Lost in Translation: An Investigation of Provider Fairness in Book Recommender Systems

Msc Thesis Computer Science & Engineering

Rareș-Dorian Boza

TU Delft

Lost in Translation: An Investigation of Provider Fairness in Book Recommender Systems

Msc Thesis Computer Science & Engineering

Thesis report

by

Rareș-Dorian Boza

to obtain the degree of Master of Science
at the Delft University of Technology
to be defended publicly on TO Update!

Thesis committee:

Chair:	Sole Pera
Supervisor:	Sole Pera
Daily Supervisor:	Sole Pera
External examiner:	Masoud Mansoury
Place:	Faculty of Electrical Engineering, Mathematics, Computer Science, Delft
Project Duration:	From 1th Oct 2024 - To 26th May 2025
Student number:	5685281

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

Faculty of EEMCS · Delft University of Technology

Abstract

Recommender Systems are key instruments that are constantly being employed in the online environment as a method of connecting users and items. Due to the resulting personalised suggestions, users benefit from quickened decision making, however, such systems can also introduce unwanted side-effects, by reinforcing existing biases or limiting exposure to diverse content. While much of the research surrounding unfairness and its effects has been conducted specifically to benefit the users, the creators of the distributed items, namely the providers, can also be subject to inequity. From the perspective of a provider, garnering visibility and exposure of their items through these systems converts into revenue. Recommender Systems can be a positive means in sectors where, historically, groups of providers have been disadvantaged from reaching their desired consumers, although if mishandled can become an additional hindrance. Amongst the many relevant domains, the book industry is a clear example where authors have been discriminated based on traits unrelated to the quality of their writing. Publishers have manifested an adversity towards translated works leading to an innate disadvantage when it comes to their distribution and marketing. Still, the fairness of how recommender systems handle translated works when competing with other books remains unexplored. To address this gap, we conduct an empirical exploration of several state of the art recommendation algorithms, evaluating their performance with respects to accuracy and provider fairness. This allows us to discern if any of the algorithms are a helpful conduit or a harmful one. Upon identifying inequity concerning foreign authors, we probe the ability of mitigating it on an algorithmic level, by applying a reranking method. Outcomes stemming from this work, reveal high representation of books available in English in recommendations, while highlighting the advantages of foreign manuscripts with translations over those without. The outlined findings inform future design of Recommender Systems and their provider fairness evaluation.

Preface

This thesis signifies a pivotal achievement in my academic career that would have otherwise not been possible without the people that supported me throughout this endeavour.

Carolina, there are no words to describe the amount of unyielding support you have offered me during this stressful months of intensive work. Thank you for reigniting my passion for reading when we first met, which inadvertently brought me to the subject of this study. Most of all, you have taken to soothing my worries and you never stopped being my #1 fan through your encouragements.

My grandma Iuliana and my parents, you spent a significant portion of your lives looking after me and making sure I receive whatever opportunities you could have offered me to achieve my dreams and be the best version of myself.

My friends, you were there when I needed a break from work and to take my mind off whatever issue was troubling me. Your excitement and curiosity about my chosen topic helped me see its true value and relevance.

Sole, thank you for the advisory role that you undertook during the supervision of my research. Not only did you offer invaluable feedback that benefitted the quality of my work, but you took your time with me during a period I struggled with burnout and helped me turn my thesis into a passion project. When I entered your office for the first time I was uncertain of what subject I was going to tackle and how, but you stuck with me and passed down your enthusiasm for research unwaveringly.

Contents

List of Figures

List of Tables

Part I

Foundations

1 Introduction

Recommender Systems (RSs) comprise algorithms and techniques that support their users by connecting them to suitable items in various contexts [1]. One of the widely used methods to achieve this is analysing historical data, such as in the form of user-item interactions, and learning through it to deduce future behaviour [2]. With an ever-increasing amount of information available to the end-user to filter through [3], the presence of RSs has soared in recent years in the online medium, as a means of presenting relevant items in a quicker, more efficient manner to customers [4, 5]. Most of the platforms employing RSs are operating a two-sided marketplace [6], where they act as intermediaries between two categories of users, consumers and producers.

Research surrounding the improvement of RSs has traditionally focused on increasing the accuracy of user rating estimation towards unobserved items [7]. Improving the quality of the recommendation lists generated by an RS was furthered by metrics such as serendipity, tracking items that are not actively looked for but prove relevant to the user, or novelty, representing content of unknown relevance [8, 9]. Even so, these methods of quantifying a recommender list's utility fail to capture perpetuating social bias stemming from the data used for facilitating RSs [10]. Consequently, when concerned with historical data employed in an RS, ingrained biases of the respective industry will inadvertently be mimicked. These characteristics systematically harm the experience of users in an RS ecosystem.

A topic that has been gaining traction in the study of RSs is the analysis of their *fairness* [11, 12]. Unfairness is a consequence of some existing biases induced or proliferated by RSs. At a consumer level, lack of fairness creates stark differences in accuracy and beyond accuracy metrics for users of different backgrounds. Shifting the viewpoint from the consumer side when addressing RS fairness to include other types of stakeholders [13] is an effort started in recent years. From a producer standpoint, unfairness leads to less exposure and lower ranking performance of items, despite having similar utility to other entries on the list. This effect is observed especially on a group level, where individuals are divided generally by any fairness related attribute such as age, race or gender. Unfortunately, commercial RS services are often not transparent in revealing the true business value of the utilised system towards their item providers [14], masking the effects on their revenue.

Impact-Oriented RS analysis [15] is a growing approach that tries to stir attention away from focusing solely on accuracy and towards more valuable perspectives. It advocates for better informed decisions that concentrate on the intended purpose of the system. Consequently, it highlights that domain-specific aspects are often overlooked and a high level abstraction for the best-performing model will not translate to every premise. Moreover, it encourages tackling unfairness on a case by case basis, with a better understanding of the scenario it is implemented in. In the case of provider fairness, endeavours to define diversity of content in a recommendation list, have failed to reach an overarching standardised definition [16]. Current consensus instead suggests focusing on different key areas of RSs perceived by stakeholders to improve this standpoint, such as the background of item producers in relation to their exposure to the end users.

By following the paradigm of tackling unfairness through tailoring our approach to each specific instance, we focus on implications of fairness in the book domain, a sector which suffers from historical inequity from various perspectives [17, 18]. As an effect stemming from market globalisation, popularisation of e-books and availability of online book communities, book authors have growing agency in reaching new readers. Equity amongst book authors in terms of having their books recommended to potential readers given equal

utility is crucial as it can directly affect the revenue of the producer. It has been previously shown that RSs in book scenarios proliferate bias concerning author gender and country of provenance [19, 20].

Figure 1.1: Top 10 books of "The 100 best books of the 21st century" as aggregated by The New York Times. The top shelf illustrates foreign books while the bottom one shows books originally written in English.

Authors are able to opt for translations to bridge the language gap between foreign readers and their books [21]. This proves advantageous especially when trying to reach audience in international, online, reader communities which prove to be largely English-centric [22]. Nevertheless, it is currently unknown whether the availability of a manuscript in a more accessible language is reflected by equity between these works and similar ones originally written in English in the context of recommendation. Commercial solutions such as GoodReads give us little insight into how the employed RS performs in terms of desired fairness measurements. Older forum posts [1][2] authored by the GoodReads service offer a study pointing out that users do not rely on their algorithmic recommendations and guide them to steps they can take to improve their feed. Instead, they are listed with mandatory actions to undertake to trigger non-English book recommendation. This suggests that certain items, especially with respect to language, encounter issues in being recommended to their target audience.

The behaviour of RSs when dealing with item providers from the two identified groups, either English natives or foreign authors, has significant ramifications. Commissioning a translation of a book implies more spent resources, be it time, financial costs or others, on top of the effort invested in creating the actual manuscript [21, 23]. The expected value of such an undertaking is equitable increase in exposure to a level of that of a similar book available in English [21]. Should this not be the case, we expect a systematic loss of income and recognition for authors of less spoken languages. Furthermore, as RSs can steer public opinion by capturing consumers in their content feeds [24, 25], discrimination of providers based on language of origin discourages existing and emerging lesser known cultural expressions, styles of writing and topics by not giving them enough visibility to the public.

From the outlook of the consumer, access to a broader catalogue of books with respect to the language of provenance enables them to discover and sympathise with other cultures [26]. Consequently, translations enable the user to benefit from this intercultural exchange, irrespective of their linguistic abilities. Especially when concerning children, diverse literature can dismantle stereotypical perceptions and enhance critical thinking regarding sensitive topics [27]. Figure 1.1 showcases what provider unfairness regarding linguistic backgrounds can look like from a consumer perspective. The editorial staff of The New York Times Book Review compiled a list following responses from their followers. The resulting "Top 100 Best Books of the 21st Century"[3] exhibits a staggering ratio of almost $9{:}1$ in favour of English originals over foreign translated books. Having this type of bias proliferate at an RS level can steer the reader away from diversifying their choice of literature.

[1] https://www.goodreads.com/blog/show/343-how-do-books-get-discovered-a-guide-for-publishers-and-authors-who-want

[2] https://help.goodreads.com/s/article/How-Can-I-Improve-my-Recommendations

[3] https://www.nytimes.com/interactive/2024/books/best-books-21st-century.html

In this work, we conduct an empirical exploration to examine the perceived advantage offered by translations to those authoring books in languages other than English. With the core principle of offering authors equal chances of popularising their work, we investigate various RSs with respect to their fairness in the context of translated literature. This effort is driven by the following research questions:

Research Question 1

What are the existing distributions of foreign literature in the examined book collection as well as the available user profiles?

Research Question 2

To what extent do RSs proliferate unfairness towards translated foreign works in favour of books originally written in English?

Research Question 3

If a bias against translated foreign books is identified, to what extent could mitigation strategies be employed in order to decrease said partiality, while balancing accuracy?

To answer **RQ1** we perform an initial exploration of the book corpus available in the GoodReads dataset [28]. We extend the available catalogue to contain information regarding the original language and translation availability into English of the books in question. We quantify the frequency of the items in the established provider groups and subsequently compare and contrast our findings to their distribution in user profiles. This effort will serve as a confirmation of the under representation of translations in book publishing [23], but also offer insight about whether the consumer activity follows the same trend.

To address **RQ2**, we focus on the presence of the target provider groups in the top-50 recommendations generated for users through the use of various recommender algorithms. We formulate a perspective on how different recommender algorithms may favour certain types of providers.

RQ3 explores the ability to combat the previously highlighted case of unfairness efficiently with existing solutions. We opt for a reranking algorithm to redistribute the generated recommendations in the user lists, analysing any changes. It is important to offer solutions that introduce equity among authors while also keeping into account the satisfaction of the users of the RS [12].

With this work[4], we advance knowledge in the scrutiny of provider fairness in RSs, particularly within the book domain. Focusing on the books' language traits and English translation availability, we emphasise the different treatment RSs exhibit with regard to various provider groups. The key contributions of this study are outlined as follows.

- We enhance a commonly used dataset with user-rating book interactions to contain the availability in English for each entry, by merging it with other public resources.
- We show that the book data corpus has a higher known proportion of manuscripts written in English. By comparing distributions of several genres to the baseline, we outline that there are significant differences in their representation to our expectations and the overall catalogue. We study the bulk of the available user profiles to find that in comparison to the book catalogue, their own rating distribution is non-homongenous, with a tendency to be dissimilar to the trend of the corpus.
- We measure the fairness of provider groups aggregated by language characteristics in an RS environment to find advantages gained by manuscripts with translation into English over those who do not. A low population class of items with ambiguity regarding origin is found to receive favourable imbalance in exposure and visibility, reinforcing the hypothesis that English available books are prevalent in recommendations.
- We observe the mitigation impact of a reranking approach and note a closer yet marginal resemblance of provider groups representation to that of the user profiles, together with a loss in performance.

[4] https://github.com/raresboza/thesis_lost_in_translation

To our knowledge, this is the first study that investigates provider fairness in RSs from the viewpoint of language accessibility and how this should reflect in increased visibility and exposure to users that have engaged with similar items.

The rest of this thesis is organised as follows. In Chapter 2 we discuss the background and related literature informing our work. Since it centres on fairness in book recommendation, we report on the challenges accentuated by different types of bias towards authors in the book sector. With regards to RSs, we document previous efforts investigating book RSs, as well as past studies into RS provider fairness that influence our research. Given the exploratory nature of this work, in Chapter 3 we pay special attention to the dataset curation process and how we set up our experiments. We describe the results achieved by carrying out the outlined experiments and analyse how they relate to provider fairness from the perspective of language and translation availability to English in Chapter 4. In Chapter 5 we answer the research questions we defined and discuss the implications of the discrepancies of provider treatment on RS fairness. Limitations and future directions are also documented rigorously. Chapter 6 contains details about the ethical conduit we followed during this process, together with a reflection on the effect of our contributions. In Chapter 7 we conclude our thesis by highlighting the key takeaways.

2

Related Work

In this chapter, we reflect on previous literature emphasising contributions on which we build upon and background informing our work.

2.1. Fairness in the Book Market

We take notice of how equity challenges in the industry tend to pose continuous and glaring obstacles for authors from various perspectives. While separate from RS research, these factors can have a trickle-down effect in the field, manifesting themselves as various type of biases [15].

With respect to gender, publishers [17] have a higher preference towards works authored by male writers in both traditional and indie settings. The revenue gap between genders persists in either scenario, with pen names still remaining a common strategy to score comparable sales to male counterparts [29]. In terms of visibility and acclaim, reviewers tend to respond more positively towards authors of the same gender, but also the male reviews are getting more attention from the community at large [18].

When it comes to language representation and accessibility of literature from diverse cultural backgrounds, the situation too presents itself as dire. In general, central languages [23], described as languages exporting the most translations to others, have the smallest number of translations as a percentage of their own book publishing, with the UK and US being as low as 5% in the early $2000s$. Even if, in the last decade, translations have received a more favourable outlook [21] through digital printing and print-on-demand as well as the rapid growth of social media as a promotional tool, authors trying to break into the international market still suffer from various issues, including lack of visibility and availability on the market. For instance, on GoodReads, users from the US make up to 40% of the readers, while roughly half of the books labelled as classics overlap with the English literature curriculum [30]. This medium is often embedded in e-readers, devices that benefit from growth in usage [31] and in reliance for book discovery by the consumers [22].

Throughout our study, we focus on the concept of works [19] rather than individual book editions. We can define a work as the abstract, intellectual content of a piece of literature, independent of any specific physical or digital form. It represents the overarching creative entity that can have multiple versions, editions, or adaptations. Thus, works are the main concern of this research since we are interested in fairness regarding a work's exposure, regardless of the number of publications it has.

2.2. Recommender Systems in the Book Domain

Focusing on advances in RSs geared towards domain specific scenarios provides clear perspectives on the applicability of the intended areas of scrutiny [15]. We cover the facets of research surrounding book RSs to provide an overview of the current literature on this topic.

The accuracy of the recommendations for the consumer is a heavily investigated direction, as Kotkov et al. [32] carried out user satisfaction surveys to assess the impact of the proposed solutions to find the best serving approach. Graph-based algorithms are introduced [33] as an option to tackle large scale recommendation scenarios and comparisons are fulfilled on the aspect of relevancy ranking [34] and topological characteristics [35]. The capabilities of large language models are considered for their added value in both collaborative and content-based algorithms [36, 37]. The explainability of the generated

recommendations with the help of large language models represents an important characteristic of these solutions [38].

Bias proliferated by RSs in the book domain is analysed and tackled from various standpoints. Consumer-oriented fairness places emphasis on sensitive user groups such as children and how they are susceptible to stereotypes in recommendations [39]. In addition, popularity bias is examined for its impact on consumers and how it can be mitigated [40, 41]. We take note of attempts to diversify [42] and introduce novel books [43] in the content recommended for readers. Concerning provider fairness, studies examine treatment of various groups of authors or books that are predetermined on traits relating to causes of discrimination [20, 19, 44, 45]. Efforts [46] were conducted to show the tendency to exacerbate determined biases and unfairness throughout the constant interaction with book RSs.

2.3. Provider Fairness in Recommender Systems

Provider fairness [12] builds upon traditional item fairness with the additional perspective of grouping the contents of the recommendation catalogue based on certain predetermined characteristics. The core issue which is under scrutiny in this context relates to the RS facilitating unwarranted advantage to one or more item groups at the expense of others. This can manifest in a systematic decrease of item visibility and exposure affecting disadvantaged provider groups [11].

Popularity bias [41, 47] is described as the phenomenon where items with more interactions available in the consumer history are prioritised over lesser-known items when generating future recommendations. Although popularity bias is not directly linked with the concept of fairness, which touches on normative ideas of what recommendations should look like [11], it was shown that it can be a direct cause of provider unfairness in book recommendation [45]. Similarly, the diversity of recommendations is negatively affected by the existence of such bias, concept which has been shown to relate to social characteristics of providers in the public perception [16]. Research [48] emphasised the concern of artists over discrepancy in visibility of tracks caused by popularity bias and continuous efforts [49] seek to establish a clear link between this factor and social imbalances in recommendations.

We present several studies of investigating provider fairness, spanning various domains and covering diverse viewpoints of provider categorisation. In the context of music recommendation, Lesota et al. [50] found that US-originating music is overrepresented in the lists generated for users, while items from sparsely portrayed countries in the available catalogue fail to reach their interested audience. Fairness from the standpoint of the provider's country of provenance was also analysed in the movie, education and book settings [51, 20, 52, 45]. In the book recommendation scenario [45], it was revealed that books authored by American writers have a higher representation in the book corpus and were favoured in the generated lists by the RSs. Gender imbalance proliferated by RSs was under scrutiny [19, 44] to unveil the data bias existing in the available catalogue, together with the observation that some algorithms push imbalanced ratios to the users, while others have a diversifying effect.

There are multiple ways to recognise and measure provider unfairness in an RS ecosystem. With regard to the item catalogue, data bias can be inferred by assessing the item distribution relating to the inspected social viewpoint [44, 45]. Additional insights are drawn by contrasting catalogue findings with the available user history to garner the potential impact an RS can have in the environment [19, 53]. The methods of evaluation regarding recommended lists for the consumers include tracking their diversity, but also visibility and exposure of each provider group. To mitigate adversities identified by these metrics, previous endeavours have successfully relied on reranking strategies [51, 52, 20, 54], which consist of an approach that adjusts the positions of the items in the lists, created by recommender algorithms, through a chosen heuristic [12, 11].

2.4. Dataset Curation for Social Fairness in Recommendation

Carrying our RS-related evaluation typically involves a dataset of user activity over an item catalogue to further research objectives. Prominent datasets are available and their widespread use in research allows for comparison of proposed techniques against established benchmarks as well as increased reproducibility of experiments [55]. Nevertheless, some sources have restricted access to their data as their user interactions are withheld under strict licensing by relevant media sources. This causes some of the datasets to not be able to be updated further since the API used to gather information is no longer

publicly available by the service providers [28]. As such, it hinders researchers from having access to latest trends in user-item interaction.

Due to the absence of attributes describing item characteristics that are under scrutiny for vulnerability to inequity, efforts centred on the provider fairness in book recommender systems often lack access to datasets that are immediately suitable to their research objective [11]. Hence, these studies have to enrich the established datasets used in RS research in order to be able to conduct experiments with respect to the proposed viewpoint or focus. These endeavours, culminating in curated datasets, are welcomed contributions in the scientific community. For instance, efforts concerning geographical imbalance in item recommendations [51, 20, 56] or gender equity [44] employed several external APIs in order to ascertain the country of production of the inspected book.

To further provider fairness in RS research regarding broad social aspects, efforts are made to provide tools that aid dataset expansion with desired metadata about items. The PIReT Book Data Tools encompass a publicly available pipeline that links together various open-source datasets in order to associate ratings with relevant book information. Ekstrand et al. [19] defines different types of data that can be aggregated by such a system: consumption data, book data and author data. Consumption data is obtained from sources that contain historical user profiles and keep a registry of their interactions with items. Book data is sourced from entries containing information on the specific manuscript, while author data relates to the writers themselves. The tool takes advantage of common identifiers between the types of data above to agglomerate them into a single curated dataset, such as ISBNs or the name of the primary author. This enables the tool to extract the desired information for a variety of use cases. In particular, Ekstrand et al. [19] focuses on author gender, and by means of the described system, links said demographic data back to the book catalogue in the consumption data.

Part II

Study

3

Methodology

This chapter outlines the data sources, and methods that are employed in this study to explore the identified research gap and address the established research questions, thus establishing an experimental setup.

3.1. Dataset

There is a wide range of datasets that contain information about user behaviour in conjunction with a book catalogue. However, none of the options we consulted offers the full information required about the books, specifically regarding the original language of writing and the availability of translations to English. In our study, we use the GoodReads(`GR-base`) [28] dataset, as it registers real user data about the consumption patterns of books on one of the most popular online platforms centred on literature discussions. We present an overview of the relevant information tracked in `GR-base` and detail the language inference and sampling processes we conduct to further our analysis.

3.1.1. Overview of User-Item Interactions

`GR-base` encompasses reviews, ratings and other interactions characteristic of the platform. These actions are attributed to anonymised users, which includes their profiles and activity history. The aforementioned records were gathered through the now unavailable GoodReads API[1] by analysing the users' public bookshelves. We take only the explicit feedback given by the ratings in our analysis. These entries record an assessment of the item given by the user on a scale of 1 to 5 together with the timestamp of the interaction. The number of users, items and transactions of `GR-base` as well as other derived datasets can be viewed in Table 3.1.

In terms of the metadata available on the items, the book entries are annotated with information containing the title, authors and the relevant ISBN of the edition the user interacted with. The editions are connected through an arbitrarily set identifier representing the manuscript in all of its formats. Subsets of ratings portrayed in `GR-base` categorised by several book genres are also precomputed and ready for use. We gloss over other details that are attributed to these books, either because they are beyond the scope of our research (e.g. publication year) or the field in question is found to be predominantly empty, such as in the case of edition language.

3.1.2. Language Inference

Since `GR-base` does not contain language information for book data, we extend the original dataset to contain both the availability of the items in English as well as the language the works were originally written in. This curated dataset shall be referred to as `GR-lang`.

To simplify our scenario, we aggregate the language into binary categories: English and other. Translations into English are also sorted in this manner, tracking their availability. Cases where the original language used during authoring cannot be established are considered ambiguous, separated from instances where no language can be inferred. As such, by considering the aforementioned cases, we reach the following labelling scheme:

- `eng-original` (EO): the book was written in English.

[1] `https://help.goodreads.com/s/article/Why-did-my-API-key-stop-working`

- `other-translated` (OT): the book was written in another language, but has at least one edition in English.
- `other-not-translated` (ONT): the book was written in another language, with no translation found in English.
- `ambiguous` (AMBG): language of authoring could not be deduced, although editions in both English and another language were found.
- `unknown` (UNKN): no language information could be traced to the specific book.

We employ the PIReT Book Data developed by Ekstrand et al. [19] in order to carry out this inference efficiently. The pipeline was initially designed with the goal of extracting author gender from the given resources. Due to the change in scope, several additions had to be made in order to adapt it to obtain the language availability of the items in the book catalogue. An outline of the sources used and the additional information extracted can be observed in Figure 3.1.

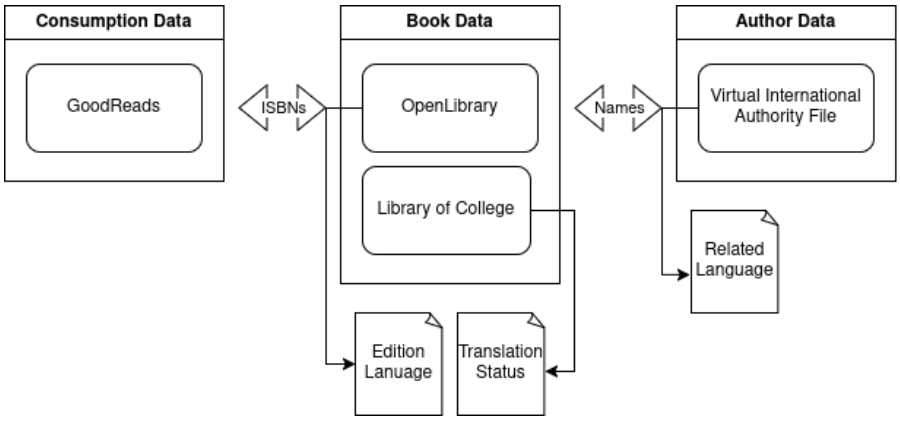


Figure 3.1: Diagram outlining the data linking process and language inference pipeline. (adapted from Ekstrand et al. [19])

The process of aggregating the book catalogue into works and inferring the translation status of a book in English is done by consulting various data sources about the work itself and about the primary author. We use the following information:

- **OpenLibrary** (OL) - for any works, we check all belonging editions that are saved in this data dump. Out of the available editions, we check if the language field is filled in and record it.
- **Library of Congress** (LoC) - all editions in this data dump are stored following the Marc21 format[2], making it convenient to easily find relevant information within the structured entries. To further our inference process, we track the availability of `MARC Authority Field 41` across all records describing the belonging language code. This datafield first marks whether the edition is a translation of the work. If so, two language codes are present, representing source and destination. Otherwise, only the original language of the work is inserted.
- **Virtual International Authority File** (VIAF) - we check the records for all the authors linked with the relevant works in `GR-base`. These entries also respect the Marc21 format. In this datadump, we check the records for the availability of `MARC Authority Field 377` and store its contents for each author. This datafield contains the main associated language for the respective individual.

We prioritise the information extracted from LoC. Apart from being able to infer the original language, this is also the only source that directly refers to the work's translation status. Then, we consider the languages gathered from OL. As these entries only specify the language of the entries, we cannot infer an original language unless there are no editions with differing languages. Lastly, if required, we check the associated language of the author as registered by VIAF. The reason we assigned the lowest priority level to it is because it does not refer to the work itself, but rather to the individual.

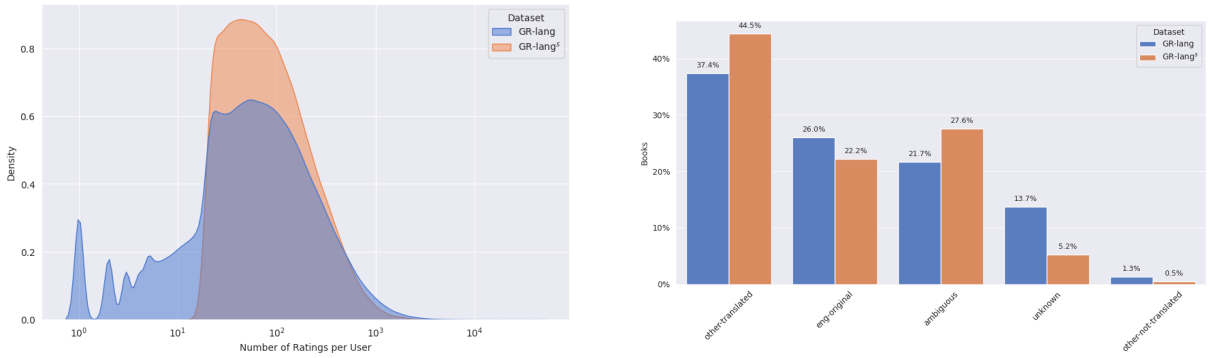
[2]https://www.loc.gov/marc/bibliographic/

3.1.3. Sampling

Due to hardware limitations and considering the size of GR-lang, we use a sample of the dataset GR-lang^s to carry out RS-related experiments. We sample 20000 users taking into account the overall user activity together with the language distribution of the entire rating corpus. This is done by fitting users in bins given their rating frequency. When sampling, users that resemble the global language rating distribution of GR-lang have increased probability of being chosen. This fosters the ability of GR-lang^s to mimic the user activity of the dataset it was derived from. Post-sampling, we apply k core filtering on both users and items in order to reduce the sparsity of the dataset. The values used are $k_{user} = 20$ and $k_{item} = 10$. The ratings of the transactions are binarised, with the threshold for book enjoyment being set to 3.

Dataset	Users	Items	Transactions
GR-base	816,371	1,500,768	104,028,929
GR-lang	816,371	1,500,768	104,028,929
GR-lang^s	18,827	42,101	1,452,684

Table 3.1: Datasets used in our research.



(a) User activity comparison.

(b) Language presence ratios comparison.

Figure 3.2: Comparisons of the sampled dataset GR-lang^s with the source dataset GR-lang.

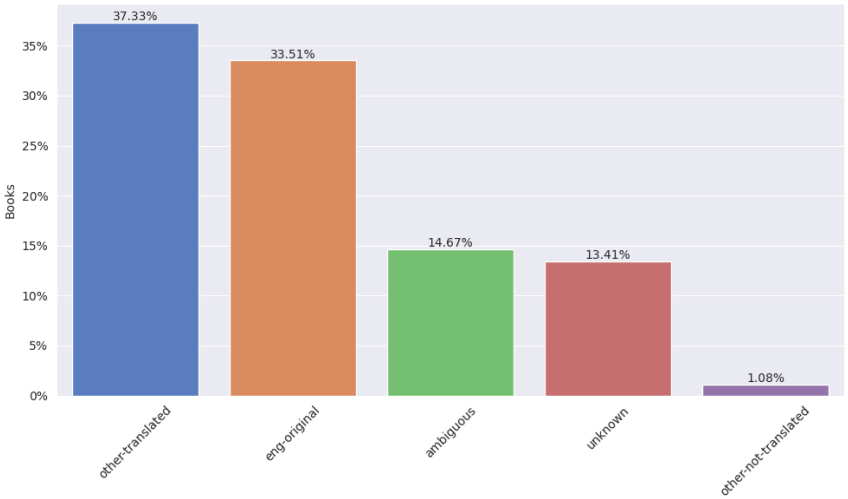


Figure 3.3: Language distribution of items in GR-lang^s.

The shape of GR-lang^s can be seen in Table 3.1. The resemblance of GR-lang^s to GR-lang in terms of user consumption and activities can be observed in Figure 3.2. The distribution of the item catalogue of GR-lang^s with respect to the defined language tags, outlined in Figure 3.3, displays a fairly equal

spread between books that are English originals (33.51%) and those that have a translation available in this language (37.33%). Items that have editions in both English and foreign language (14.67%) are tied alongside those without any retrievable information to this aspect (13.41%). Only 1.08% are items of foreign literature without translations.

We clarify our considerations when selecting a representative sample of the user base found in `GR-lang` to undergo scrutiny. The user activity in `GR-lang` strongly resembles a power law distribution, supported by the heavy tail distribution observed in Figure 3.4. To prevent user trend analysis from being polluted by the assessment of very small user profiles, we remove 15% of the users due to their low activity. This assures that every user has at least 11 items rated. Power users with more than 2500 ratings have also been excluded from this analysis as they were categorized as outliers. As every user, no matter the amount of activity, is weighted equally amongst its peers, the aforementioned measures prevent skewing user profiles analysis conducted in this study.

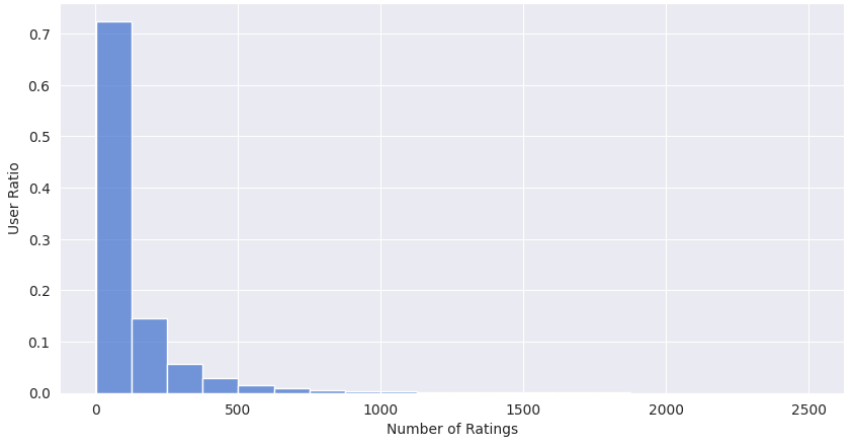


Figure 3.4: Distribution of users by number of ratings in `GR-lang`.

3.2. Algorithms

This section covers details involving the RS pipeline considered in our paper, focusing on both the generation of recommendation and their post-processing.

3.2.1. Recommender Algorithms

We consider a wide range of recommender algorithms with diverse underlying architectures.

For a non-personalised baseline that provides a minimum benchmark to compare other techniques against, we choose the Random Recommender, which is a naive solution that adds items to the recommendation list at random. In terms of neighbourhood-based approaches, we opt for ItemKNN [57] which deduces similarity of items to generate relevant recommendations for users.

Most of the collaborative filtering algorithms we employ are based on latent factor models. MF2020 [58] is a competitive matrix factorisation approach that learns user similarities using the dot product. BPRMF [59] generates recommendations by employing a Bayesian optimisation criterion for pairwise ranking in matrix factorisation. SLIM [60] is a sparse linear method that utilises regression for top-n recommendation problems. PMF [61] employs a probabilistic model for user interests aimed at improving effectiveness of recommendations in sparse data scenarios.

We opt for autoencoder methods to complement the list of recommeder algorithms inspected in this study. MultiVAE [62] uses multinomial likelihood to model user-item feedback data. ItemAutoRec [63] is built to predict missing user ratings for a particular item, while remaining computationally advantageous.

3.2.2. Reranking Recommendations

For the post-processing of recommendation lists generated by the recommender algorithms, we turn to Calibrated Popularity (CP) as a reranking mitigation strategy.

CP [40] is a reranking approach arguing that items on the recommendation list should mimic the popularity proportions of items consumed by the user in the past. CP creates a recommendation list L_u by determining the optimal set of items L_u^*, as

$$L_u^* = \underset{L_u, |L_u|=n}{\text{argmax}} \ (1-\lambda) \cdot \text{Rel}(L_u) - \lambda \cdot JSD(P, Q(L_u)), \tag{3.1}$$

where λ is the weight between relevance and popularity in the calibration, Rel(L_u) is the sum of predicted scores for items in L_u, J is the Jensen-Shannon Divergence, P is the distribution of popularity of items in the user's profile, Q is the distribution of popularity of items in the user's recommendation list.

The popularity of items is computed by quantifying the number of rating interactions related to them [64, 65, 50]. Items are clustered into three categories: high (top 20%), tail (bottom 20%), and medium, comprising the rest of the catalogue. The reranking weight managing relevance and popularity is set to $\lambda = 0.90$, as we seek to outline the strength of mitigating such bias. The choice for a high λ is grounded in previous studies that employed this reranking strategy with similar goals [41, 40].

The CP reranking method was successful in a movie recommendation scenario [66], where it included more tail items in the list while conserving relevancy for the user. Furthermore, Ungruh et al. [41] show that fairer recommendations achieved through CP does not undermine the effectiveness of recommendations regarding user satisfaction. Apart from the past performance of this reranking strategy, we motivate our choice to tackle popularity bias on the basis that translations have a difficult time breaking into the mainstream market of English readers [21], thus leading to disproportionate activity surrounding these books.

3.3. Metrics

In this section, we detail the metrics considered for the evaluation of the recommendation lists and the scrutinised book catalogue.

3.3.1. Distribution Similarity

In order to assess the similarity between the book catalogue and the user profiles in `GR-lang`, we adopt the Jensen-Shannon Divergence (JSD) metric [67]. This is defined as

$$JSD(P||Q) = \frac{1}{2}KL(P||M) + \frac{1}{2}KL(Q||M), \tag{3.2}$$

where M is the mixture distribution of P and Q and KL represents the Kullback-Leibler Divergence between the two distributions.

A resulting JSD value of 0 signifies that the compared distributions were identical, while the more dissimilar they are, the higher it becomes. We normalise the results of JSD in the range $[0, 1]$ for clarity during analysis.

3.3.2. Relevancy of Ranking Measurement

The performance of the recommender algorithms in terms of relevancy of their output to the consumer is evaluated with the following criteria, based on previous literature [12, 39, 68]:

- Normalised Discounted Cumulative Gain (nDCG): indicates the ranking quality of the recommendation list based on the relevancy and positions of the items in comparison with the expected outcome.
- Mean Average Precision (MAP): assesses the ability of the RS to place relevant items in higher positions of the recommendation list.
- Mean Reciprocal Rank (MRR): measures how soon the first relevant result appears in the recommendation list.

3.3.3. Fairness Measurement

To analyse provider fairness in this study, we opt for the following fairness metrics:

- Disparity Exposure [51] (DE): measures the difference in exposure of across interest groups of providers. Namely,

$$DE(G) = \left(\frac{1}{|U|} \sum_{u \in U} \frac{\sum_{pos=1}^{k} \frac{1}{\log_2(\tilde{r}_G^u(pos)+1)}}{\sum_{pos=1}^{k} \frac{1}{\log_2(\tilde{r}_I^u(pos)+1)}} \right) - R_I(G), \tag{3.3}$$

where G is a specific interest group, U is the set of users, $\tilde{r}_G^u(pos)$ is the relevance of the item belonging to G for the user, $\tilde{r}_I^u(pos)$ relevance of the item for the user and $R_I(G)$ is the ratio of items belonging to G in the whole item catalogue.

DE is bounded by $[-R_I(G), 1 - R_I(G)]$. A group has a DE value of 0 when there is no disparate exposure present, any other values across the domain pointing to lower or higher exposure in the recommendations given the group's representation.

- Ranking-based Equal Opportunity [69] (REO): determines how similar are the true positive rates of the interest groups, given the recommendation lists of users. Per interest group, the REO value determines the probability of belonging items to be ranked in the recommendation list, given that the respective user likes them. A lower overall REO value emphasises a lower bias of recommendation. Specifically, this highlights that all item groups have the opportunity to reach interested users.
- Gini Coefficient [12, 11] (GC): indicates the inequality between individual items in terms of the frequency distributions. The domain of GC is bounded by $[0, 1]$, where a lower value underlines fairer recommendations. We study this metric group-wise [70], but also for the whole sample of items. A maximum GC value entails that a given item or group is recommended to all users.
- Language Ratio Difference (LRD): illustrates the mean difference between group ratios in the user profiles and recommendation lists. It is adapted for our multi-group setting from the Misinformation Ratio Difference definition formulated by Pathak, Spezzano, and Pera [71]. This value ranges in the interval $[-1, 1]$. A negative LRD value denotes a higher presence of that item group in the recommendation list compared to the user profiles, while a positive value highlights the opposite.

3.3.4. Statistical Significance

To grasp the statistical significance of our experimental results, we adopt the Wilcoxon Signed-Rank Test. This is a non-parametric test that ranks the difference between two result samples and checks if their distribution is symmetric around zero [72]. The p threshold used is 0.01 and we apply the Bonferroni correction when we compare multiple hypotheses at once [53].

3.4. Experimental Setup

This section describes the framework underlying the study of RSs with respect to provider fairness that guides this study.

3.4.1. Overview of Experiments

To address the RQs of this study, we conduct several experiments, which we describe below.

Language & Translation Presence in Source Data. We focus on identifying the representation of language but also translation availability into English in the book catalogue which would be leveraged by RSs to generate recommendations. Such an analysis enables us to establish existing trends in the corpus and create a hypothesis about propagated data bias in the event of recommendation. To achieve this, we use the `GR-lang` dataset to quantify the relevant labels regarding these traits. The proportions of these characteristics is further checked on a genre level, under the premise that provider groups have better resemblance in specific categories. Similarity of the user profiles' language and translation status distribution to the book catalogue is computed using JSD, explained in Section 3.3.1.

Performance of Recommender Algorithms. We assess the ability of various recommender algorithms, described in Section 3.2.1, to generate equitable recommendations with respect to the outlined provider groups, while they still remain relevant to the specific user. Training and testing of the algorithms is done using $\texttt{GR-lang}^s$. To get a grasp of the behaviour exhibited by the recommender algorithms, we employ both user relevancy and fairness metrics on the top-50 recommendations in each list. The utilised metrics are defined in Section 3.3.2 and Section 3.3.3 respectively. Statistical significance is performed using the Wilcoxon Signed-Rank Test, as described in Section 3.3.4, in order to compare metric scores across a variety of algorithms.

Performance of Reranking Strategy. We aim to mitigate possible inequities found between provider groups, while maintaining agreeable relevance scores for the target users. To accomplish this, CP, introduced in Section 3.2.2, is employed as a reranking strategy operating on the recommendation lists of each user, generated by the algorithms evaluated in the previous experiment. Consequently, fairness and relevance metrics presented in Section 3.3 are recomputed for the reranked lists. In this case, we assess the top-50 as well as the top-10 reranked results for each user, in order to grasp changes at the top level of the item lists. The significance of the results is determined using Wilcoxon Signed-Rank Test following Section 3.3.4, centred on the differences to the initial metrics for each algorithm. We continue the discussion by analysing the trade-off between the two types of metrics but also about any changes that transpired at provider group level.

3.4.2. Implementation

In terms of deploying the algorithms enumerated in Section 3.2.1, we use the Elliot framework [73] to carry out the tuning, train-test routine and part of the evaluation. All of the relevance metrics, as well as REO are computed through this pipeline. Consequently, the rest of the metrics are implemented as described in their reference material. As far as reranking is concerned, we use the implementation[3] of the CP reranking strategy as defined by Ungruh et al. [41]. To facilitate the reproducibility of our findings, a code repository is publicly available for inspection[4].

We carry out a temporal split of the data, where interactions of the users are withheld for testing based on the relevant timestamp. This order-aware method allows the RS to predict a user's future based on their past, as we are respecting each individual timeline of interactions [74, 20]. As such, $\texttt{GR-lang}^s$ is partitioned in an 80-20 train-test split where each user has their latest 20% of ratings in the latter dataset.

The hyper-parameter tuning of the chosen algorithms is heavily based on the experiments defined by Anelli et al. [68], relying on their hyper-parameter value ranges for most algorithms. In the cases where the setup for any of the outlined algorithms were missing, we turned to the respective papers pertaining to their implementation for setup. Tree of Parzen Estimators is used as the method to choose hyper-parameters. The validation set is composed of the 20% most recent ratings of the test set, while the entries are excluded from the train set for this step. The optimization metric for this process is nDCG@10.

[3] https://github.com/rUngruh/mitigatingPopularityBiasInMRS

[4] https://github.com/raresboza/thesis_lost_in_translation

4

Results

This chapter contains the findings of the experiments described in Section 3.4. The abbreviations for the provider groups and metrics can be consulted in Section 3.1.2 and Section 3.3 respectively.

4.1. Experiment 1: Language & Translation Presence in Source Data

The scope of the experiment is two-fold, focusing on the language traits of the items in the book catalogue under scrutiny in Section 4.1.1 and contrasting the user activity to said catalogue based on the same characteristics in Section 4.1.2.

4.1.1. Language Inference Book Catalogue

To understand the language characteristics of our curated dataset, `GR-lang`, we turn to an analysis of the book catalogue. We examine the presence of varied books from the perspective of original language and accessibility in English. To achieve this, we quantify the spread of each language types established in Section 3.1.2 amongst the existing books. Furthermore, we juxtapose our findings with several genre types in order to visualise how language is represented in each category. An overview of the book catalogue is shown in Figure 4.1. The filtering of language and translation characteristics per a sample of genres is outlined in Figure 4.2.

We highlight the strong differences between the presence of each category in the `GR-lang` book corpus. The majority of the items (54.01%) could not have a language label attributed to them. The highest appearance between the items that were identified to have related language information are books originally written in English with 31.91%. Books coming from other languages account for 12.38% of the corpus. Specifically, 9.18% of the books are written in a language other than English but with an available translation, while 3.20% cannot be found in English. A small percentage of books (1.70%) were found to have editions in English and at least one more language, although the original one could not be identified.

In addition to the catalogue analysis of `GR-lang`, we inspect the presence of language characteristics and translation availability to English in specific genres. For this, we group the books according to their respective genres, classifying them by the same language criteria. We motivate this decision by rationalizing that in certain genres there should be greater parity between English and other languages, as well as demand for translations. For instance, in non-fiction we should account for biographies of historical personalities that should cover the whole world, as well as works of science from foreign academic centres. The incentive for looking into children books is guided by the intrinsic need of every culture to prepare its young readers, introducing them to elements of the surrounding society and boosting their language comprehension, endeavours which cannot be completely fulfilled by non-native literature. Nonetheless, critically acclaimed titles are often translated and present in the commonly read books of people of different cultures and this is no different for children books. We chose the romance genre for its growing popularity, speculating that it could offer the chance to plenty of foreign authors to break into the English-speaking market.

Aggregating the book catalogue into genres reveals various differences in their language characteristics. For all of the classified genres, the trend of the English originals being larger than the rest of the categories remains consistent with the general distribution of `GR-lang`, in the cases where language information is available. We emphasise on romance as being the most inconsistent genre in comparison to the rest of the

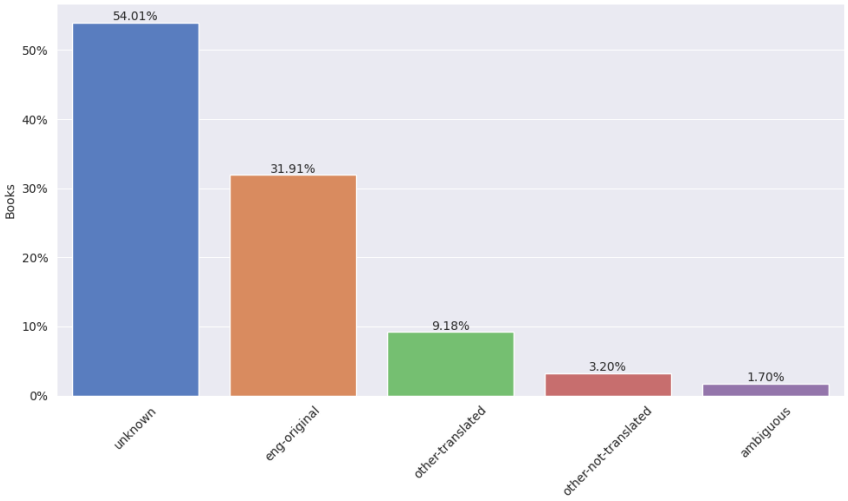


Figure 4.1: Language distribution of items in `GR-lang`.

distributions. A majority of its items (71.74%) do not have any language information. Another discrepancy is centred on English originals, which consist of only 18.53% of the items in the romance genre. This presents a stark contrast with nonfiction and children books, where the unknown entries account for 33.57% and 29.46% of the category respectively. The nonfiction and children genres have a higher proportion of English originals than the overarching distribution, closer to 49%. The similarity between these two genres persists when shifting focus on other languages, where they constitute ~15% in nonfiction and ~19% in children oriented books. Still, 11.37% of the items in the nonfiction book catalogue are foreign works with availability of English editions, while the value is considerably higher for children' books, reaching 18.02%. Both books without translations into English and those without discernable language of origin have limited presence in all genres, ranging from ~1% to ~4%.

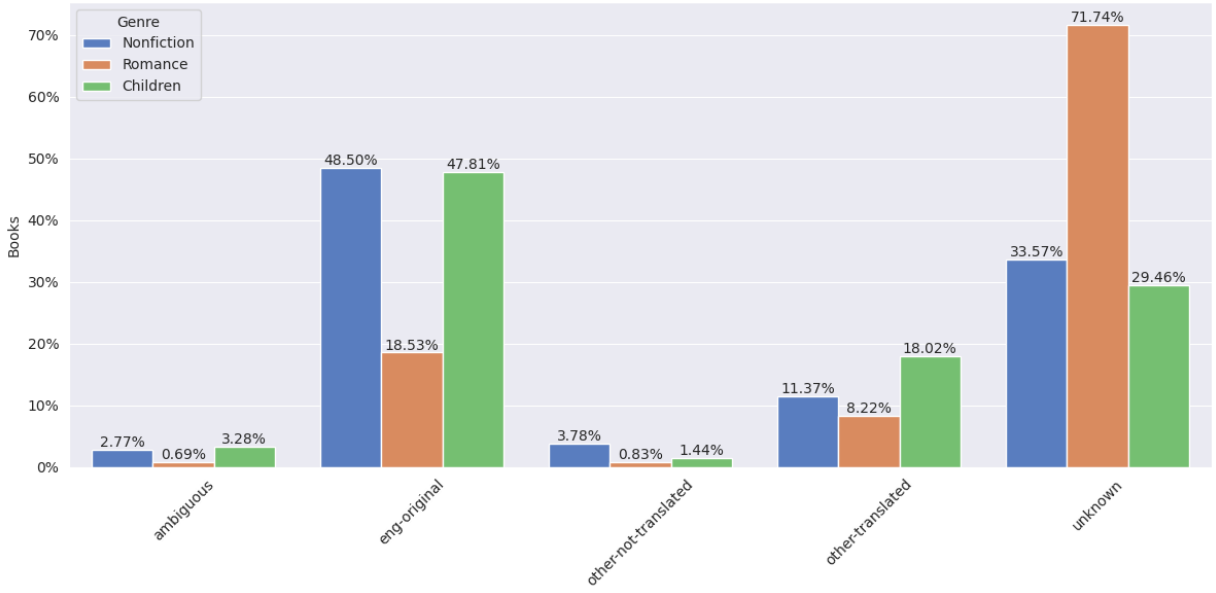


Figure 4.2: Language distribution in `GR-lang`, grouped by their genre. Only contains nonfiction, romance and books catered to children.

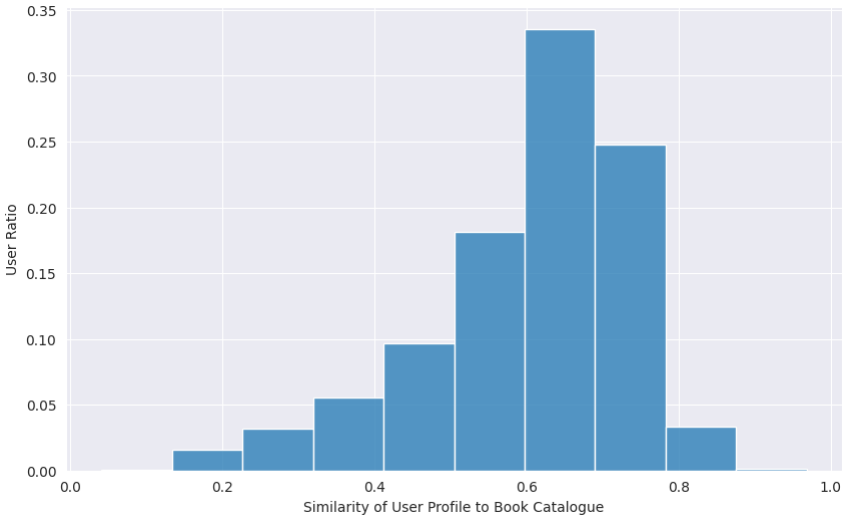


Figure 4.3: Closeness of user profiles to the item catalogue language distribution in `GR-lang`.

4.1.2. Catalogue and User Profiles Similarity

To gain insight into how the user base tracked in `GR-lang` interacts with the items in our book catalogue, we opt for a comparison between the distribution of their user profiles and the base distribution featured in Figure 4.1. We want to understand how homogenously the users interact with the book corpus from the viewpoint of language and translation availability to English. For this, we quantify individual profile distribution considering the established language types and report the JSD value to assess the extent of its deviation. Given that JSD is a continuous variable, we visualise user profiles into 10 bins along its domain. A summary of the similarity of user activity to the book catalogue with regard to language characteristics is illustrated in Figure 4.3.

The computed user profile bins emphasise a noticeable deviation from the language type distribution of the available items. Almost 80% of the users recorded in `GR-lang` reach a JSD value of at least 0.5, highlighting trend of deviation from the available catalogue. When excluding mild dissimilarity, 33.57% of the users are placed in the 0.6-0.7 bin while 24.76% are situated in the 0.7-0.8 range, together still comprising a majority. Complete deviation and very high similarity both reach values lower than 5%. Concerning users with past activity relatively alike to that of the book catalogue distribution, we quantify 5.52% with a JSD value between 0.3 and 0.4 and 9.67% in the next bin until 0.5.

4.2. Experiment 2: Performance of Recommender Algorithms

We undertake the task of evaluating the performance of recommender algorithms from the perspective of the relevancy of recommendations to the user, but also group provider fairness, assessed by the metrics established in Section 3.3.3. Our findings are based on the generation of the top-50 recommendations using each algorithm. The focus lies on evaluating the ability of the inspected algorithms to treat the providers groups categorised from a language standpoint with equity. In this endeavour, we observe this trait in conjunction with whether the users are presented with items of interest. This is what prompts the analysis of both relevancy and fairness metrics. We summarise these results in Table 4.1. Figure 4.4 portrays the significant differences in the global metric values across every algorithm, while Figure 4.6 illustrates significant differences between fairness metrics group-wise.

With respect to ranking quality of the recommendation lists, we examine the appropriate beyond-accuracy metrics. We exclude the results of Random and ItemAutoRec in the immediate overview that follows, due to the scores achieved by them being considerably lower than their peers. The nDCG values for the recommendations created by the algorithms under scrutiny are between 0.0867 and 0.1845. These values underline the that most relevant items are placed in the lower positions of the generated lists. This is further reinforced by the MAP scores that are within the range of 0.0558 and 0.1118, indicating that there are few useful items retrieved in the recommendations, which are also ranked low. MRR values for the inspected algorithms vary in the 0.1863-0.3780 interval, emphasising that the first relevant item to the user

is encountered in the lower end of the generated lists.

It is worth noting that the conducted Wilcoxon signed-rank testing reveals that every improvement of beyond-accuracy metrics achieved by an algorithm over another, including that of ItemAutoRec over Random, is statistically significant. Taking into account the presented overview, the best performing algorithms, from the standpoint of any beyond-accuracy metrics are: Slim, ItemKNN and MultiVAE. Figure 4.5 presents the spread of the nDCG score for each user. All three algorithms present a wide interquartile range, with ItemKNN presenting a tighter core and a lower median. This shows that user performance varies heavily even within the distribution of each algorithm. There are many outliers at the upper tail of each algorithm, suggesting that a small portion of users do get served correctly.

	Random	ItemKNN	PMF	BPRMF	MultiVAE	Slim	MF2020	ItemAutoRec
nDCG	0.0009	0.1712	0.0867	0.1060	0.1474	0.1845	0.1216	0.0017
MAP	0.0006	0.1018	0.0558	0.0646	0.0818	0.1118	0.0737	0.0013
MRR	0.0026	0.3359	0.1863	0.2089	0.2759	0.3780	0.2357	0.0054
REO_{EO}	0.0011	0.1212	0.0870	0.0935	0.08567	0.1148	0.0858	0.0007
REO_{OT}	0.0012	0.1322	0.0610	0.0810	0.1132	0.1431	0.0981	0.0011
REO_{ONT}	0.0014	0.1691	0.0021	0.0517	0.0954	0.1381	0.0482	0.0000
REO_{AMBG}	0.0012	0.1612	0.1175	0.1437	0.1653	0.1847	0.1605	0.0032
REO_{UNKN}	0.0011	0.0581	0.0009	0.0078	0.0172	0.0313	0.0049	0.0012
$REO_{Overall}$	0.0839	0.3065	0.8613	0.5971	0.5012	0.4153	0.6535	0.8653
DE_{EO}	0.0011	−0.1519	−0.0915	−0.0994	−0.1704	−0.1746	−0.1686	0.0470
DE_{OT}	−0.0001	0.0772	−0.0275	−0.0277	0.0759	0.0844	0.0763	0.0529
DE_{ONT}	0.0001	−0.0028	−0.0106	−0.0086	−0.0047	−0.0057	−0.0077	−0.0108
DE_{AMBG}	0.0001	0.1807	0.2626	0.2641	0.2154	0.2159	0.2292	−0.0471
DE_{UNKN}	−0.0010	−0.1031	−0.1330	−0.1284	−0.1160	−0.1200	−0.1291	−0.0419
GC_{item}	0.1426	0.8827	0.9980	0.9858	0.9258	0.9370	0.9851	0.9991
GC_{group}	0.4637	0.5906	0.5815	0.5784	0.6151	0.6268	0.6318	0.5711
LRD_{EO}	−0.2748	0.0476	−0.0306	−0.0251	0.0695	0.0848	0.0715	−0.3593
LRD_{OT}	0.1583	−0.0003	0.1543	0.1417	0.0011	−0.0186	0.0086	0.0676
LRD_{ONT}	−0.0074	−0.0012	0.0143	0.0097	0.0026	0.0033	0.0076	0.0146
LRD_{AMBG}	0.3272	−0.0459	−0.1981	−0.1768	−0.0995	−0.0992	−0.1394	0.4172
LRD_{UNKN}	−0.2034	−0.0001	0.0601	0.0505	0.0264	0.0296	0.0517	−0.1400

Table 4.1: Metrics computed for the top-50 recommendations generated by the analysed recommender algorithms.

We turn to the results centred on provider group fairness exhibited in the top-50 recommendations generated by the chosen algorithms. The property of recommending items with specific language characteristics to the interested user-base is projected by the attributed group REO values. Zooming in on items with unknown language characteristics, the REO_{UNKN} scores show a great inability of all algorithms to include this type of item in lists where they would be desired by the user. REO_{ONT} displays a similar trend, where PMF, BPRMF and MF2020 struggle to correctly connect such books with their target audience. Once again, we choose to exclude Random and ItemAutoRec from the described ranges, as their values are close to 0. REO_{EO}, REO_{OT} range from 0.0858 to 0.1212, and from 0.0610 to 0.1431, respectively. Although similar in value, REO_{EO} is higher in the case of PMF, BPRMF whereas the others and lower in the rest. Furthermore, REO_{AMBG} shows the best performance out of all target provider groups, with scores in the 0.1175-0.1847 interval.

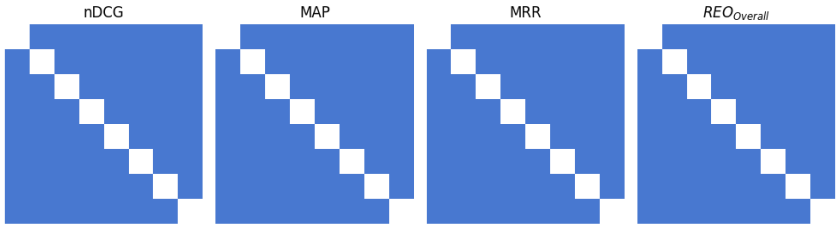


Figure 4.4: Wilcoxon signed-rank test with Bonferonni correction on performance metrics and REO$_{Overall}$ calculated on the top-50 recommendation lists generated by the chosen algorithms. $p < 0.01$ is indicated by the blue colour. The order of the recommender algorithms is the same as in Table 4.1

.

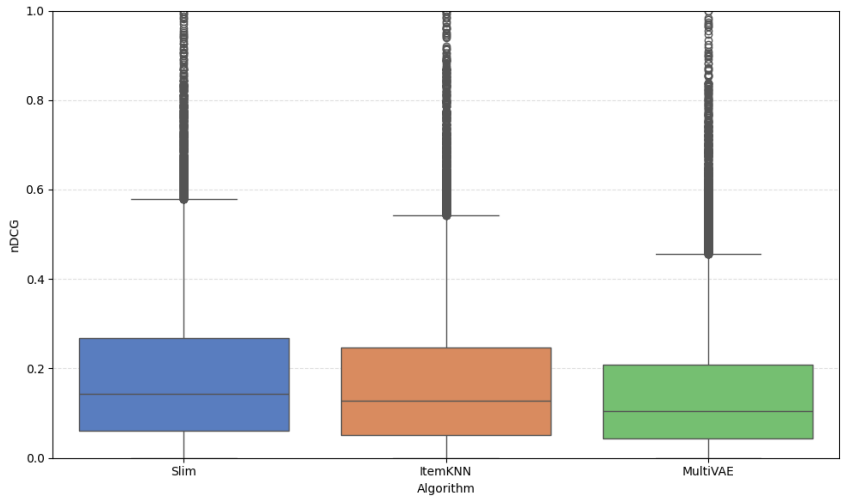


Figure 4.5: nDCG scores per user for the 3 best performing algorithms in terms of user relevancy, on top-50 recommendation lists.

.

We capture the exposure of each provider group and how it differs from their presence in the catalogue. DE$_{EO}$ and DE$_{UNKN}$ are predominantly negative, indicating disparity in exposure from their representation in the item corpus. These values range from -0.1746 to 0.0470 and from -0.1330 to -0.0010. Concentrating on DE$_{OT}$, we find negligible penalisation, when it comes to PMF (-0.0275) and BPRMF (-0.0277). The other algorithms lean towards a mild benefit in exposure to this class of items in the $[0.0529, 0.0844]$ interval. While DE$_{ONT}$ shows little deviation of exposure from expectations, DE$_{AMBG}$ has a large number of scores above 0.2. This conveys a strong increase given the limited presence of this class of items in the catalogue.

Scrutinising the visibility of the inspected provider groups in the recommendation lists in comparison to the appropriate user profiles, we find that the proportions generally hold. An exemption is made by the AMBG item group, where LRD$_{AMBG}$ is lower in most cases, denoting that this provider group is recommended much more than it is observed in the user profiles. Algorithms with better performance in user relevance metrics have a slightly positive LRD$_{EO}$ (0.0476-0.0848) and an LRD$_{OT}$ value close to 0, meaning that the former category has 4-8% lower proportions of recommendations, while the latter remains stable. PMF and BPRMF have a raise in LRD$_{OT}$ scores meaning these providers appear less in the recommendations than in the history of the respective users. In contrast, LRD$_{AMBG}$ achieves the lowest 2 scores (-0.1981 and -0.1768) of our trials, continuing the trend of featuring AMBG items with a higher ratio than that observed in the recommendations.

Irrespective of the effects on individual groups of providers, we seek to present an overview of the unfairness generated by these algorithms. Apart from the naive baseline, GC$_{group}$ shows a consistently lower score compared to GC$_{item}$. This emphasises that a reduced set of items garners the most exposure

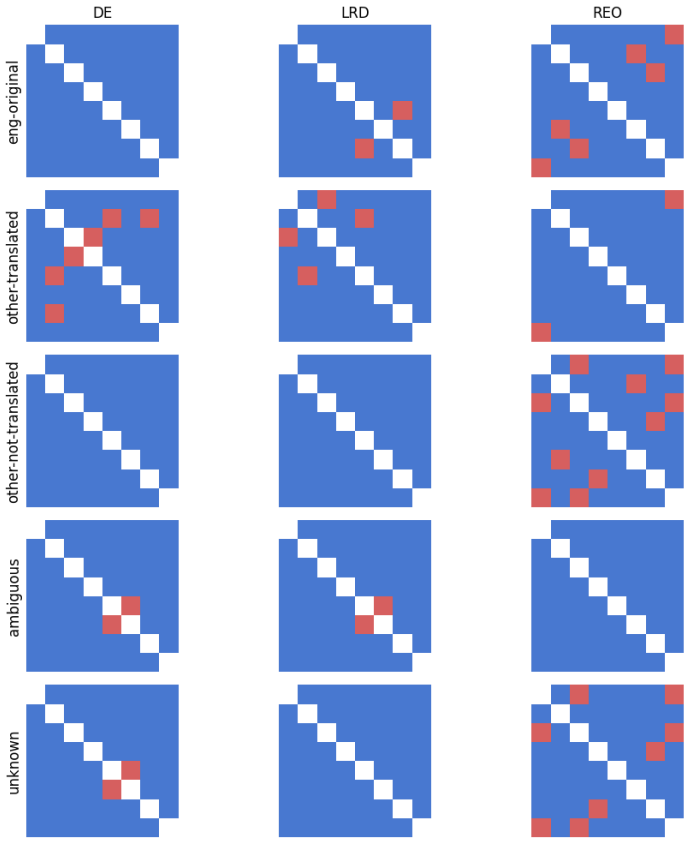


Figure 4.6: Wilcoxon signed-rank test with Bonferonni correction on group fair metrics (DE, LRD, REO) calculated on the top-50 recommendation lists generated by the chosen algorithms. $p < 0.01$ is indicated by the blue colour. The order of the recommender algorithms is the same as in Table 4.1

.

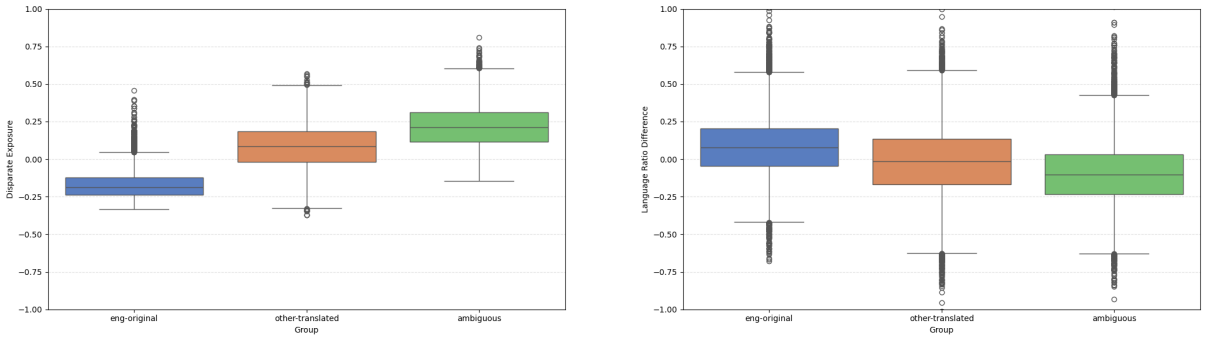
across recommendation lists, even though they are not necessarily part of the same provider group. Nevertheless, with values between 0.4637 and 0.6318 there is still a skew toward favouring some provider characteristics. In contrast, $\text{REO}_{\text{Overall}}$, which tracks the coefficient of variance between the successful recommendation rates per group, indicates higher bias in worse performing algorithms (0.5971-0.8653) in terms of nDCG scores. With regard to the best performing algorithms, ItemKNN (0.3065) has the lowest bias, followed by Slim (0.4153) and MultiVAE (0.5012). This denotes that there is moderate dispersion between the analysed provider groups in the recommendations of the aforementioned algorithms. The probability of recommending UNKN items to interested users lags behind the most as the value REO_{UNKN} is much lower than the rest.

The differences in provider group fairness metrics are observed to be overall significant, as shown in Figure 4.6. Cases where several algorithms achieve scores that are not significantly different are found primarily when computing REO_{EO}, REO_{ONT} and REO_{UNKN}, although they still represent a minority. We highlight that Slim and ItemKNN do not present significant differences in the scores of REO_{EO} and REO_{ONT}, meaning that the successful recommendation rate of items belonging to the respective groups does not improve for one algorithm over another. Slim records no significant differences over MultiVAE in DE_{UNKN}, DE_{AMBG}, and LRD_{AMBG} despite having a significant difference in user relevancy, also in terms of the magnitude of nDCG. Therefore, the difference in this performance can be pinpointed to the treatment of other provider groups given visibility and ranking. The comparison between BPRMF and PMF reveals that their DE_{OT} scores are not statistically different, thus maintaining a similar, balanced exposure with a slight 2.7% decrease with respect to the group's presence in the book corpus.

To grasp the treatment of individual groups within the generated lists, we inspect the distributions of DE and LRD for EO, OT and AMBG in the case of Slim, available in Figure 4.7. We identify large

variance in all classes of LRD scores between the user base. However, LRD_{EO} has the tightest core, with a shorter lower tail and outliers that do not reach maximum negative values. The inclusion of EO items in the recommendations is higher, but their exposure remains slightly low with respect to to their representation. This is reinforced by the spread of DE_{EO}, which also demonstrates that several outlier users are subject to increased exposure to this group. In comparison, the AMBG group receives more visibility but also disproportionately higher exposure. The LRD_{OT} scores emphasise that this class of items varies in representation for each individual user. However, through DE_{OT}, it is clear that it ranks higher than is expected of its coverage in the item catalogue.

ItemAutoRec and the naive baseline Random achieve similar poor performance in terms of nDCG, MAP and MRR. Nevertheless, the two exhibit various differences at the composition level of their respective recommendations to the user base. By randomly sampling out of the available item catalogue, the naive baseline includes more EO and UNKN items in the generated lists, at the expense of OT and AMBG, as observed through the LRD scores. The exposure, as described by DE, remains consistent across every provider group, being balanced to their representation in the data. In contrast, for ItemAutoRec, the values for LRD_{EO} and LRD_{OT} are even more accentuated, and we notice an exposure change of ~5% for most categories. The GC_{item} score of 0.9991 reflects that this algorithm focuses on the recommendation of few unique items, while the naive baseline (0.1426) has generated lists benefiting of diversity.



(a) Disparate exposure comparison.

(b) Language ratio difference comparison.

Figure 4.7: Comparisons of group providers w.r.t. fairness in the top-50 user recommendations generated by Slim.

We summarise the results of the fairness and user relevance metrics computed for the top-10 recommendations in Table 4.2. The top 3 algorithms maintain the original order from the perspective of relevance to the user, although the scores are slightly reduced. This effect projects itself further to the REO values, as a reduction in the ability to connect items with interested users records lower probabilities of these provider groups to be matched with their audience. Nevertheless, the groups are affected disproportionately as the values for $REO_{Overall}$ increase with around 0.02-0.1, emphasising slightly more fluctuation around the average success rate. We highlight that this comes from an even worse performance for UNKN to be recommended correctly and a REO_{ONT} that has either similar probability or worse to EO, OT and AMBG depending on the top-50 recommendations of the algorithm.

In terms of visibility of the providers, given that there are 5 analysed groups, any increase in representation in the list will alter the tendencies by a large margin. By looking at Slim, ItemKNN and MultiVAE, we observe a consistency in the trends fostered within the larger recommendation lists. This results in proportional representation to the user profiles in the case of OT, ONT and UNKN, with a heightened presence for AMBG and a mild decrease in the case of EO as shown through LRD scores. DE_{EO} is below -0.12, with values close to 0 for DE_{ONT} and DE_{AMBG} being greater than 0.16. We remark that that the deviations of exposure of the provider groups from the expectations given their representation in the catalogue maintain similar values to those in the top-50 lists.

Other algorithms manifest a different treatment of the provider groups when narrowing down to top-10 recommendations. PMF and BPRMF attain a DE_{EO} score closer to 0, meaning that their exposure is balanced to their representation in the item catalogue. The OT is ranked much lower, lessening its

	Random	ItemKNN	PMF	BPRMF	MultiVAE	Slim	MF2020	ItemAutoRec
nDCG	0.0006	0.1530	0.0786	0.0871	0.1191	0.1734	0.1028	0.0019
MAP	0.0006	0.1590	0.0821	0.0884	0.1193	0.1827	0.1050	0.0017
MRR	0.0016	0.3222	0.1723	0.1928	0.2608	0.3659	0.2202	0.0046
REO_{EO}	0.0003	0.0570	0.0443	0.0400	0.0319	0.0549	0.0331	0.0003
REO_{OT}	0.0002	0.0451	0.0136	0.0152	0.0370	0.0529	0.0310	0.0002
REO_{ONT}	0.0000	0.0592	0.0000	0.0148	0.0269	0.0334	0.0127	0.0000
REO_{AMBG}	0.0002	0.0523	0.0405	0.0539	0.0571	0.0695	0.0514	0.0027
REO_{UNKN}	0.0002	0.0178	0.0001	0.0018	0.0046	0.0055	0.0010	0.0002
$REO_{Overall}$	0.5068	0.3252	0.9771	0.7542	0.5376	0.5103	0.6744	1.5504
DE_{EO}	0.0019	−0.1266	0.0004	−0.0394	−0.1686	−0.1633	−0.1537	0.1053
DE_{OT}	−0.0008	0.0712	−0.1237	−0.1554	0.0665	0.0726	0.0772	0.0441
DE_{ONT}	0.0001	−0.0026	−0.0107	−0.0091	−0.0044	−0.0075	−0.0086	−0.0108
DE_{AMBG}	−0.0005	0.1624	0.2674	0.3336	0.2227	0.2257	0.2159	−0.0627
DE_{UNKN}	−0.0007	−0.1045	−0.1336	−0.1298	−0.1162	−0.1276	−0.1307	−0.0758
GC_{item}	0.2993	0.9234	0.9992	0.9935	0.9473	0.9751	0.9914	0.9998
GC_{group}	0.4634	0.5760	0.5814	0.6242	0.6091	0.6255	0.6278	0.6200
LRD_{EO}	−0.2768	0.0059	−0.2022	−0.1271	0.0676	0.0728	0.0258	−0.4846
LRD_{OT}	0.1591	0.0047	0.3786	0.4467	0.0237	−0.0005	0.0159	0.1591
LRD_{ONT}	−0.0072	−0.0018	0.0143	0.0107	0.0018	0.0070	0.0093	0.0146
LRD_{AMBG}	0.3293	−0.0119	−0.2521	−0.3839	−0.1204	−0.1271	−0.1062	0.4225
LRD_{UNKN}	−0.2044	0.0031	0.0615	0.0538	0.0273	0.0477	0.0552	−0.1116

Table 4.2: Metrics computed for the top-10 recommendations generated by the analysed recommender algorithms.

exposure further by 10% compared to the DE_{OT} score recorded for the top-50 recommendation. LRD_{EO}, LRD_{AMBG} emphasise the amplification of these providers in the generated lists while LRD_{OT} underlines an even greater reduction in the presence of this group in the recommendations compared to past user activity. Given these changes in magnitude that happen to the tracked fairness metrics, we note the tendency of PMF and BPRMF to propagate their disproportionate treatment even more at the top of the recommendations.

4.3. Experiment 3: Performance of Reranking Strategy

To understand the extent of which existing mitigation approaches can reduce imbalances in fairness found in the recommendation listing, we delve into the effect of reranking the recommendation lists evaluated in Section 4.2 using CP. Besides an analysis of fairness for the established provider groups centred around the language characteristics they represent, changes in scores of relevance of the recommendations are also recorded. Joint assessment of the two categories of metrics allows for observing trade-offs between the capability to diversify items presented to the user, and the utility of the recommendations. We report our findings for the reranking of the top-50 generated recommendations in Table 4.3, containing the metrics for which changes can be registered. Similarly, Table 4.4 presents the scores computed for the top-10 lists, this time for every metric, as new items can be introduced. To facilitate a clear comparison between the initial recommendations and the reranked counterparts, we highlight the values of each metric for which statically significant differences have been recorded.

The scores of the relevancy metrics in both the top-50 and top-10 scenarios reveal statistically significant differences in relation to their non-reranked equivalents. Out of the evaluated algorithms, only the worst-performing ones, the naive baseline Random and ItemAutoRec improve, although they remain far behind the others. The order of the algorithms, ranked by nDCG does not change for either size of the list. We note that differences with regard to this metrics are slim, within approximately 0.0100, as they would be hard to perceive in practice by end-users.

We remark on the impact of the newly introduced items in the top-10 recommendations on the interest provider groups in terms of fairness. The scores of REO change significantly for some of the algorithms, which for the most part comprise both the best and worst performing methods in terms of accuracy. In the case of Slim, we observe an increase in the scores of REO_{EO}, REO_{ONT} and a decrease of REO_{AMBG}. This emphasises that English originals as well as translated books are reaching the users that like them with more success, while the opposite happens for books that could not be properly identified as belonging to these categories. $\text{REO}_{\text{Overall}}$ is distincly smaller than its non-reranked value, emphasising smaller discrepancy in correct recommendation between the provider categories (0.4324 vs 0.5103), although it is not statistically significant. For ItemKNN and MultiVAE, all significant differences (besides REO_{UNKN} for MultiVAE) show smaller values in the reranked scores. The MultiVAE $\text{REO}_{\text{Overall}}$ indicates lower bias (0.5007 vs 0.5376), albeit a joint analysis of the respective group values shows that worsened performance for most providers is the reason for reaching better equity in this aspect. In terms of LRD scores, the effects of reranking predominantly push the values toward 0 which indicates equal ratios of representation as in the user histories. LRD_{AMBG} continues to show an increase in the amount of items recommended that belong to this group, having negative values, while LRD_{EO} signals an up to 7% reduction in English originals in comparison to their presence in the user histories for the algorithms with the highest user relevancy scores.

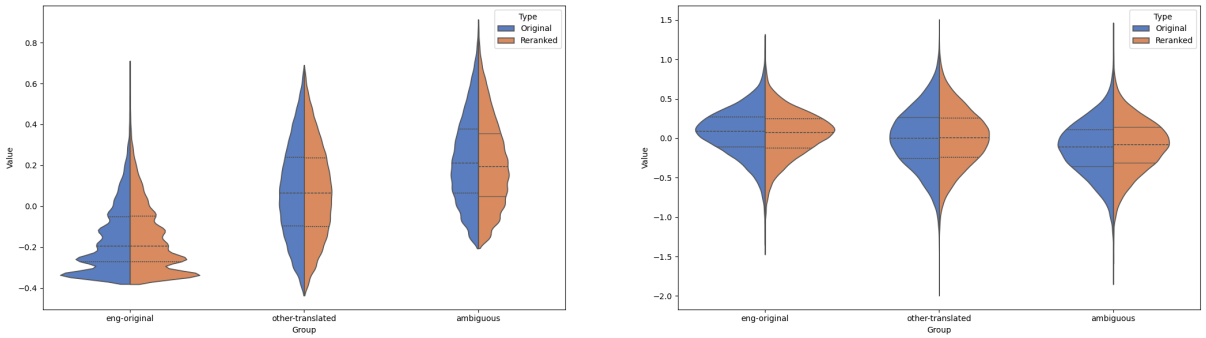
The exposure of the item groups under scrutiny are shown to display statistical differences in relation to their initial scores after the reranking process. In the case of the top-50 recommendations, the changes in magnitude of the values remain small and the trend is inconsistent. PMF and BPRMF present a slight decrease in exposure of EO compared to the expectancy resulting from this group's representation, while MultiVAE, Slim and MF2020 record a slight increase. The DE_{ONT} scores are getting closer to 0, which brings the exposure closer to the expected amount given by the group's representation in the corpus. Inspecting DE_{AMBG} reveals that the disparity is reduced in all cases but for MF2020 and the naive baseline, in which it increases. This trend of marginal differences in the overall value without a conclusive direction manifests itself across most provider groups. The scores of DE_{ONT} and DE_{UNKN} for PMF are found to have statistically significant differences between the values for specific users, while the overall value remains identical. This indicates that reranking has mixed effects on the exposure of each provider group in the recommendation list of users.

Shifting the viewpoint to the top-10, the scores for DE, irrespective of the class of items, mainly indicate a slight improvement towards balance of exposure with respects to the groups' proportions in the catalogue. The magnitude of the shift in value is mostly within 0.100. We note exceptions for BPRMF where for the negative DE_{EO} and positive DE_{AMBG} become more severe. PMF also reports an exacerbation of the disproportionate exposure given to AMBG after reranking. The naive baseline is the only case where the exposure of groups gets considerably less balanced in comparison to the previous values in the initial list.

To visualise how the distribution of the fairness metrics evolved after reranking we turn to Figure 4.8, which presents a comparison between the original and reranked top-10 recommendation lists generated using Slim with respect to LRD and DE. We take note that no drastic changes can be observed with regard to the shape of the distributions. The peaks of the DE_{EO} distribution suggest that there are clusters of users where the exposure to this provider group is negated. In the cases where statistically significant differences were recorded, both the LRD and DE values shift towards 0, highlighting proportional visibility to the user profiles of the provider groups and closer exposure to their level of representation in the catalogue.

With respect to general fairness of recommendation, in the reranking of top-50 recommendations, the changes in values are negligible. We report that ItemKNN, MultiVAE, and Slim all have slight reductions in the bias of recommending either solely a small set of items defined by GC_{item} or out of one provider group, emphasised by GC_{group}. In the case of the top-10 recommendations, the GC_{item} and GC_{group} scores for most of the algorithms present a trend of slight improvement in the bias. This relates to better diversity in individual items and provider groups compared to the initial recommendation lists. The reranked version of

ItemKNN improves from both perspectives (0.9087 as opposed to 0.9234 for GC_{item} and 0.5722 from 0.5760 for GC_{group}). Similarly, BPRMF has a lower GC_{item} (0.9840 from 0.9935) and GC_{group} (0.5852 from 0.6242) respectively. Slim and MF2020 benefit from similar decreasing scores, while PMF, MultiVAE and Random become more biased from both standpoints. Evaluating these metrics for ItemAutoRec reveals that its severe bias for items does not change, while the group selection improves. As with to the larger size of the list, the changes in scores are small unless you consider the worse performing algorithms in terms of serving relevant items to users (Random and ItemAutoRec).



(a) Disparate exposure comparison.

(b) Language ratio difference comparison.

Figure 4.8: Comparisons of group providers w.r.t. LRD and DE in the initial and reranked top-10 user recommendations generated by Slim.

	Random	ItemKNN	PMF	BPRMF	MultiVAE	Slim	MF2020	ItemAutoRec
nDCG	**0.0012**	**0.1668**	**0.0861**	**0.1051**	**0.1447**	**0.1816**	**0.1207**	**0.0023**
MAP	**0.0010**	**0.0993**	**0.0555**	**0.0637**	**0.0793**	**0.1110**	**0.0733**	**0.0017**
MRR	**0.0048**	**0.3125**	**0.1803**	**0.1986**	**0.2631**	**0.3604**	**0.2277**	**0.0095**
DE_{EO}	**−0.0168**	−0.1521	**−0.0932**	**−0.1106**	**−0.1676**	**−0.1731**	**−0.1681**	**0.0211**
DE_{OT}	**0.0156**	**0.0775**	−0.0274	**−0.0234**	**0.0752**	0.0842	**0.0750**	**0.0579**
DE_{ONT}	**−0.0007**	**−0.0028**	**−0.0106**	**−0.0081**	**−0.0044**	**−0.0052**	**−0.0074**	−0.0108
DE_{AMBG}	**0.0168**	**0.1799**	**0.2642**	**0.2696**	**0.2117**	**0.2123**	0.2295	**−0.0291**
DE_{UNKN}	**−0.0147**	**−0.1026**	**−0.1330**	**−0.1275**	**−0.1148**	**−0.1182**	**−0.1289**	**−0.0391**
GC_{item}	0.1981	0.8793	0.9979	0.9840	0.9190	0.9310	0.9850	0.9990
GC_{group}	0.4781	0.5902	0.5830	0.5852	0.6117	0.6233	0.6303	0.5618

Table 4.3: Metrics computed for the top-50 recommendations generated by the analysed recommender algorithms and reranked with CP. Bolded scores (excluding GC) denote significant difference from the the initial recommendation list.

	Random	ItemKNN	PMF	BPRMF	MultiVAE	Slim	MF2020	ItemAutoRec
nDCG	**0.0014**	**0.1455**	**0.0776**	**0.0844**	**0.1130**	**0.1690**	**0.1013**	**0.0031**
MAP	**0.0014**	**0.1490**	**0.0807**	**0.0846**	**0.1116**	**0.1756**	**0.1029**	**0.0034**
MRR	**0.0040**	**0.2986**	**0.1663**	**0.1825**	**0.2475**	**0.3485**	**0.2121**	**0.0089**
REO_{EO}	0.0004	0.0565	0.0443	0.0400	**0.0313**	**0.0562**	0.0332	0.0003
REO_{OT}	**0.0006**	**0.0440**	0.0136	0.0150	**0.0353**	0.0527	0.0310	**0.0005**
REO_{ONT}	0.0004	0.0551	0.0000	**0.0193**	0.0286	**0.0430**	0.0110	0.0000
REO_{AMBG}	**0.0007**	**0.0512**	0.0405	**0.0526**	**0.0539**	**0.0675**	0.0514	**0.0031**
REO_{UNKN}	0.0002	0.0181	0.0001	0.0023	**0.0055**	**0.0094**	0.0010	**0.0005**
$REO_{Overall}$	**0.4463**	0.3139	0.9760	**0.6993**	**0.5007**	0.4324	0.6946	**1.3091**
DE_{EO}	**−0.0834**	**−0.1254**	**−0.0045**	**−0.0697**	**−0.1554**	**−0.1554**	**−0.1524**	**−0.0809**
DE_{OT}	**0.0714**	0.0719	−0.1233	**−0.1420**	**0.0640**	0.0708	**0.0732**	**0.1021**
DE_{ONT}	**−0.0039**	**−0.0020**	**−0.0106**	**−0.0070**	**−0.0031**	**−0.0052**	**−0.0075**	−0.0108
DE_{AMBG}	**0.0787**	**0.1574**	**0.2717**	**0.3443**	**0.2053**	**0.2097**	0.2166	**0.0406**
DE_{UNKN}	**−0.0627**	**−0.1019**	**−0.1334**	**−0.1257**	**−0.1107**	**−0.1199**	**−0.1298**	**−0.0511**
GC_{item}	0.6648	0.9087	0.9992	0.9884	0.9229	0.9576	0.9911	0.9998
GC_{group}	0.5278	0.5722	0.5831	0.6156	0.5939	0.6096	0.6225	0.5610
LRD_{EO}	**−0.1048**	**−0.0032**	−0.2024	**−0.1335**	**0.0411**	**0.0529**	0.0261	**−0.0121**
LRD_{OT}	**0.0154**	0.0063	0.3786	**0.4425**	**0.0272**	**0.0064**	**0.0173**	**−0.1111**
LRD_{ONT}	**0.0008**	**−0.0036**	**0.0141**	**0.0068**	**−0.0006**	**0.0029**	**0.0071**	0.0146
LRD_{AMBG}	**0.1683**	**0.0028**	**−0.2515**	**−0.3617**	**−0.0834**	**−0.0947**	**−0.1044**	**0.2501**
LRD_{UNKN}	**−0.0796**	**−0.0023**	**0.0611**	**0.0460**	**0.0157**	**0.0324**	**0.0538**	**−0.1415**

Table 4.4: Metrics computed for the top-10 recommendations generated by the analysed recommender algorithms and reranked with CP. Bolded scores (excluding GC) denote significant difference from the the initial recommendation list.

Part III

Closure

5 Discussion

In this chapter we reflect on the results obtained by carrying out the defined experiments. We address the research questions that guide our empirical study, and provide insight on the design and evaluation of RSs from the perspective of fairness of the involved provider groups. Notes on the limitations of this paper, as well as proposals for future directions are also documented.

5.1. Prevalence of English Books in the Catalogue

Concerning **RQ1**, we take note of the distribution of language types and translation availability to English in the book catalogue that is the basis for our study. Identified foreign literature is scarce, especially when it comes to manuscripts without translation in English. Books originally written in English heavily dominate the part of the catalogue that has retrievable language information.

We investigate how the language representation and translation availability is distributed in romance, non-fiction and books catering to children. Romance was chosen as a popular genre across multiple cultures, while children require reading material representative of their surroundings. History studies and biographies pertaining to nonfiction tend to be first covered in the language of origin of the country of person in question. The outlook we provide on these genres generally supports the same observations made on the whole catalogue, despite the rate of language identification varying drastically, specifically in the case of romance books. Consequently, in genres where we theorised a better balance between English originals and foreign items, this did not manifest in our analysis. There was an increase in the percentage of books with an English translation in areas where more language information was available, although with a strong discrepancy in numbers to native literature. Untranslated foreign literature, has a marginal increase in the case of nonfiction, but plummets in representation in the other genres.

With regard to the consumer's activity, we highlight that the activity of a strong majority of the user population deviates from the distribution of the book catalogue. Despite user profiles showing this tendency, we note the thin tail indicating complete divergence from the item distribution. This shows that the number of users whose reading patterns cannot be satisfied by the available items from the perspective of their identified language characteristics is marginal.

5.2. Recommendations Propagate Books Available in English

When examining the effects of RSs on the proliferation of unfairness with respect to language of writing and translation availability, we emphasize a distinct presence of books accessible in English in the generated recommendations. Furthermore, books with editions in English benefit from much higher visibility and exposure to their untranslated counterparts.

We tackle the viewpoint established by **RQ2** of how RSs include translated foreign items in the generated lists when these compete with English originals. We find that translations are treated advantageously by a subset of the algorithms. Concerning algorithms that reach better scores in terms of the relevance of their top-50 recommendation lists, translations rank higher than their size of total items would suggest. The opposite is true for English originals, which fail to garner exposure proportional to their numbers in the user catalogue, but also have a slight reduction in presence when compared to user histories. Furthermore, translated items are more likely to reach their audience through these generated recommendations.

However, algorithms with moderate efficacy reveal significant contrast between the fairness treatment of the outlined provider groups, since English items gain visibility and reach interested users better than other groups, while still lacking proportional exposure to their catalogue representation. In this scenario, books with an available translation are subject to a reduced presence in the user profiles.

A large part of the user activity is centred on the minority group of ambiguous items, with editions in both English and other foreign languages, but without a clear tie to which one constitutes the source material. This item class is treated advantageously by most of the tested algorithms, exhibiting increased visibility and being ranked close to the top in the recommendation lists. Due to the disproportional imbalance in how the aforementioned items are consumed and presented, any claim made about the unfair recommendation tendencies towards either English originals or translated items would be unsubstantiated. Consequently, the identification of these manuscripts could heavily skew perception of exposure, representation, and correctness in recommendation of English originals and translated manuscripts.

Upon combining all the provider groups which share the property of books having editions in English (EO, OT, AMBG), we highlight the fact that most recommendations are of items accessible in English. This is derived from the results obtained from the Language Ratio Difference metric. Namely, we highlight a balanced LRD_{OT}, the mild penalty found in LRD_{EO} and the advantage flagged by LRD_{AMBG}. This is a consequence of English-translated literature and English originals making up over half of the ratings in the user profiles. As such, these scores emphasise that this majority is amplified even further in the recommendations.

In terms of foreign books without translation, the characteristics of this item group remain balanced according to its size in the user profiles. Its rate of recommendation to users that are interested varies greatly depending on the evaluated algorithm. In contrast to items of foreign origin, which present increased exposure to the expectancy given the representation in the catalogue, this group does not benefit from this treatment. By referencing the user activity relating to these two provider groups, we outline that a balanced LRD_{OT} results in a much larger presence of this group in the recommendations, than the LRD_{ONT} close to 0 entails for its own class of items. As such, we note the preferential treatment given by RSs, offered for items with a translation into English among foreign manuscripts.

Nevertheless, it is important to consider that the experiment is carried out on a sample of the dataset, which has different proportions of language characteristics in its item catalogue compared to the initial corpus. Therefore, the data bias observed in Section 5.1, emphasising the disproportionate number of items being English originals does not manifest itself in this particular case to such a drastic extent as in the `GR-lang` dataset.

5.3. Mixed Results Turn Reranking Mitigation Inconclusive

The reranking of the recommendation lists aimed at mitigating unfairness propagated through RSs, achieved marginal changes in the way the established provider groups are treated. In an effort to tackle **RQ3**, we found that the effectiveness of the reranking strategy utilised varied on a case by case basis depending on the algorithm and recommendation size. The impact of provider unfairness mitigation is low and with negative effects to the relevance of the lists to the users.

In the top-10 reranked recommendations, we notice that provider groups get slightly more balanced representation compared to that in the user profiles as well as proportional exposure to their presence in the item catalogue. We emphasise that if the visibility of provider groups gets calibrated to the user history, the recommendations would therefore reinforce the trends outlined in the user activity. Similarly, exposure deserving of the proportions of the provider classes in the item catalogue would cause bias in ranking akin to the established data bias. For better treatment of disadvantaged provider groups in the lists presented to users to be achieved, some advantageous behaviour of algorithm needs to be recorded for minority groups such as OT in terms of given visibility and exposure, and preserved throughout reranking.

Even if there is less variance in the success rate of presenting items of the established provider groups to users who are interested in them, we find this to be misleading the effects of the mitigation strategy employed. When delving into the individual probabilities of each group, we highlight that this parity stems from lower scores overall, rather than an improvement of the ones that failed to capture exposure and visibility on the lists of users that like the respective items. From an exposure standpoint, regarding the top-50 generated lists, there is no clear trend identified across all the sets of recommendations. Overall,

the differences in exposure given the representation of the item groups are slim and while most of the scores result in better balance, there are cases where the disparity worsens. However, the latter described effect is not consistent across a subset of provider groups.

The choice of reranking strategy for the mitigation experiment, CP, has led to a marginal reduction in the popularity bias observed in the initially evaluated recommendation lists. We note that the improvements to GC scores are minor, even if it increases when looking at the values obtained for the reranked top-10 recommendations. Consequently, the recommendations are slightly more diverse, focusing less around a small number of items and provider groups. This assessment is tied to the curtailing of the group exhibiting the most imbalance in terms of being overrepresented in the lists, namely the AMBG provider group which encompasses books with both foreign and English editions but no identified source language. Even so, some algorithms do not follow this trend and either remain unchanged from this perspective, record an exacerbation of bias or show different behaviour between popularity bias of items and that of provider groups.

The performance analysis indicates that a tradeoff has to be made in order to benefit from the highlighted unfairness mitigation between provider groups. We found that reranking led to worse satisfiability of the users, regardless of whether fairness in the recommendations improved.

5.4. Implications

Throughout this thesis, we made extensive use of the curated dataset `GR-lang` and its derived sample $\texttt{GR-lang}^s$ as the basis for the user profiles analysed in our experiments. We identified the language characteristics of the items found in the respective catalogue, as well as the availability of translations to English in the case of foreign books. Regarding this process of inference, we have inevitably run into issues with finding this type of information for more than half the catalogue, and discerning the language of origin for a small subset of these items. We theorise that the sources used in carrying out the inference of these characteristics offer scarce information about lesser-known manuscripts, which include the class of unknown items, judging from the deviation of the user profiles from the catalogue. As for the books that have editions in both English and at least another foreign language, but an unclear provenance, we note their disproportionate and advantageous consumption, in comparison to other categories, by the observed users. We expect that these popular items may have more records in the inspected sources, a factor which increases the chances for an erroneously filled-in document that inadvertently causes ambiguity.

Previous research [45] focused on group provider fairness in book recommendation found that the majority of items belong to American authors. Given that these providers represent a subset of the authors writing books in English, we notice that the highlighted trend holds in the identified catalogue. Although the quantity of native literature is better represented than translations from other languages in the data, the difference observed is not as drastic as stipulated by Heilbron [23]. The user's activity deviates heavily from the item catalogue from the point of view of the inferred characteristics likely due to the high percentage of unidentified books in the distribution. This is reinforced by Figure 3.2, which highlights a consumption pattern for literature with availability in English akin to initial expectations [30].

We note that the user profiles examined in this study are not homogenous and present deviations in consumption patterns from the available book catalogue. This reinforces the use of collaborative filtering based recommender algorithms to capture user tendencies which were shown to maximise user satisfaction [71] through personalisation. Even so, the algorithms evaluated in this thesis exhibited different behaviours with respect to the distribution and ranking of the provider groups in the generated recommendations backing past observations [19]. As user satisfaction increases, so does the treatment of the item categories which become more consistent. In fact, the top-3 best performing collaborative filtering algorithms display the lowest coefficient of variance between the probability of provider of being connected to interested users. This is emphasised by the $\text{REO}_{\text{Overall}}$ scores. ItemKNN offers the best parity from this viewpoint while being the second most accurate recommender. This displays a scenario where simple approaches over-perform more complicated methods [58]. This assessment of ItemKNN is in stark contrast to its observed fairness benchmarking in Music RSs with a focus on country representation [75] where the algorithm was found to calibrate the popularity of items, but exacerbate inequity between provider groups.

Inspecting the fairness and user relevancy performance of the evaluated algorithms, we acknowledge differences in the behaviour towards provider groups as highlighted by the metrics computed on the

recommendation lists. For instance, ItemAutoRec and Random both score poorly in all indices tracking relevancy of the items to the consumer (nDCG, MAP, MRR) but showcase a different composition of their recommendations. Following an analysis of the GC scores, we point out that ItemAutoRec tends to focus only on a few items in its recommendations process, while the naive baseline, mainly due to its random selection, offers more diversity both in terms of individual entries and groups. PMF, BPRMF and MF2020, while being significantly different in regards to their capabilities of recommending relevant books to users, share a similar struggles concerning the provider groups that fail to be included in the lists to their audience. This is exemplified by poor performance when it comes to UNKN and ONT provider groups, as defined by their REO scores. Furthermore, LRD scores show that PMF and BPRMF tend to penalise the representation of OT while MF2020 does not. Such observations attest that algorithms manifest various tendencies with regard to the composition of the recommendation and display different weaknesses when it comes to both provider fairness and relevance to the user.

Foreign books that lack translation to English, but also manuscripts with unidentifiable language characteristics show great improvement in how they are being presented to users who show interest in these items in MultiVAE, ItemKNN and Slim in clear contrast to the rest of the algorithmic choices. We speculate that catering to such minority provider groups, given what can be observed in the user activity, will contribute significantly to the algorithm's performance with respect to the relevance of recommendations to consumers. By observing the shifts in DE and LRD that occur when limiting the analysis to the top-10 recommendations, we note that algorithms that achieve lists of lesser relevance to the user exacerbate the biases targeting provider groups at the top of their recommendations. In comparison, Slim, ItemKNN and MF2020 maintain the recorded disparity towards exposure and visibility of books depending on their language characteristics, or balance this behaviour.

Reflections drawn from our experimental results emphasise disparity in the treatment of the outlined provider groups by the chosen recommendation algorithms. The reranking through CP renders improvements from the point of view of calibrating the visibility and exposure of each provider group in most cases. However, it decreases the quality of connecting the items to users which they appeal to. This happens disproportionately, heavily downgrading more successful provider groups, without uplifting the rest, reaching a situation of equal parity under the premise that every provider is worse off. The highlighted behaviour of the examined reranking strategy fails to capture the intrinsic needs of a two-sided marketplace, where the needs of both consumers and providers have to be considered [6]. We observe that the results of the reranking procedure reflect a focus on fairness disregarding the cost of utility to the consumer [12], whereas in a real scenario, an ideal approach would incentivise finding a balance between both stakeholders [11]. As in Gómez, Boratto, and Salamó [20], we found that the reranking of the recommendation lists had varied effects, which led to reducing exposure and visibility of privileged groups in conjunction with more fairly treated providers, or, in the case of BPRMF, even reinforcing already overrepresented provider groups at the expense of others.

In terms of adopting CP to improve provider fairness in recommendation lists, we found that tackling popularity bias led to a reduction in visibility and a disproportionate gain of exposure for the popular provider group AMBG. This reaffirms previous claims [75, 45] that connect unfairness treatment of provider categories with popularity bias. We followed the insights drawn from Ferraro, Ekstrand, and Bauer [54] stating that recommender algorithms have higher agency in proliferating unfairness than user activity itself, and resorted to mimicking user tendencies in terms of popularity to achieve improvements from both consumer satisfaction and provider fairness perspectives. CP reranking was shown to enforce a slight reduction in popularity bias [40], and our findings confirm this property, on top of demonstrating an improvement of provider group diversity in the recommendations.

Reranking through CP is shown to propagate its effects across the entire user base, based on the small shift of the entire distributions pertaining to individual LRD and DE scores observed for reranked Slim. A possible explanation for why the magnitude of the reranking adjustments is low can be the low diversity in the recommendations to single users. Therefore, changes in the ranking of items will not lead to meaningful observable differences on a provider level. We suggest the reranking of larger lists as a way of mitigating the prevalence of a provider group in a list, as it allows new items to be introduced in the top-10 recommendations. In our scenario, this action helped increase the magnitude of balancing scores for the first recommendations, which are known to benefit of the most consumer attention [76].

The evaluation of fairness remains a complex issue and quantifying the bias proliferated through an RS

is not a trivial task. We analysed the fairness of outcome [11] as described by the provider groups' chances of reaching an interested audience. On top of that, we observe the exposure these groups receive in comparison to the amount of entries in the recommendations, as well as the differences in representation between the user profiles and the recommendations. Following these documented scores, we emphasise the diverging behaviour towards foreign works depending on whether a translation to English exists. While this is an expected consequence of availability in central languages [23], we cannot assess whether these changes are enough of an improvement or entail an overcorrection. To address this, we propose gathering valuable perspectives, akin to efforts in RS diversity [16, 48], from authors and other agents in the book domain. This would define expectations of the stakeholders surrounding fair recommendations from language and translation perspectives to be taken into account in RS development.

The provider fairness evaluation within an RS context should be conducted with a thorough understanding of the user activity and the available item catalogue. In the analysis carried out in this manuscript, we employ metrics that relate to the distribution of the groups in these cases. DE analyses whether the exposure a provider group receives is expected of their representation in the corpus. As such, balanced scores pinpoint the proliferation of any data bias previously ascertained in the catalogue. In the case of LRD, this metric reports the difference in the ratio of a provider group betewen user history and the recommendations. Therefore, values close to 0 underline that the generated recommendations follow the trends in the user profiles with respect to provider group presence.

We attempt to portray our findings in a real-life palpable scenario, in which a reader turns to an RS for suggestions about books to read. Whereas the list curated by the New York Times editorial, mentioned in Chapter 1, is static and does not necessarily seek to cater to the consumer's preferences, the recommendation list generated by the algorithm attempts to satisfy their interests. Consequently, the bias manifested against foreign books is tied to their representation in the user history. This is beneficial to readers who already consume foreign literature, but presents a series of newfound issues. As can be observed in Figure 5.1 and Figure 5.2, the two users benefit from different recommendations: one is limited to English originals, while the other receives entries exclusively from a single foreign author. This raises questions about the ability of RSs to diversify the contents of such lists with respect to the language traits, in conjunction with the prospect of receiving novel items of unknown relevance.

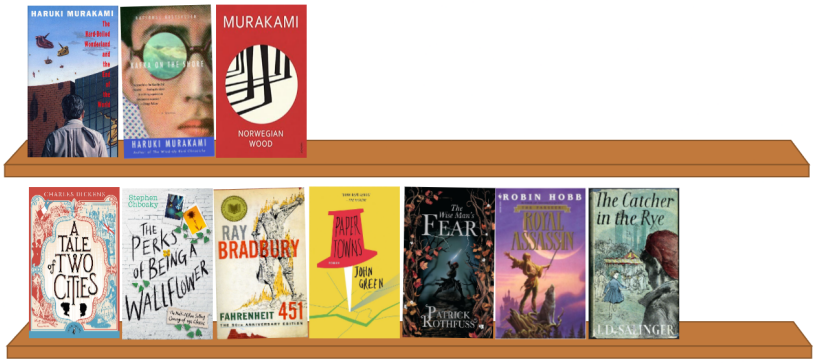


Figure 5.1: Top 10 book recommendations given to a user using Slim with 3 foreign books. The top shelf illustrates foreign books while the bottom one shows books originally written in English, verified manually.

5.5. Limitations

The empirical exploration that is carried out in this study suffers from several drawbacks that need to be recorded and considered when weighing the insights drawn from it.

The process to gather language traits and information about translation availability to English for each book, culminating in the creation of the `GR-lang` dataset is not without faults. We base our inference on fields that are not mandatory and can also be erroneously filled in either in terms of content or expected structure. The datafield inspected in LoC is optional, meaning that it is rarely completed whenever the book edition is not a translation. Even in the cases when the record refers to a translation, it can be completed erroneously, by concatenating the language codes or not respecting the indices described by the Marc21

Figure 5.2: Top 10 book recommendations given to a user using Slim with no foreign books. The top shelf illustrates foreign books while the bottom one shows books originally written in English, verified manually.

format. In the case of VIAF, the datafield is filled-in wrongly at times. Some glaring examples include not using the agreed upon language codes or inserting information that belongs in other datafields (e.g. author gender). Similar to Ekstrand et al. [19], we simplify the provider groups by only assigning one language to each provider, whereas in reality an author can publish books in multiple known languages.

The data sample $\texttt{GR-lang}^s$ was taken due to hardware constraints that prevented us from conducting an evaluation on the full dataset. Although we have tried to imitate the characteristics of GR-lang, it is possible that the traits exhibited in $\texttt{GR-lang}^s$ which are then examined in some of our experiments are not entirely representative. As can be seen from Section 3.1.3, we have decided to prioritise the authenticity of the user activity to the parent set to the detriment of the distribution of the book catalogue in terms of language characteristics. In addition, the outcomes discussed in this study can be impacted by the true language attributes of the items in unidentified groups such as UNKN and AMBG. False positives aggregated to the OT group belonging to EO can also change the representation and exposure of these groups in the recommendation lists.

In terms of algorithmic diversity, we attempted to choose a wide range of collaborative filtering algorithms depending on their underlying architecture. Nevertheless, we had to give up evaluating some deep learning solutions such as NeuMF [58] due to reduced hardware capabilities. This limitation also prompted the reduction of some of the costly hyperparameters in PMF during tuning. Concerning reranking strategies, we only consider CP, which does not offer a full picture of the mitigation potential of existing options in this domain. The item popularity was computed with respect to the $\texttt{GR-lang}^s$ using clearly defined proportions for each category, and without an in-depth analysis of how to accurately depict this property. Furthermore, we performed reranking with only one chosen ratio of balancing relevance recommendations and similarity to the popularity spread of the user profile, which can fail to capture the full performance of the approach.

Reflecting on the metrics that we employed in order to quantify the fair recommendation towards the involved provider groups, we note that our only measurements of diversity in the lists is given by GC_{item} and GC_{group}. This offers us a perspective about whether recommendations are centred on a single individual item or group respectively. Nevertheless, our study does not track the extent to which a reader encounters both novel and diverse content. An example of this would be that a user who has read only English originals is exposed to translated literature by the RS, while still offering books of the previous provider group in the list. We also do not assess how diverse the recommendation is within each group. There could be cases where a user only gets recommendations consisting of a small sample of foreign authors.

5.6. Future Work

Looking ahead, we define several directions that can further the research on fairness of the provider groups in RSs. We tackle various aspects of our evaluation process and propose extensions based on this study's contributions.

Concentrating on book recommendation, the umbrella term of foreign books is a label that incorporates vast amounts of diverse literature in our study. Our work restricts itself to the performance of this group

against items originally written in English, which are perceived to be an advantaged group. Nevertheless, the languages of these foreign items could be examined at a more granular level, which would enable the detection of disparity on a language basis. This would also reveal whether items belonging to other central languages [23], such as French or German mimic the representation rate of English in the item catalogue. More specifically, this process would uncover which languages benefit the most out of a translation and whether this reflects in their popularity with the user base.

Moving forward, we consider the identification of translation availability and language traits of books through the use of prompts to large language models, analogous to work on stereotype detection [39]. A more robust inference process of these characteristics would surmount limitations defined by the sources consulted in this study.

The nDCG scores for users reveal that a minority of users is better served in terms of the recommendations they receive. Building on this finding, we consider the approach of Pathak, Spezzano, and Pera [71] to conduct an analysis of the recommendations served to super readers. For instance, a possibility for these consumers could be the RS exhausting the items of a certain provider group before opting to diversify. Another viewpoint could focus on the languages readers speak and whether recommendations cover items in all of them. A possible scenario is when RSs forward books which the consumer does not have access to from a language proficiency point of view.

To address the questions raised in Section 5.5 about the lack of knowledge regarding the diversification and introduction of novel content in the recommendations from the perspective of the outlined provider groups, we turn to a possible method for measuring this aspect. From Kunaver and Požrl [77] a possible solution would entail computing the item diversity by aggregating their similarity, relevance and ranking. The established provider groups, as well as the author of a book could be part of the similarity computation.

In terms of mitigation, we propose a comparison including other reranking strategies (e.g FA*IR [41, 54] or calibrating disparity [20]) in the context of language and translation availability, as a means of understanding their capabilities to combat unfairness in the established scenario. Feedback loops [54] could be used as viable method for observing how the generated recommendations evolve over time in their fairness characteristics and to reflect the impact that reranking has on them. Calibrating the recommendations [78] based on the proportions of the book language characteristics observed in the user profile could lead to improvement in the disparate LRD scores we highlighted.

6

Ethical Considerations

The empirical analysis described in this thesis is performed following a rigorous ethical conduit as enforced by the institution of the TU Delft. In our research, we base our results on interactions between two groups of users formed by readers on one side and the authors of the analysed books on the other. This chapter covers our considerations for handling the data of the aforementioned individuals, as well as insights on the design of the experiments in our research and the ethical perspectives that can be drawn from the discussion outlined in Section 5.4.

6.1. Data Management

We used a publicly available dataset [28] containing user-item interactions that were recorded on the GoodReads platform and collected towards the end of 2017. The gathered information is free to access and consult on the respective platform. We note that the identity of the consumers is anonymised in the data we consulted. The GoodReads Terms of Use[1] define that the audience using the platform must be at least 13 years old. Keeping in mind that this directive is difficult to apply in practice, we make no attempt to infer the background of the readers [79]. The items under scrutiny include metadata of the author of the books which are presented under the identity or pseudonym of their choice.

To satisfy the demand for language traits of the books rated by users, we turn to publicly available data dumps that allow their usage in scientific research. These resources are offered by their respective curators (VIAF[2], OL[3], LoC[4]) and contain information about books and their respective authors. A detailed overview of how this information was linked with the item catalogue of `GR-base` to annotate the language of provenance and translation availability of the manuscripts to English can be found in Section 3.1.2. Due to redistribution restrictions[5] stipulated in the conditions of usage of the GoodReads dataset, we ensured that there are no traces of the inferred data in this project's code repository[6].

6.2. Research Process Reflection

In the interest of transparency and reproduction of our findings, the methodology outlined in this paper is meant to guide the reader through the process leading up to our experiments. We present the full extent of the results achieved in this study regardless of the prospect of rejecting our initial hypotheses. A critical analysis of our work from the perspective of its limitations is reported to help build an informed context around our conclusions.

Any external tool that we use in the implementation of our experiments is thoroughly documented. We include any changes and configurations that were made in order to tackle the research goal. Although we do not offer the dataset with the inferred language characteristics of books, we describe the steps that need to be taken in order to reconstruct it and share insights on its structure. This ensures that future

[1] `https://www.goodreads.com/about/terms`

[2] `https://viaf.org/en/viaf/data`

[3] `https://openlibrary.org/developers/dumps`

[4] `https://www.loc.gov/cds/products/marcDist.php`

[5] `https://cseweb.ucsd.edu/~jmcauley/datasets/goodreads.html`

[6] `https://github.com/raresboza/thesis_lost_in_translation`

research can reuse our evaluation scripts irrespective of the data source used.

6.3. Ethical Use and Study of Recommender Systems

Through the efforts outlined in this study, we reflect on the impact of RSs on any of the stakeholders engaged in this ecosystem. RSs are already entities known to be able to steer public opinion and influence the preference of their users [25]. A taxonomy of ethical challenges in the field of RSs [80] identifies the issue of disparity between utility to any type of user as threats to a system's integrity. In our scenario, the danger of the systematic decrease in exposure and visibility of providers based on the language of writing satisfies this condition.

We recognise that the mislabelling of books in our research can result in the downplay of the gravity of unfairness proliferation in RSs from the perspective of linguistic traits of providers, just as it can lead to its exaggeration. This paper serves as a contribution to a growing sense of responsibility in RS development and adoption to monitor the continuous threat to fairness in its use. Through the discrepancies found in the treatment of various provider groups, we emphasise the need for open discourse to identify the issues perceived by affected individuals in RSs.

7

Conclusion

The book market greatly benefitted from the addition of manuscripts in electronic format and digitalisation on the Internet, by reducing the costs of book creation and their distribution, together with lowering the barriers imposed by traditional publishing [81]. Readers also adopted this style of consuming literature and are met with further recommendations on their devices [82]. A survey [83] focused on the reading tendencies of teenagers emphasises that social media in its many forms is a popular method of discovering the next book to read by this age group. Nevertheless, this industry is known to suffer from historical biases that affect authors from social perspectives, which raises questions about how this is projected online. Our study in particular is centred on translations and how they contribute to the competition between English and the traditionally disadvantaged class of foreign books [23].

Recommender Systems constitute a central piece of online two-side markets [84] which are incorporated in Internet book stores and dedicated social media for reading. The core task of RSs is matching relevant items to users that find utility in them. Whereas past research [47, 41] identified how different types of biases can affect RS behaviour, growing attention is paid to the fairness of recommendations to users, grouped on sensitive traits that are subject to discrimination [12, 11, 48]. Provider fairness scrutinisation has already revealed the proliferation of inequitable treatment in the book domain with regards to gender [19, 44] and country of origin [45]. We contribute to this endeavour by analysing the composition of generated recommendations from the angle of translated foreign books and how this group fares in comparison English originals and its foreign, untranslated counterpart. Therefore, our aim is to identify and quantify the language diversity in the available catalogue, as well as frame the user activity trends with regard to this topic. We question whether RSs exhibit differences in their recommendation of the outlined provider groups and seek to mitigate these disparities.

In this thesis, we conducted an empirical exploration of the trends observed in book recommendation relating to provider fairness anchored in an RS ecosystem and centred on the language of origin of the books in the catalogue as well as their translation availability to English. We analysed the representation of the aforementioned traits in the item catalogue used for recommendation, as well as the user's tendency to mimic the inferred proportions in their own activity. The hypotheses of English originals garnering disproportionate attention and translations offering foreign books a better chance of competing were jointly tested by investigating their performance in an offline recommendation setting. We assessed the ability to mitigate the disparity between the treatment of established provider groups through a reranking strategy specifically targeting a reduction of popularity bias.

Outcomes of our study highlight the data bias found within the identified linguistic traits in the book corpus, leaning heavily towards a prevalence of English originals. The representation of translated books outweighs that of literature where such an edition is unavailable, while users' consumption pattern steers away from books with unidentified characteristics. At a recommendation level, when inspecting the best performing algorithms in terms of user satisfaction, we find that the group of translated books receive additional representation in the generated lists compared to its own presence in user history. This emphasises a clear gain over untranslated foreign manuscripts which tend to maintain resemblance to reader history. We observe that a popular group of works where language of provenance could not be established, but identified editions are both written in English, and at least one foreign language is offered unfairly, advantageous exposure and visibility by the recommendation algorithms. Due to the uncertainty

of these ambiguous items, we could not establish a difference in the treatment between translations and English originals. The behaviour of RSs towards the three aforementioned provider groups denotes the widespread presence and implicit favouring of English available books in the generated recommendations. Although reranking had marginal impact on the fairness of the recommendations at a loss of relevancy to the user, it reduced previously given advantages and calibrated the groups towards a closer representation of that in the user profiles.

We outlined several limitations in this study that could be used as a foundation for future work. The provider groups under investigation in this research match the original language only in a binary manner - either English or foreign. We encourage further granularity in the inference of these language characteristics, which would enable the capturing of different trends and inequalities [23] amongst various languages. The within-group diversity, but also the introduction of novel provider groups in the lists generated by RSs is not measured in this thesis. Nevertheless, the analysis of the latter phenomenon could spark interest about whether translations foster the introduction of foreign literature in recommendations to users that have not read any before. Besides Calibrated Popularity, more reranking strategies can be evaluated to emphasize their impact on provider fairness. Future research could also introduce feedback loops [54] to assess how reranking performs over multiple user sessions.

The findings presented in this study underline practices that can be used by RS developers and researchers to conduct fairness evaluations from the viewpoint of the providers. We expand on previous efforts meant to capture cultural representation and diversity in book recommendations and the treatment of such groups. To the best of our knowledge, the work examining linguistic traits in the context of provider fairness in RSs is novel, even though this perspective is strongly tied to the country of origin of the author [20, 45]. The contributions highlighted constitute an advancement towards book RSs serving both the needs of the providers, while tracking the satisfaction of the consumers through the lens of language accessibility of the manuscripts to an international audience.

References

[1] Deuk Hee Park et al. “A literature review and classification of recommender systems research”. In: *Expert systems with applications* 39.11 (2012), pp. 10059–10072.

[2] Francesco Ricci et al. “Recommender Systems: Introduction and Challenges”. In: *Recommender Systems Handbook*. Ed. by Francesco Ricci et al. Boston, MA: Springer US, 2015, pp. 1–34. DOI: 10.1007/978-1-4899-7637-6_1. URL: https://doi.org/10.1007/978-1-4899-7637-6_1.

[3] Yunqi Li et al. “Fairness in recommendation: Foundations, methods, and applications”. In: *ACM Transactions on Intelligent Systems and Technology* 14.5 (2023), pp. 1–48.

[4] Jinfei Yang et al. “A novel technique applied to the economic investigation of recommender system”. In: *Multimedia Tools and Applications* 77 (2018), pp. 4237–4252.

[5] Matthias Bogaert et al. “Evaluating multi-label classifiers and recommender systems in the financial service sector”. In: *European Journal of Operational Research* 279.2 (2019), pp. 620–634.

[6] Rishabh Mehrotra et al. “Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems”. In: *Proceedings of the 27th acm international conference on information and knowledge management*. 2018, pp. 2243–2251.

[7] Asela Gunawardana et al. “Evaluating recommender systems”. In: *Recommender systems handbook*. Springer, 2012, pp. 547–601.

[8] Mouzhi Ge et al. “Beyond accuracy: evaluating recommender systems by coverage and serendipity”. In: *Proceedings of the fourth ACM conference on Recommender systems*. 2010, pp. 257–260.

[9] Marius Kaminskas et al. “Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems”. In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7.1 (2016), pp. 1–42.

[10] Jiawei Chen et al. “Bias and debias in recommender system: A survey and future directions”. In: *ACM Transactions on Information Systems* 41.3 (2023), pp. 1–39.

[11] Yifan Wang et al. “A survey on the fairness of recommender systems”. In: *ACM Transactions on Information Systems* 41.3 (2023), pp. 1–43.

[12] Di Jin et al. “A survey on fairness-aware recommender systems”. In: *Information Fusion* 100 (2023), p. 101906.

[13] Yashar Deldjoo et al. “Fairness in recommender systems: research landscape and future directions”. In: *User Modeling and User-Adapted Interaction* 34.1 (2024), pp. 59–108.

[14] Markus Zanker et al. “Measuring the impact of online personalisation: Past, present and future”. In: *International Journal of Human-Computer Studies* 131 (2019). 50 years of the International Journal of Human-Computer Studies. Reflections on the past, present and future of human-centred technologies, pp. 160–168. DOI: https://doi.org/10.1016/j.ijhcs.2019.06.006. URL: https://www.sciencedirect.com/science/article/pii/S107158191930076X.

[15] Dietmar Jannach et al. “Towards More Impactful Recommender Systems Research.” In: *ImpactRS@ RecSys*. 2019.

[16] Sanne Vrijenhoek et al. “Diversity of what? On the different conceptualizations of diversity in recommender systems”. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2024, pp. 573–584.

[17] Dana B Weinberg et al. “Comparing gender discrimination and inequality in indie and traditional publishing”. In: *PloS one* 13.4 (2018), e0195298.

[18] Mike Thelwall. “Reader and author gender and genre in Goodreads”. In: *Journal of Librarianship and Information Science* 51.2 (2019), pp. 403–430.

[19] Michael D. Ekstrand et al. “Exploring author gender in book rating and recommendation”. In: *Proceedings of the 12th ACM conference on recommender systems*. 2018, pp. 242–250.

[20] Elizabeth Gómez et al. “Provider fairness across continents in collaborative recommender systems”. In: *Information Processing & Management* 59.1 (2022), p. 102719.

[21] Marion Dalvai. “Translating literature into English in the twenty-first century: Opportunities and challenges”. In: *Orbis litterarum* 74.6 (2019), pp. 392–410.

[22] Kathryn Zickuhr et al. “Libraries, Patrons, and E-Books.” In: *Pew Internet & American Life Project* (2012).

[23] Johan Heilbron. “Towards a sociology of translation: Book translations as a cultural world-system”. In: *European journal of social theory* 2.4 (1999), pp. 429–444.

[24] Dan Cosley et al. “Is seeing believing? How recommender system interfaces affect users’ opinions”. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2003, pp. 585–592.

[25] Gediminas Adomavicius et al. “Do recommender systems manipulate consumer preferences? A study of anchoring effects”. In: *Information Systems Research* 24.4 (2013), pp. 956–975.

[26] Carol Maier et al. “Literature in translation: teaching issues and reading practices”. In: (2010).

[27] Rhoda Myra Garces-Bacsal. “Diverse books for diverse children: Building an early childhood diverse booklist for social and emotional learning”. In: *Journal of Early Childhood Literacy* 22.1 (2022), pp. 66–95.

[28] Mengting Wan et al. “Item recommendation on monotonic behavior chains”. In: *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*. Ed. by Sole Pera et al. ACM, 2018, pp. 86–94. DOI: 10.1145/3240323.3240369. URL: https://doi.org/10.1145/3240323.3240369.

[29] Nettie Finn et al. “Pseudonymous Disguises: Are Pen Names An Escape from the Gender Bias in Publishing?” In: (2016).

[30] Melanie Walsh et al. “The goodreads “classics”: a computational study of readers, Amazon, and crowdsourced amateur criticism”. In: *Journal of Cultural Analytics* 6.2 (2021), pp. 243–287.

[31] Rodrigo Fernandes Malaquias et al. “Understanding the Effect of Culture on E-Book Popularity during COVID-19 Pandemic.” In: *Turkish Online Journal of Educational Technology-TOJET* 20.2 (2021), pp. 182–188.

[32] Denis Kotkov et al. “Clusterexplorer: enable user control over related recommendations via collaborative filtering and clustering”. In: *Proceedings of the 14th ACM Conference on Recommender Systems*. 2020, pp. 432–437.

[33] Huiyuan Chen et al. “Hessian-aware Quantized Node Embeddings for Recommendation”. In: *Proceedings of the 17th ACM Conference on Recommender Systems*. 2023, pp. 757–762.

[34] Vito Walter Anelli et al. “Challenging the myth of graph collaborative filtering: a reasoned and reproducibility-driven analysis”. In: *Proceedings of the 17th ACM Conference on Recommender Systems*. 2023, pp. 350–361.

[35] Daniele Malitesta et al. “A Novel Evaluation Perspective on GNNs-based Recommender Systems through the Topology of the User-Item Graph”. In: *Proceedings of the 18th ACM Conference on Recommender Systems*. 2024, pp. 549–559.

[36] Chiyu Zhang et al. “Embsum: Leveraging the summarization capabilities of large language models for content-based recommendations”. In: *Proceedings of the 18th ACM Conference on Recommender Systems*. 2024, pp. 1010–1015.

[37] Gustavo Penha et al. “What does bert know about books, movies and music? probing bert for conversational recommendation”. In: *Proceedings of the 14th ACM conference on recommender systems*. 2020, pp. 388–397.

[38] Alessandro Petruzzelli et al. “Instructing and prompting large language models for explainable cross-domain recommendations”. In: *Proceedings of the 18th ACM Conference on Recommender Systems*. 2024, pp. 298–308.

[39] Robin Ungruh et al. “Mirror, Mirror: Exploring Stereotype Presence Among Top-N Recommendations That May Reach Children”. In: *ACM Transactions on Recommender Systems* (2025).

[40] Himan Abdollahpouri et al. “User-centered evaluation of popularity bias in recommender systems”. In: *Proceedings of the 29th ACM conference on user modeling, adaptation and personalization*. 2021, pp. 119–129.

[41] Robin Ungruh et al. “Putting Popularity Bias Mitigation to the Test: A User-Centric Evaluation in Music Recommenders”. In: *Proceedings of the 18th ACM Conference on Recommender Systems*. 2024, pp. 169–178.

[42] Kohei Hirata et al. “Solving diversity-aware maximum inner product search efficiently and effectively”. In: *Proceedings of the 16th ACM Conference on Recommender Systems*. 2022, pp. 198–207.

[43] Yuncong Li et al. “Modeling User Repeat Consumption Behavior for Online Novel Recommendation”. In: *Proceedings of the 16th ACM Conference on Recommender Systems*. 2022, pp. 14–24.

[44] Shrikant Saxena et al. “Exploring and mitigating gender bias in book recommender systems with explicit feedback”. In: *Journal of Intelligent Information Systems* 62.5 (2024), pp. 1325–1346.

[45] Savvina Daniil et al. “Hidden author bias in book recommendation”. In: *arXiv preprint arXiv:2209.00371* (2022).

[46] Samira Vaez Barenji et al. “User and Recommender Behavior Over Time: Contextualizing Activity, Effectiveness, Diversity, and Fairness in Book Recommendation”. In: *arXiv preprint arXiv:2505.04518* (2025).

[47] Savvina Daniil et al. “Reproducing popularity bias in recommendation: The effect of evaluation strategies”. In: *ACM Transactions on Recommender Systems* 2.1 (2024), pp. 1–39.

[48] Karlijn Dinnissen et al. “Amplifying artists’ voices: Item provider perspectives on influence and fairness of music streaming platforms”. In: *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. 2023, pp. 238–249.

[49] Karlijn Dinnissen. “Fairness and Transparency in Music Recommender Systems: Improvements for Artists”. In: *Proceedings of the 18th ACM Conference on Recommender Systems*. 2024, pp. 1368–1375.

[50] Oleg Lesota et al. “Analyzing item popularity bias of music recommender systems: are different genders equally affected?” In: *Proceedings of the 15th ACM conference on recommender systems*. 2021, pp. 601–606.

[51] Elizabeth Gómez et al. “Disparate impact in item recommendation: A case of geographic imbalance”. In: *European Conference on Information Retrieval*. Springer. 2021, pp. 190–206.

[52] Elizabeth Gómez et al. “The winner takes it all: geographic imbalance and provider (un) fairness in educational recommender systems”. In: *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 2021, pp. 1808–1812.

[53] Michael D Ekstrand et al. “All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness”. In: *Conference on fairness, accountability and transparency*. PMLR. 2018, pp. 172–186.

[54] Andres Ferraro et al. “It’s not you, it’s me: the impact of choice models and ranking strategies on gender imbalance in music recommendation”. In: *Proceedings of the 18th ACM Conference on Recommender Systems*. 2024, pp. 884–889.

[55] Savvina Daniil et al. “On the challenges of studying bias in Recommender Systems: A UserKNN case study”. In: *arXiv preprint arXiv:2409.08046* (2024).

[56] Savvina Daniil. “Bias in Book Recommendation”. In: *Proceedings of the 18th ACM Conference on Recommender Systems*. 2024, pp. 1376–1381.

[57] Greg Linden et al. “Amazon. com recommendations: Item-to-item collaborative filtering”. In: *IEEE Internet computing* 7.1 (2003), pp. 76–80.

[58] Steffen Rendle et al. “Neural collaborative filtering vs. matrix factorization revisited”. In: *Proceedings of the 14th ACM Conference on Recommender Systems*. 2020, pp. 240–248.

[59] Steffen Rendle et al. “BPR: Bayesian personalized ranking from implicit feedback”. In: *arXiv preprint arXiv:1205.2618* (2012).

[60] Maurizio Ferrari Dacrema et al. “A troubling analysis of reproducibility and progress in recommender systems research”. In: *ACM Transactions on Information Systems (TOIS)* 39.2 (2021), pp. 1–49.

[61] Andriy Mnih et al. “Probabilistic matrix factorization”. In: *Advances in neural information processing systems* 20 (2007).

[62] Dawen Liang et al. “Variational autoencoders for collaborative filtering”. In: *Proceedings of the 2018 world wide web conference*. 2018, pp. 689–698.

[63] Suvash Sedhain et al. “Autorec: Autoencoders meet collaborative filtering”. In: *Proceedings of the 24th international conference on World Wide Web*. 2015, pp. 111–112.

[64] Harald Steck. “Item popularity and recommendation accuracy”. In: *Proceedings of the fifth ACM conference on Recommender systems*. 2011, pp. 125–132.

[65] Himan Abdollahpouri et al. “The unfairness of popularity bias in recommendation”. In: *arXiv preprint arXiv:1907.13286* (2019).

[66] Anastasiia Klimashevskaia et al. “Evaluating The Effects of Calibrated Popularity Bias Mitigation: A Field Study”. In: *Proceedings of the 17th ACM Conference on Recommender Systems*. 2023, pp. 1084–1089.

[67] María Luisa Menéndez et al. “The jensen-shannon divergence”. In: *Journal of the Franklin Institute* 334.2 (1997), pp. 307–318.

[68] Vito Walter Anelli et al. “Top-n recommendation algorithms: A quest for the state-of-the-art”. In: *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. 2022, pp. 121–131.

[69] Ziwei Zhu et al. “Measuring and mitigating item under-recommendation bias in personalized ranking systems”. In: *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 2020, pp. 449–458.

[70] Masoud Mansoury et al. “A graph-based approach for mitigating multi-sided exposure bias in recommender systems”. In: *ACM Transactions on Information Systems (TOIS)* 40.2 (2021), pp. 1–31.

[71] Royal Pathak et al. “Understanding the contribution of recommendation algorithms on misinformation recommendation and misinformation dissemination on social networks”. In: *ACM Transactions on the Web* 17.4 (2023), pp. 1–26.

[72] Joao Vinagre et al. “Statistically robust evaluation of stream-based recommender systems”. In: *IEEE Transactions on Knowledge and Data Engineering* 33.7 (2019), pp. 2971–2982.

[73] Vito Walter Anelli et al. “Elliot: A Comprehensive and Rigorous Framework for Reproducible Recommender Systems Evaluation”. In: *SIGIR ’21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*. Ed. by Fernando Diaz et al. ACM, 2021, pp. 2405–2414. DOI: 10.1145/3404835.3463245. URL: https://doi.org/10.1145/3404835.3463245.

[74] Lien Michiels. “Methodologies to evaluate recommender systems”. PhD thesis. University of Antwerp, 2024.

[75] Oleg Lesota et al. “Oh, behave! country representation dynamics created by feedback loops in music recommender systems”. In: *Proceedings of the 18th ACM Conference on Recommender Systems*. 2024, pp. 1022–1027.

[76] Mathew S. Isaac et al. “The Top-Ten Effect: Consumers’ Subjective Categorization of Ranked Lists”. In: *Journal of Consumer Research* 40.6 (Dec. 2013), pp. 1181–1202. DOI: 10.1086/674546. eprint: https://academic.oup.com/jcr/article-pdf/40/6/1181/9500561/40-6-1181.pdf. URL: https://doi.org/10.1086/674546.

[77] Matevž Kunaver et al. “Diversity in recommender systems–A survey”. In: *Knowledge-based systems* 123 (2017), pp. 154–162.

[78] Harald Steck. “Calibrated recommendations”. In: *Proceedings of the 12th ACM conference on recommender systems*. 2018, pp. 154–162.

[79] Tomasz Huk et al. “Use of Facebook by children aged 10-12. Presence in social media despite the prohibition”. In: *The New Educational Review* 46.1 (2016), pp. 17–28.

[80] Silvia Milano et al. “Recommender systems and their ethical challenges”. In: *Ai & Society* 35 (2020), pp. 957–967.

[81] Joel Waldfogel et al. “Storming the gatekeepers: Digital disintermediation in the market for books”. In: *Information economics and policy* 31 (2015), pp. 47–58.

[82] Andreja Zubac et al. “A research of e-book market trends: North America and the European Community”. In: *Knjižnica: revija za področje bibliotekarstva in informacijske znanosti* 58.1-2 (2014).

[83] Leonie Rutherford et al. “Discovering a good read: Exploring book discovery and reading for pleasure among Australian teens”. In: (2024).

[84] Arpita Biswas et al. “Toward fair recommendation in two-sided platforms”. In: *ACM Transactions on the Web (TWEB)* 16.2 (2021), pp. 1–34.