grammatical features. As we should expect, the corresponding part of speech measure followed suit (0.74). By contrast, verbs had a similar and corresponding status as the ‘signature’ part of speech for female writing (probability of superiority for authors 0.66, for books 0.75).

If we look at other measures it seems that male writing gravitated towards noun phrase complexity, as the high determiner density suggested. This hypothesis is further supported by the high probability of superiority values for noun density in male-authored writing, and for the prominence of dependency relations of the kind connecting elements inside noun phrases, viz. adjectival modifier, numeric modifier, and nominal modifier. Definite form in adjectives, which also appeared in the toplist, only occurs together with modifier (attributive) function and is consequently in line with the dependency measures. Furthermore, high values for cardinal and ordinal numbers and the fraction of nouns being in the genitive case were distinctive for male writing (in line with the results for nominative in female writing). These tendencies speak in favour of the hypothesis that male authors tend to package a larger fraction of their text into noun phrases.

If we look at other measures it seems that male writing gravitated towards noun phrase complexity, as the high determiner density suggested. This hypothesis is further supported by the high probability of superiority values for noun density in male-authored writing, and for the prominence of dependency relations of the kind connecting elements inside noun phrases, viz. adjectival modifier, numeric modifier, and nominal modifier. Definite form in adjectives, which also appeared in the toplist, only occurs together with modifier (attributive) function and is consequently in line with the dependency measures. Furthermore, high values for cardinal and ordinal numbers and the fraction of nouns being in the genitive case were distinctive for male writing (in line with the results for nominative in female writing). These tendencies speak in favour of the hypothesis that male authors tend to package a larger fraction of their text into noun phrases.

If we look at other measures it seems that male writing gravitated towards noun phrase complexity, as the high determiner density suggested. This hypothesis is further supported by the high probability of superiority values for noun density in male-authored writing, and for the prominence of dependency relations of the kind connecting elements inside noun phrases, viz. adjectival modifier, numeric modifier, and nominal modifier. Definite form in adjectives, which also appeared in the toplist, only occurs together with modifier (attributive) function and is consequently in line with the dependency measures. Furthermore, high values for cardinal and ordinal numbers and the fraction of nouns being in the genitive case were distinctive for male writing (in line with the results for nominative in female writing). These tendencies speak in favour of the hypothesis that male authors tend to package a larger fraction of their text into noun phrases.

In the same way as the male-authored texts had an orientation towards noun phrase complexity, female-authored works tended to be richer in structure built directly around verbs, i.e., clausal constructions. This is as expected: When a smaller number of the words are part of nominal constructions, a larger fraction of the words will be more directly (i.e.

---

**Table 3.** Grammar-based text property measures whose values tend to be higher in male-authored text, otherwise as Table 2.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Group</th>
<th>Authors</th>
<th>Books</th>
<th>Crime</th>
<th>Non-crime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Determiner (UD)</td>
<td>NP complexity</td>
<td>0.75</td>
<td>0.79</td>
<td>0.80</td>
<td>0.77</td>
</tr>
<tr>
<td>Determiner</td>
<td>NP complexity</td>
<td>0.74</td>
<td>0.80</td>
<td>0.80</td>
<td>0.78</td>
</tr>
<tr>
<td>Masculine</td>
<td>Adjective</td>
<td>Same-gender (M)</td>
<td>0.72</td>
<td>0.72</td>
<td>0.74</td>
</tr>
<tr>
<td>Noun</td>
<td>NP complexity</td>
<td>0.70</td>
<td>0.72</td>
<td>0.66</td>
<td>0.78</td>
</tr>
<tr>
<td>Numeric modifier (UD)</td>
<td>NP complexity</td>
<td>0.70</td>
<td>0.79</td>
<td>0.83</td>
<td>0.72</td>
</tr>
<tr>
<td>Nominal modifier (UD)</td>
<td>NP complexity</td>
<td>0.70</td>
<td>0.76</td>
<td>0.75</td>
<td>0.77</td>
</tr>
<tr>
<td>Cardinal number</td>
<td>NP complexity</td>
<td>0.69</td>
<td>0.75</td>
<td>0.80</td>
<td>0.68</td>
</tr>
<tr>
<td>Comparative</td>
<td>Adverb</td>
<td>Degree</td>
<td>0.69</td>
<td>0.70</td>
<td>0.70</td>
</tr>
<tr>
<td>Ordinal number</td>
<td>NP complexity</td>
<td>0.67</td>
<td>0.73</td>
<td>0.72</td>
<td>0.75</td>
</tr>
<tr>
<td>UTR/NEU</td>
<td>Adjective</td>
<td>NP complexity</td>
<td>0.66</td>
<td>0.69</td>
<td>0.67</td>
</tr>
<tr>
<td>Masculine</td>
<td>Pronoun</td>
<td>Same-gender (M)</td>
<td>0.66</td>
<td>0.71</td>
<td>0.72</td>
</tr>
<tr>
<td>Participle</td>
<td>NP complexity</td>
<td>0.64</td>
<td>0.64</td>
<td>0.61</td>
<td>0.68</td>
</tr>
<tr>
<td>Genitive</td>
<td>Noun</td>
<td>NP complexity</td>
<td>0.64</td>
<td>0.68</td>
<td>0.66</td>
</tr>
<tr>
<td>Adjectival modifier (UD)</td>
<td>NP complexity</td>
<td>0.63</td>
<td>0.68</td>
<td>0.71</td>
<td>0.66</td>
</tr>
<tr>
<td>IND/DEF</td>
<td>Adjective</td>
<td>NP complexity</td>
<td>0.63</td>
<td>0.65</td>
<td>0.63</td>
</tr>
<tr>
<td>Definite</td>
<td>Adjective</td>
<td>NP complexity</td>
<td>0.62</td>
<td>0.66</td>
<td>0.63</td>
</tr>
<tr>
<td>Superlative</td>
<td>Adverb</td>
<td>Degree</td>
<td>0.62</td>
<td>0.67</td>
<td>0.68</td>
</tr>
</tbody>
</table>


The relative frequency of determiner dependencies was the strongest indicator of gender on the author level (probability of superiority 0.75) of all measures based on grammatical features. As we should expect, the corresponding part of speech measure followed suit (0.74). By contrast, verbs had a similar and corresponding status as the ‘signature’ part of speech for female writing (probability of superiority for authors 0.66, for books 0.75).
not as embedded in noun phrases) constituents of clauses. So, we found higher values for verbs, pronouns, and the fraction of subject form in pronouns in female-authored writing. Pronouns are, in comparison to noun phrases, more often used without modifiers to refer to things and to people. The high fraction of subject form pronouns (in Swedish these are: jag, du, han, hon, vi, ni, and de corresponding to English, I, you, he, she, we, you, and they), which are used for reference to persons, suggest that female writing has an orientation towards reference to people. The higher values for the verb-headed dependency relations nominal subject and auxiliary (verb) in female writing also support the idea that these parameters reflect an emphasis on clausal construction. The top-listing for female writing of nouns being in the nominative case is also in line with their being used directly in clauses, as opposed to being used as modifiers in the genitive case. (The opposite tendency, for genitive case in nouns, was seen for male text.) The top-listed dependency relation ‘clausal complement’ also creates clausal complexity unmediated by noun phrases.

One of the most basic inflectional distinctions in Swedish, that between singular and plural, also turned out to be strongly associated with gender differentiation. The values for singular forms in adjectives tend to be higher among female writers. (The higher values for indefinite, common, and neuter forms are a consequence of the fact that these properties are only marked for forms in the singular.) A follow-up inspection of the data showed that the female/singular and male/plural tendency generalizes to nouns, pronouns, adjectives, and the adjective-like past participles (see the Supplementary Materials).

Interestingly, grammatical number in determiners was almost independent of author gender. The high ranking for male authors of the fractions of adjectives underspecified for grammatical gender and for definiteness seem to reflect the fact that the tagger labels adjectives as such when they appear as modifiers (attributes). So, this is a somewhat artificial consequence of noun phrase complexity. Masculine forms in adjectives (e.g. gamle, old) and in pronouns with adjectival inflection (e.g. denne, this) were also relatively more frequent in male writing. This is a stylistically marked feature, as the corresponding feminine forms are used gender-neutrally in many registers of Swedish. (There is no explicit feminine counterpart to the masculine in these cases in the tagset.)

Another group of strongly attested differences is related to degree in adjectives and adverbs: A high fraction of positive forms in adjectives is associated with female writing, whereas the relative frequencies of comparative and superlative forms of adverbs tend to be higher for the male group. A follow-up inspection of the data showed that this generalizes to adjectives and adverbs across the board (see the Supplementary Materials).

4.2. Gender and text property measures based on graphic words

As mentioned above, 4754 measures met the condition that restricted the attention to those graphic words which were found in at least every second book in the corpus. The criterion of strongly attested correlation with author gender extracted 113 ‘female’ and 92 ‘male’ graphic words from the corpus. The twelve top-ranking entries are listed in Tables 4 and 5. (The complete tables are found in the Supplementary Materials.) Many of the top-listed graphic words fell naturally into conceptual categories. Even if this step in the analysis is
qualitative application of lexical semantics, these classes were robustly not overlapping between the genders.

Almost twenty graphic words meeting the threshold criterion in favour of female-authored fiction represented concepts covering emotional states and processes. Three fourths of them stood for positive emotions, notably different words for love, joy, and gratitude. On the opposite polarity side I found, e.g., sad, angry, and embarrassed. Another large group of words (14 entries) stood for properties of inducing emotions, i.e., evaluative adjectives and adverbs. Positive polarity examples included beautiful, fun, and wonderful. Negative evaluation was expressed by words like terrible, troublesome, and boring. A number of words representing other mental states were also more common in female language use: think (tänka, two forms), hold an opinion (tycka, two forms), want (v) or will (n), pretend, and dare.

Some concepts for body parts were also prominent in female fiction. Examples included forms of the nouns, tear, hair (two forms), ear, hairstyle, cheek, curls (Swedish lockar, which also means attract), heart, and eyebrow. The majority of these concepts cover parts of the face. If we take physical appearance to be a more general category, a group of words standing for clothes and personal accessories could be added here. The examples include feminine items like klänning (dress) and handväska (handbag).

Another salient theme behind words more frequently used by female writers was food and beverage, as reflected in a dozen words like eat, kitchen, chocolate, food (two forms), coffee, plate, mug, kitchen, eat, and the verb for preparing food, laga (which also means mend). Social relations and interaction was the theme behind a somewhat heterogeneous group of words: hug, marriage, talk, and hang out. (It is plausible that the Swedish verb störa, disturb, to a high degree has been used in interpersonal contexts and should be added here.) Concepts related to homes seemed to be another category emphasized in

Table 4. Top twelve graphic words by author-based probability of superiority whose relative frequencies tend to be higher in female-authored text.

<table>
<thead>
<tr>
<th>Swedish</th>
<th>Translation</th>
<th>Group</th>
<th>Authors</th>
<th>Books</th>
<th>Crime</th>
<th>Non-crime</th>
</tr>
</thead>
<tbody>
<tr>
<td>åta</td>
<td>eat</td>
<td>food</td>
<td>0.76</td>
<td>0.75</td>
<td>0.74</td>
<td>0.79</td>
</tr>
<tr>
<td>glad</td>
<td>happy</td>
<td>emotion (positive)</td>
<td>0.75</td>
<td>0.73</td>
<td>0.64</td>
<td>0.82</td>
</tr>
<tr>
<td>hem</td>
<td>home/homewards</td>
<td>home</td>
<td>0.74</td>
<td>0.73</td>
<td>0.69</td>
<td>0.77</td>
</tr>
<tr>
<td>bo</td>
<td>live</td>
<td>home</td>
<td>0.74</td>
<td>0.71</td>
<td>0.69</td>
<td>0.74</td>
</tr>
<tr>
<td>passa</td>
<td>fit</td>
<td>verb</td>
<td>0.74</td>
<td>0.68</td>
<td>0.65</td>
<td>0.72</td>
</tr>
<tr>
<td>laga</td>
<td>prepare (food)/mend</td>
<td>food</td>
<td>0.73</td>
<td>0.72</td>
<td>0.67</td>
<td>0.79</td>
</tr>
<tr>
<td>hon</td>
<td>she</td>
<td>same-gender (F)</td>
<td>0.73</td>
<td>0.77</td>
<td>0.82</td>
<td>0.72</td>
</tr>
<tr>
<td>tårarna</td>
<td>the tears</td>
<td>body</td>
<td>0.72</td>
<td>0.76</td>
<td>0.81</td>
<td>0.70</td>
</tr>
<tr>
<td>snygg</td>
<td>beautiful/handsome</td>
<td>evaluation (positive)</td>
<td>0.72</td>
<td>0.70</td>
<td>0.62</td>
<td>0.79</td>
</tr>
<tr>
<td>roligt</td>
<td>fun</td>
<td>evaluation (positive)</td>
<td>0.72</td>
<td>0.73</td>
<td>0.75</td>
<td>0.71</td>
</tr>
<tr>
<td>háret</td>
<td>the hair</td>
<td>body</td>
<td>0.71</td>
<td>0.82</td>
<td>0.84</td>
<td>0.79</td>
</tr>
<tr>
<td>kär</td>
<td>in love</td>
<td>emotion (positive)</td>
<td>0.71</td>
<td>0.67</td>
<td>0.64</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Otherwise as Table 2.
female writing as expressed in the high scores for words with the senses (at) home/homewards, live (in a place, two forms), and move (to a new home).

Fiction by female authors was also richer in four interjections: åh (oh), herregud (oh my god), oj (oh), and hej (hello). This is probably related to more use of dialogue by female authors (see Section 5.3).

Five intensive adverbs, so, very (two synonyms), unbelievably, and precisely, were used to a larger extent by female authors. (One downtoner word, the verb form seemed, tycktes, is more common in male-authored text. This fact can possibly be interpreted as an expression of a communicative inclination opposite to that of the adverbs associated with female-authored fiction.)

The third person singular pronoun for female referents (hon) also scored higher for female writing in all its forms (hon, henne, hennes, i.e. she, her, her[s]). A follow-up inspection of the data (see the Supplementary Materials) showed that a corresponding situation holds for male writing and male-reference personal pronouns. The tendency for same-gender reference was also reflected by the nouns mom (two forms) and lady, which were top-listed for female writing. Correspondingly, a masculine form of a demonstrative pronoun with adjectival inflection, denne (this), was top-listed for male language use, as were two forms (mannaen and männnen) of the noun man.

A number of verb forms also came out as relatively more frequent in female writing. This might, in particular for the high-frequency verbs, be seen as a reflection of the general tendency to higher verb density. The same explanation might tentatively be applied to the higher relative frequency of the negation inte. As the word attaches to verbs, its relative frequency is likely to be dependent on that of verbs.

When we look at the 92 graphic words which came out as top-listed for male-authored writing (see Table 5 for the twelve top-ranking entries) we find a number of quite obvious classes represented by several words: There is a striking group of more than a dozen words related to weapons and animosity: shoot (v), weapon (two forms), shot (n), arrest, gun

<table>
<thead>
<tr>
<th>Swedish</th>
<th>Translation</th>
<th>Group</th>
<th>Authors</th>
<th>Books</th>
<th>Crime</th>
<th>Non-crime</th>
</tr>
</thead>
<tbody>
<tr>
<td>meter</td>
<td>meter</td>
<td>geometry</td>
<td>0.77</td>
<td>0.78</td>
<td>0.80</td>
<td>0.78</td>
</tr>
<tr>
<td>hundra</td>
<td>hundred</td>
<td>numeral</td>
<td>0.75</td>
<td>0.76</td>
<td>0.79</td>
<td>0.70</td>
</tr>
<tr>
<td>fran</td>
<td>from</td>
<td>preposition</td>
<td>0.74</td>
<td>0.76</td>
<td>0.70</td>
<td>0.83</td>
</tr>
<tr>
<td>Ett</td>
<td>a, an, one</td>
<td>determiner</td>
<td>0.72</td>
<td>0.71</td>
<td>0.72</td>
<td>0.68</td>
</tr>
<tr>
<td>rök</td>
<td>smoke</td>
<td>x</td>
<td>0.72</td>
<td>0.66</td>
<td>0.63</td>
<td>0.72</td>
</tr>
<tr>
<td>tredje</td>
<td>third</td>
<td>numeral</td>
<td>0.71</td>
<td>0.76</td>
<td>0.76</td>
<td>0.74</td>
</tr>
<tr>
<td>vänster</td>
<td>left</td>
<td>direction</td>
<td>0.71</td>
<td>0.72</td>
<td>0.75</td>
<td>0.67</td>
</tr>
<tr>
<td>riktad</td>
<td>directed</td>
<td>direction</td>
<td>0.70</td>
<td>0.65</td>
<td>0.65</td>
<td>0.64</td>
</tr>
<tr>
<td>två</td>
<td>two</td>
<td>numeral</td>
<td>0.70</td>
<td>0.76</td>
<td>0.80</td>
<td>0.70</td>
</tr>
<tr>
<td>Norrut</td>
<td>North</td>
<td>direction</td>
<td>0.70</td>
<td>0.68</td>
<td>0.69</td>
<td>0.66</td>
</tr>
<tr>
<td>fört</td>
<td>led</td>
<td>x</td>
<td>0.70</td>
<td>0.64</td>
<td>0.61</td>
<td>0.67</td>
</tr>
<tr>
<td>riktade</td>
<td>directed</td>
<td>direction</td>
<td>0.70</td>
<td>0.64</td>
<td>0.64</td>
<td>0.62</td>
</tr>
</tbody>
</table>

As Table 4.
(two forms), war, enemies, and police (man) (two forms). The authority theme, it seems, also manifests itself in the nouns leader, order, and the verb control.

An emphasis on geometric concepts also seem to be distinctive for male writing. A dozen graphic words stand for direction-related concepts: directed (three forms); left, right, and up; north (three forms), west, and south. Ten forms stand for concepts related to location and distance: meter, height, places, far, middle, line, and closest. Four words for parts of buildings might be mentioned in this context: corridor and entrance (or gate), door, and wall.

Cardinal and ordinal numbers, and the word for quantity (two forms), was another prominent category for male fiction, as shown by a dozen entries in the toplist. It is also noteworthy that graphic words associated with quantification of time were more common in male writers, viz. hour, minutes, and seconds. These time units are, of course, often used together with cardinal numbers. The comparative and superlative adjective forms closest, largest, smaller, larger, and lower are also evidence for the quantitative tendency in male writing.

As noted in Section 4.1, male language use has a tendency towards a higher determiner density. This is also reflected in the use of the specific graphic words en and ett, which can be used both as number words (one) and as indefinite determiners (a, an).

I found that two prepositions have a clearly gender-dependent distribution: från (from) and av (of, by) are ‘male’ prepositions. This tendency is of particular strength for från. Its semantics obviously aligns with the direction theme, discussed above.

In addition to the graphic words I have discussed, there are a few solitary ones, such as flowers (female) and name and smoke (male), which cannot in any obvious way be associated with underlying themes given the other top-listed words.

5. Discussion

A summary of the most important findings of the present study on the differences between female and male language use in fiction is given in Table 6.

5.1. Verb and noun phrase density

A fundamental stylistic dimension of Swedish and many other languages is the relative amount of text that is embedded in noun phrases. I saw clearly that male writing tends to be placed higher on that scale (see Section 4.1). This result chimes with the observation of Newman et al. (2008: 229) that men have an inclination ‘to talk about concrete objects, which require nouns and, of course, articles’. For female-authored text there was an opposite tendency, in line with grammatical expectations, with higher frequencies for verbs and pronouns. These observations can be seen as supporting the ‘people/things theory’, which holds that women are more interested in people and men in things (Luoto, 2021; Thelwall et al., 2019), since discourse about what people do and how they interact often refer repeatedly to them by means of single pronouns.

We can take a closer look at the stylistic situation by plotting all books according to verb and determiner density, which were the two measures with the highest book-based probabilities of superiority for the two genders, as in Figure 2. We see that female-
authored books tend to be high in verb density and low in determiners, whereas works by male authors exhibit the inverse tendency. We should note that the patterns emerge on cohort level. So, distributions like those in Figure 2 do not show that each individual author writes in some gender-typical style. For these two text property measures, there are

Table 6. A summary of the findings.

<table>
<thead>
<tr>
<th>‘Dimension’</th>
<th>High in ‘female’ texts</th>
<th>High in ‘male’ texts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammatical structures</td>
<td>Verbs, clausal construction</td>
<td>Noun phrase complexity</td>
</tr>
<tr>
<td>Reference</td>
<td>Pronouns referring to persons</td>
<td>Singular indefinite articles</td>
</tr>
<tr>
<td>Third person reference (only singular; plural gender-neutral in Swedish)</td>
<td>To female referents</td>
<td>To male referents</td>
</tr>
<tr>
<td>Grammatical number</td>
<td>Singular</td>
<td>Plural</td>
</tr>
<tr>
<td>Degree in adjectives and adverbs</td>
<td>Positive</td>
<td>Comparative and superlative</td>
</tr>
<tr>
<td>Adjectives and adverbs</td>
<td>Evaluative concepts</td>
<td>Weapons and animosity</td>
</tr>
<tr>
<td></td>
<td>Intensive adverbs</td>
<td>Numbers and quantification</td>
</tr>
<tr>
<td>Semantic themes</td>
<td>Positive emotions</td>
<td>Geometry- and direction-related concepts</td>
</tr>
<tr>
<td></td>
<td>Emotional and other mental states and processes</td>
<td>Location and distance</td>
</tr>
<tr>
<td></td>
<td>Social relations, home, eating and food</td>
<td></td>
</tr>
</tbody>
</table>

Features which stand in a contrary relation to each other are aligned.

Figure 2. How books distribute according to verb and determiner density. As recorded in Tables 2 and 3, books by female authors contain a higher fraction of verbs than male-authored books in 80% of the cases and vice versa for determiners in 75% of the cases.
dense regions of the distributions characterized by a high degree of gender balance. These regions represent, as it were, an ‘average’ style. (Of course, ‘style’ is a much richer concept than what can be captured by only two measures.) At the same time there are more extreme regions which are clearly dominated by either gender.

Figure 3 shows the same book data points as Figure 2, but with the books by a number of authors highlighted. They include the top five authors in terms of their number of books in the corpus along with the authors who occupy ‘fringe’ positions in the region created by the two measures. Most of these authors are remarkably consistent as regards the two aspects of style over their different books. Jonas Hassen Khemiri, known as a stylistically creative author, is an exception: The three novels by him are clearly more dispersed than the other authors’ work. If we take the data in Figure 2 to define a scale of ‘stylistic femininity’, so to speak, a female author (Karina Berg Johansson) appears as the most ‘feminine’ author, but the least ‘feminine’ author (Olga Tokarczuk) is also a female author. The plot also shows that some productive authors, in particular Jan Guillou, consistently had more extreme styles.

5.2. Thematic differences

Both previous research and the present results suggest that clear differences between female and male language use to a large extent are seen in the amount of attention given to different content-based themes. It is also plausible, as argued above, that differences concerning grammatical properties primarily depend on which kinds of semantic content authors emphasize in their writing, rather than on more abstract stylistic preferences.
Data concerning third person singular pronouns and some other features showed that there is a same-gender reference tendency for both female and male authors. Female authors refer more to female individuals than male ones do. For male authors there is the opposite pattern, but to a lesser degree.

A striking tendency was the wealth of concepts covering emotive states and processes which are more common in female discourse, reflected in the graphic word measures. It was also clear that female authors deal much more with positive emotions than male ones do, but that the corresponding differentiation is weaker for negative emotions.

Concepts related to social interaction exhibited a similar pattern: Female writing tends to highlight friendship and other positive interpersonal emotions, in agreement with the female inclination towards positive feelings. A prominence of the theme of eating, food, and beverage was also evident in the word top-lists, as was the theme of having a home. Both themes are often associated with positive, or at least ordinary, social interaction. The evidence for an opposite tendency in male writing was also strong. In the narratives of male authors, animosity and the use of weapons are much more prominent. Similar results have been among the strongest conclusions of previous studies.

Another previously noted gender-related tendency that was also seen in the present study is that female authors use intensive adverbs more often than male authors do. I could also observe that a group of evaluative adjectives and adverbs were top-listed for female-authored fiction, as were a number of concepts standing for basic cognitive processes, e.g., thinking and holding an opinion. Some concepts relating to body parts and clothes were also more common in texts by female authors. The body theme had also been found by Koolen (2018: 139), who studied a corpus in Dutch which was otherwise quite similar to mine.

Describing things in quantitative terms is clearly one of the themes more common in male writing. This tendency seems to go hand in hand with male authors paying more attention to directions and spatial conditions. The differences I saw concerning grammatical number and degree in adjectives and adverbs seem to be new findings. They are however related to quantitative language as plural has a more-than-one semantics and as comparative and superlative, so to speak, express how much of a property a referent has in comparison with other objects.

The differences between the genders concerned linguistic features across the spectrum from very frequent grammatical categories to rarer content-related themes. The results also showed that the gender differentiation in recent popular Swedish fiction to a high degree match what has been reported for other languages and situations of language use.

5.3. Limitations

A corpus study like the present one focuses on certain parameters and submits them to a predefined statistical treatment. Other text properties could obviously have been explored. Additional dimensions of authorial gender performance which deserve closer inspection are, for instance, related to the genders of the narrator, the narratee, and the characters in the story. This is interesting not least in the context of dialogue, which represents a variety of language use which is different from frame narration (Dahllöf, 2022). Using data from Dahllöf (2022), I can see that there is more dialogue in female-authored fiction: the
probability of superiority for the fraction of words being embedded in quoted utterances is 0.61 if we compare female- and male-authored books. It is possible that some of the gender-related differences in grammar and vocabulary can be explained by that circumstance.

Apart from the discussion about noun phrase density in Section 5.1, I have not systematically analysed statistical dependencies among the various text property measures. A limitation of the current analysis is thus that it considered each measure in isolation. Methods such as factor analysis and topic modelling would most likely reveal meaningful statistical connections among groups of measures.

As mentioned in the introduction, some scholars argue that there is a danger that research of this kind ‘treasures differences over similarities’ (Bamman et al., 2014: 153) and that ‘results are interpreted using gender stereotypes’ (Koolen, 2018: 129). In response to potential objections of this kind, I would like to claim that the set-up of the present study was based on methodologically straightforward and simple components. Furthermore, the main lines of my interpretation of the quantitative results have been justified by run-of-the-mill grammatical analysis and lexicology. The procedure has not been tweaked to deliver results that would agree with given preconceptions. With other data it could have delivered negative results or found gender-related differences of completely different kinds.

### 6. Conclusions

The present study has in a systematic way addressed the question of to what extent and how authors’ gender is reflected in grammatical and lexical text properties of bestselling fiction in Swedish 2015–2020. As far as I can tell, this is the first study of its kind applied to Swedish literature. The corpus was compiled for this purpose according to popularity-based criteria and ended up well balanced between female and male authors, measured by word count. This suggests that female authors are not marginalized in contemporary popular Swedish literature in the same manner as they have been in many other times and places by an overwhelmingly stronger presence of male authors.

My results showed that a substantial fraction of the grammatical and lexical properties of texts correspond to highly significant differences, on the cohort level, between books written by female and by male authors. These gender-related differences were in most cases easy to interpret as conforming to patterns that have been reported for other languages and categories of language use. In particular, the gender performance inherent in the present corpus agrees with traditional stereotypes about the interests of women and men. This is, for instance, evident in the domain of concepts covering emotions. We can
almost directly recycle an observation like the following, which was originally supported by data from natural conversations among American students: ‘Women used more references to positive emotion, but men referred more to anger—a finding that is perfectly consistent with gender stereotypes’ (Newman et al., 2008: 229, restating the findings by Mehl and Pennebaker, 2003). The second part must however be modified somewhat for my results: The most common, but also quite mild, Swedish word for angry (arg) is in fact top-listed for female-authored fiction. What agrees with the stereotype is that male authors pay more attention to the more extreme expressions of aggression, i.e., violence.

Even if they largely seem to reproduce well-known patterns, these results are remarkable in view of the fact that Sweden in many ways is a society characterized by a high degree of gender equality. According to the European Institute for Gender Equality (EIGE, 2022: 121), Sweden is top-ranking among EU countries for five (‘Work’, ‘Knowledge’, ‘Time’, ‘Power’, and ‘Health’) out of six ‘domains’ of gender equality (‘Money’ being the exception), and has occupied the overall top rank since 2010. Modern Swedish fiction should be the place to find gender-equal language use, if ever there was one.

Modern Swedish fiction still to a high extent illustrates the old feminist observation that many authors perform gender in ways that ‘reproduce[…] the classic representations of women’ (Cixous, 1976: 878) by highlighting stereotypical themes which corroborate ‘women’s confinement to the household and forms of service associated with motherhood’ (Armstrong, 2006: 103). Further work is needed to assess to what degree modern Swedish authors treat these themes in conventional ways or whether they rather challenge stereotypes in their storytelling.

Some scholars regard the language use of authors as mainly a matter of self-expression: ‘[T]he contents of novels, by and large, reflect the products of authors’ imaginations, life experiences, and personal interests rather than what is present in their immediate environments’ (Luoto, 2021: 2). There are, however, many social factors which might influence the choices authors make in their writing and the ways in which their careers develop. As we saw in Section 2.1, popular fiction is a cultural domain where gender performance enters the picture from all directions, e.g., in the presentation of authors, the behaviour of readers, and in the way characters are depicted. It is consequently likely that many readers and publishers think that female and male authors have different abilities to write well in different areas of fiction, encouraging stereotypical alignments between authors’ gender and the character of their work. In view of that, it is not surprising that the texts themselves reflect the gender of their authors.

Another gender-related factor which points in the same direction is the choice of performing narrator when publishing companies produce audiobooks. If we tabulate author gender against narrator gender, as in Table 7, we see that there is a high degree of agreement. If the author is female, so is the narrator in 89% of the cases. For works by male authors, the corresponding percentage is 78%. This correlation provides another kind of evidence for the gender performativity associated with popular fiction. It shows that publishers consider the gendered qualities of these texts important for the way they should be presented to audiobook consumers.

It would be interesting to explore how the kind of gender-related differences I have seen in this study affect readers. Classifier experiments suggest that corpus statistics in some cases can be more sensitive than ordinary readers. For instance, in one study (Burger
et al., 2011), automatic prediction of author gender from tweet data reached 76% accuracy, which was a better score than that earned by 95% of the human raters. However, let me just add that stylistic dispersion of the magnitudes that was exhibited in Figure 2, i.e., in the relative frequencies of verbs and determiners, is likely to influence the style of novels in ways that give readers quite different impressions about the character of the texts, in particular if we consider that there is also internal variation in the texts. Graphic word frequencies to a high degree reflect content, and can in many cases be assumed to quite directly correspond to a reader’s experience of which themes are highlighted in a story.

Today’s online digital distribution systems change the ways in which literature (as well as other cultural products) is handled in society. They give readers much more freedom to access and read whatever ‘content’ they like. By contrast, printed books are physical objects which come with considerable costs and practical limitations. In the pre-digital era people could only access books that they had at home, could afford to buy, had an opportunity to borrow, or otherwise lay their hands on. This probably contributed to a more unified reading culture, e.g., in the area of Swedish fiction.

Audiobook and e-book streaming services, which encourage consumers to sign up for subscription plans allowing them to access thousands of books, have achieved an increasingly dominant standing in recent years. Given the widespread use of sophisticated user profiling for content promotion and feed ranking algorithms, the production and marketing of literary works are becoming more and more tuned towards catering for individual interests and preferences. Text classification components in ‘gender inference modules’ (Zhou and Moreels, 2012) are likely to play a prominent role in the feedback loops among readers, authors, publishers, and streaming services. Features capturing gender and different attitudes to gender are likely to be central parameters in the models driving this development. A consequence of this might be that differences between the genders are amplified in the ‘echo chambers’ created by the algorithmic feedback loops. In Swedish party politics, another arena of storytelling (which just like popular fiction must stir the interest of its audience), there are ‘clear gender gaps in political priorities and preferences, and these have rather increased in the last decade. Women are today more left leaning than men are’ (Oskarson and Ahlbom, 2021: 1). It is not unlikely that the digital ‘echo chambers’ of contemporary culture contribute to widening gaps between the genders across the different branches of culture.

Acknowledgements

The PI of the present project, Karl Berglund, contributed to the design of the study and provided feedback along the way. In particular he compiled the data on bestseller and beststreamer status and book-trade category, as well as the gender labelling of authors and performing narrators.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was carried out in the project ‘Patterns of Popularity: Towards a
Holistic Understanding of Contemporary Bestselling Fiction’ funded by Vetenskapsrådet (Swedish Research Council) (2019-02829).

**ORCID iD**
Mats Dahllöf  [https://orcid.org/0000-0002-4990-7880](https://orcid.org/0000-0002-4990-7880)

**Supplemental Material**
Supplemental material for this article is available online.

**Notes**

1. Summers (2013) saw that ‘of the 60 authors chosen as favorites by [29] male participants, 57 were male and 3 were female’. By contrast, the 29 female participants listed 44 male authors and 19 female authors as their favourites. Summers also saw that 24 (of the 29) male informants preferred books with male protagonists, as did 6 of the female ones. Only 4 female informants preferred female protagonists. The rest, 5 male and 19 female participants, reported that they had no protagonist gender preference.

2. LIWC is a proprietary software, whose procedures and repository of dictionaries are not freely available to scholars. This is a disadvantage from a research point of view.

3. The studies exploring topic modelling relied on latent Dirichlet allocation as implemented by the Mallet software to extract both semantic topics and associated dictionaries.

4. Koppel et al. (2003) trained a Balanced Winnow classifier to tell texts by female authors from those written by men, working with both fiction and non-fiction prose. Argamon et al. (2009) used Support Vector Machines (SVM) as their classifier framework, and extracted the feature weights in order to find ‘the most influential words in the model’. Dahllöf (2012) trained SVMs as a means for measuring and analysing to what degree age, gender, and political affiliation (in binary operationalizations) influence language use in speeches from the Swedish parliament. The results suggested that age is more difficult to predict than gender, and that political leaning is the easiest to predict of the three kinds of personal trait.

5. Most of the studies using LIWC have not directly analysed the relative frequency of nouns, as ‘the LIWC articles category, which consists of determiners like ‘the’, ‘a’, ‘an’ […] serves as a proxy for the use of more nouns’ (Schwartz et al., 2013: 8).

6. The term of ‘probability of superiority’ was introduced by Grissom (1994), but the concept had previously been promoted by Wolfe and Hogg (1971), and, as ‘common language effect size’, by McGraw and Wong (1992). By defining effect size in terms of probability of superiority ‘we can easily avoid the trap of using normal distributions when they are obviously inappropriate’ (Wolfe and Hogg, 1971: 27). McGraw and Wong (1992: 361) argue that ‘[t]he primary value of the proposed statistic is that it is better than the available alternatives for communicating effect size to audiences untutored in statistics’.

7. Cohen’s $d$ (Cohen, 1988: 66–67) is ‘calculated by dividing the mean difference by the pooled standard deviation’ (Newman et al., 2008: 220). Cohen (1988: 26–27) operationalizes the thresholds for small, medium, and large effect sizes as $d > 0.2$, $d > 0.5$, and $d > 0.8$, respectively.

8. McGraw and Wong (1992: 361) argue that Cohen’s $d$ suffers from the disadvantage that it ‘yields a value that has no anchor points in the experience of nonstatisticians.’ However, as
Cohen’s $d$ is a commonly used measure, I included it among the scores reported in the Supplementary Materials.

9. The relative frequency of the lexeme *pyramid* is an example. A work like Brown’s *Da Vinci Code* by itself makes the mean for the male-authored text population larger than the value for the vast majority of its members.

10. Biber et al. (2007), discussing extraction of keywords, i.e., words ‘used more frequently in the selected group of texts’, put it as follows: ‘The significance ($p$ value) represents the probability that this keyness is accidental’ (Biber et al., 2007: 138).

11. The Talbanken corpus comprises data from informative text genres. I considered the Stanza resource a good state-of-the-art off-the-shelf solution for grammatical tagging, even if it is not directly tailored to literary text.

12. Extraction of graphic words does not necessitate the use of language-specific software. From the point of view of linguistic analysis, graphic words combine lemma and morphological information, and they do not achieve a separation of senses in cases where a graphic word is ambiguous (homonymy). An example of this is Swedish *banan*, which can be singular indefinite of *banan* (*banana*) as well as singular definite of *bana* (*the track*). However, cases of homonymy are not frequent enough to generally undermine the robustness of graphic words as indicators of content.

13. These cases were however slightly below the threshold of strongly attested.

14. The remaining three cases were however below my threshold for strongly attested.

15. The probabilities of superiority for male-reference pronoun forms (Swedish: *han*, *honom*, *hans*, English: *he*, *him*, *his*) indicated smaller differences, for *han* and *hans* just below the threshold and for *honom* less significant.

16. In a patented invention (currently assigned to Meta Platforms Inc) ‘Inferring user profile attributes from social information’ (Zhou and Moreels, 2012) there is a ‘gender inference module’ which might use fiction preferences in the following way: ‘Another factor used to infer gender of a user includes interests of the user. Typical biases of users of a particular gender towards particular interests are analysed based on historical data […] For example, correlations with particular types of books, movies, television programs may be used to infer gender of users. For example, users with a particular gender may be more likely to be interested in romantic novels as compared to action novels.’ (Zhou and Moreels, 2012: 5).

References


**Author biography**

Mats Dahllöf holds a PhD in General Linguistics and is Senior Lecturer/Associate Professor in Computational Linguistics at Uppsala University, Sweden. He has worked in semantics and in different areas in natural language processing, most recently focusing his research on digital philology, in particular on modern literary texts.