

RESEARCH ARTICLE

Understanding Editing Behaviors in Multilingual Wikipedia

Suin Kim¹✉, Sungjoon Park¹✉, Scott A. Hale², Sooyoung Kim¹, Jeongmin Byun¹, Alice H. Oh¹*

1 School of Computing, KAIST, Daejeon, Republic of Korea, **2** Oxford Internet Institute, University of Oxford, Oxford, United Kingdom

✉ These authors contributed equally to this work.

* alice.oh@kaist.edu



OPEN ACCESS

Citation: Kim S, Park S, Hale SA, Kim S, Byun J, Oh AH (2016) Understanding Editing Behaviors in Multilingual Wikipedia. PLoS ONE 11(5): e0155305. doi:10.1371/journal.pone.0155305

Editor: Eduardo G. Altmann, Max Planck Institute for the Physics of Complex Systems, GERMANY

Received: August 13, 2015

Accepted: April 27, 2016

Published: May 12, 2016

Copyright: © 2016 Kim et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Data are available from Figshare at <https://dx.doi.org/10.6084/m9.figshare.3218728.v1>.

Funding: This work was supported by Korea Ministry of Science and ICT and Future Planning, Grant 10041313, UX-oriented Mobile Software Platform; John Fell Oxford University Press (OUP) Research Fund (<https://www.admin.ox.ac.uk/pras/jffi>); and University of Oxford's Economic and Social Research Council Impact Acceleration Account and Higher Education Innovation Fund (HEIF) (<http://www.esrc.ac.uk/collaboration/knowledge-exchange/opportunities/ImpactAccelerationAccounts.aspx>), reference: IAA/HEIF-DIA-013. The funders had no

Abstract

Multilingualism is common offline, but we have a more limited understanding of the ways multilingualism is displayed online and the roles that multilinguals play in the spread of content between speakers of different languages. We take a computational approach to studying multilingualism using one of the largest user-generated content platforms, Wikipedia. We study multilingualism by collecting and analyzing a large dataset of the content written by multilingual editors of the English, German, and Spanish editions of Wikipedia. This dataset contains over two million paragraphs edited by over 15,000 multilingual users from July 8 to August 9, 2013. We analyze these multilingual editors in terms of their engagement, interests, and language proficiency in their primary and non-primary (secondary) languages and find that the English edition of Wikipedia displays different dynamics from the Spanish and German editions. Users primarily editing the Spanish and German editions make more complex edits than users who edit these editions as a second language. In contrast, users editing the English edition as a second language make edits that are just as complex as the edits by users who primarily edit the English edition. In this way, English serves a special role bringing together content written by multilinguals from many language editions. Nonetheless, language remains a formidable hurdle to the spread of content: we find evidence for a complexity barrier whereby editors are less likely to edit complex content in a second language. In addition, we find that multilinguals are less engaged and show lower levels of language proficiency in their second languages. We also examine the topical interests of multilingual editors and find that there is no significant difference between primary and non-primary editors in each language.

Introduction

Wikipedia is the world's largest general reference work, and it depends on active editors to generate and maintain up-to-date and accurate information. Wikipedia is also one of the top ten websites in terms of traffic volume, and its articles are often among the top results for many search queries on Google [1].

role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

There are 288 language editions of Wikipedia hosted by the Wikimedia Foundation, providing easy access to information for many Internet users globally, but there are high levels of inequality and asymmetry in the information available in the different language editions.

Information Inequality and Asymmetry

The English edition contains more than 4.9 million articles as of June 2015, which is 13.8% of all the articles in the 288 language editions [2]. In comparison, the Chinese edition has 827,273 articles and the Arabic edition has 373,064 articles. When looking at the number of edits or active editors, the inequality is even greater. The number of active editors is greatest for the English edition of Wikipedia, which has 6.4 times more active editors than the second most active edition, German. About 38.4% of all edits ever made to Wikipedia were to the English edition. This results in situations where users search for information on Wikipedia, only to find that it is not available in their own languages. For example, searching for PLOS ONE on Wikipedia finds the longest and most comprehensive article in English, but no results in Russian and only very short articles in German and Arabic. In addition to the inequality shown by the large differences in the number of articles and active editors, there is also an asymmetry of information between the different language editions: many topics are available in only one language or a small number of languages [3]. This applies even to the English edition. Although the English edition is the largest edition, it contains, for example, only 51% of the articles that exist in the German edition [3]. Such asymmetry is especially pronounced for Wikipedia articles about local places and events, which are mostly written only in the local languages of those locations [4]. For example, KAIST is a major science and technology university and research institution in Korea with many international students and faculty, but there is no article about it in the Spanish edition of Wikipedia.

There are several reasons for this information inequality and asymmetry in Wikipedia. First, Wikipedia was only available in English when it started in January 2001. The German and the Catalan editions were added two months later, and other language editions followed after a few years, but English has always remained the largest edition [5]. Second, English is a *de facto* standard language of the Internet and the hub among all global languages [6–8]. Lastly, many editors tend to contribute local content and that leads to an asymmetry in the information available in the different language editions [9]. For example, on average 23% of edits by unregistered users of the English edition of Wikipedia were to articles less than 100km from their location [9].

As awareness of this information inequality and asymmetry has increased there have been efforts to grow or distribute more information for several language editions. In 2010, Google sponsored a contest to encourage students in Tanzania and Kenya to contribute to the Swahili edition of Wikipedia [10], and in 2007 the German government allocated funds to support the creation of articles in the German edition of Wikipedia [11]. Even with these efforts, the English edition remains the largest and the most extensive with 2.5 times more articles compared to the second-largest edition, Swedish.

Multilingual Editors in Wikipedia

Previous research has shown that multilingual users play a key role in information diffusion across languages in social media [12–16]. In Wikipedia, where editors are not explicitly linked in a social network, multilingual editors can still play a key role in mitigating the asymmetry by transferring information across languages [7]. Thus, we present in-depth research to understand the editing behaviors of multilingual editors in Wikipedia.

Approximately 15% of active Wikipedia editors are multilingual, contributing content to multiple language editions of the encyclopedia [7]. Wikipedia provides a global account system for a single login across all Wikimedia sites, as well as interlanguage links connecting articles on the same concepts across languages. However, other than a small mixed-methods study of the contributions of Japanese–English bilingual editors on articles about Okinawa, Japan [17], little is known about the content contributions of these multilingual editors at a larger scale or across different language pairs.

Unlike the studies on monolingual Wikipedia editors [18–21], we must consider that a multilingual editor's behavior may differ for each language. Studies have shown that multilingual speakers feel differently in each of the languages they speak [22, 23] and tend to use a different language depending on the purpose, domain, and conversational partner [24]. In addition, multilingual individuals rarely possess equal and perfect fluency in all their languages [25]. Given this, we expect multilingual Wikipedia editors to also behave differently in each language edition of Wikipedia they edit. More specifically, they may contribute to one language more than another, they may edit articles on different topics in each language, and their edits may demonstrate different levels of proficiency in each language. While comparing multilingual editors' behaviors within each language may be ideal, different languages cannot be compared directly in terms of topics and proficiency because of inherent differences among them. Instead, we look at the behaviors of the multilingual editors within each language. For each language edition, we group the editors into those who edit that language edition more than any other language edition (*primary editors* of that language) and those who edit some other language edition more (*non-primary editors* of that language) and then compare the behaviors of the primary and non-primary multilingual editors. We then ask the following three research questions about multilingual editors in the English, German, and Spanish language editions.

RQ1. Do primary and non-primary editors show different levels of *engagement*?

RQ2. Do primary and non-primary editors show different levels of topical *interest*?

RQ3. Do primary and non-primary editors show different levels of *language proficiency*?

Each editor's engagement is quantitatively measured by their text contributions and the time spent revising articles. The interests of editors are identified by the topics of the articles they edited. Lastly, the language proficiencies of editors are measured by various language complexity measures applied to their contributions.

Our results can be summarized in three parts. First, there are significant differences in the levels of engagement and language proficiency in the German and Spanish editions: multilingual users who primarily edit either of these editions show higher levels of engagement and language proficiency than multilingual users who edit these editions as non-primary languages. Second, there was no notable difference between the degrees of engagement and proficiency of editors who edited the English edition as a primary or non-primary language, verifying the common assumption of English as *lingua franca* of the Web. Third, primary and non-primary editors show similar levels of interest for most topics with the exception that primary editors are significantly more interested in local topics and non-primary editors in global topics.

Our contributions to the field of Web-based study of multilingualism are as follows:

- We construct an extensive dataset of multilingual Wikipedia edit history, which comprises more than 5 million edits and make it publicly available for future research.
- We define and analyze three relevant aspects of multilingual editors' behavior: engagement, interest, language complexity.

- We define and validate several language-independent measures for quantifying the language complexity of edits.
- We show that multilingual editors indeed have potential to help mitigate the inequality and asymmetry in the information available in different languages.

Methods

In this section, we describe the data collection and analysis methods. To analyze the editors' engagement, interests, and language proficiency levels, we first start with the edit metadata from the English, German, and Spanish editions for one month and construct article edit sessions to identify consecutive contributions to articles by the same editor. Based on these article edit sessions, we extract multilingual editors and their contributions. By analyzing their contributions, we (1) measure their degrees of *engagement* through the number of contributions they made to articles, (2) discover their *interests* through the topics of the articles they edited, (3) estimate their *language proficiency* levels for each language by measuring the language complexity of their edits.

Dataset

To extract only the contributions of multilingual editors to articles from the unstructured edit history data, we conduct a data processing pipeline consisting of the following steps. We first start with metadata of English, German, and Spanish language editions from July 8 to August 9, 2013. We then construct article edit sessions by grouping together consecutive edits to the same article by the same editor. Based on the identified sessions, we define multilingual editors as those who are involved in article edit sessions for two or more language editions. Finally, we download the actual revision text for the article edit sessions of those multilingual editors. We describe these steps in more detail below.

Identifying Edits from Metadata. We begin with the edit metadata collected by Hale [7] for the largest 46 language editions of Wikipedia and extract the data for three of the largest Wikipedia editions, in terms of the number of articles: English (en), German (de), and Spanish (es). The metadata includes article titles, language editions, timestamps, editor ids, and URLs to the content of each edit captured from a near real-time broadcast on Internet Relay Chat. The extracted data comprises 2,799,729 edits by 146,616 distinct editors. We discard edits to non-article pages including article talk pages and user pages. We further discard edits that are indicated as being made by Wikipedia (ro)bots. In contrast to Hale [7], we retain edits marked as “minor” because what one considers as minor may vary from person to person and language to language.

Article Edit Session. Even though the easiest way to measure the edit activity on Wikipedia is simply counting each edit when changes are submitted, this does not accurately reflect editors' behavior because of individual differences in activity patterns. For example, some editors may submit a few large edits while others may make a series of smaller edits saving the pages more frequently as they work. To account for these individual differences, we adopt the idea of *edit sessions* [26], which measure the labor of Wikipedia editors whose contributions tend to occur in bursts. In our work, we limit each edit session to a single document, and so we rename it *article edit session*. We use *one hour* as the cutoff between intra-session and inter-session edit activities.

In other words, from the collected dataset, we define an *article edit session* as a sequence of continuous editing activity without a break of more than one hour for a specific article by a single editor.

Primary vs. Non-Primary Editors. Next, we identify multilingual editors from the metadata and retrieve the content of all the edits made by multilingual editors from Wikipedia

using the Wikipedia API. We define a multilingual editor as an editor who edits two or more language editions. Using this definition, we identified 12,577 multilingual editors with 427,529 total article edit sessions composed of 622,766 distinct edits. Following Hale [7], we operationally define primary and non-primary users as follows:

- We identify the *primary* language of an editor as the language of the edition that the multilingual editor edits most frequently.
- We identify the *non-primary* languages of an editor as all the other languages that they edit.
- For each language edition X, we define *primary editors* as the set of all users with primary language X.

Thus, an editor can be a primary editor in only one language edition, but can be a non-primary editor in multiple language editions.

Fig 1a shows the distribution of the number of languages per user in our multilingual editor dataset. We found that most multilingual editors contribute to two or three language editions. There are also a small number of editors with edits in more than 10 languages. Table 1 contains the statistics of the edit history data we use, and Fig 1b shows the proportions of the primary languages of editors who are contributing to the English (en), German (de), and Spanish (es) editions. The data for the three languages are available online [27].

Collecting Edit Text. Wikipedia provides the previous and current versions of each article. To look at the actual edited text, we download the “diffs,” which are the parts of the text that have changed from the previous version of the article for each article edit session we identified. Fig 2 shows an example of the diff provided by the Wikipedia Web interface. For each article edit session, we extract the pairs of the changed paragraphs and convert the edit from Wiki markup to plain text.

In this way, we retain only the visible text from edits, discarding all non-visible and non-text information including URL, multimedia, metadata, and document structure. We regard an edit as *non-visible* if there is no visible text change. An example of such an edit is adding a link to existing text but otherwise making no other changes.

Distribution of Article Edit Sessions. We examine the distribution of multilingual editors’ contributions by language and the number of article edit sessions. Fig 3 shows the distribution of editors by the number of article edit sessions. The plot, on a log-log scale, shows that the distributions for primary and non-primary users in all three editions is heavy-tailed; most users perform only a few edits while a few users perform many edits.

Quantifying Editors’ Behavior

Our approach is to analyze and compare the behavior of primary and non-primary multilingual editors for the English, German, and Spanish editions in terms of *engagement*, *interest*, and *language proficiency*. We measure editors’ engagement by looking at the editing patterns quantitatively within article edit sessions. Then, we investigate editors’ interests by focusing on the content of edits in terms of the main topics of the articles they edited. Lastly, we measure levels of language proficiency by defining and computing several language complexity metrics.

Examining Edits in Four Aspects. We define an edit paragraph as a line of Wiki markup in an article, utilizing the difference of the text before and after the edit from Wikipedia as shown in Fig 2. In an article edit session, all of the edit paragraphs before revisions, shown in the left side of Fig 2, are paired with their corresponding revised edits, shown in the right side of Fig 2. With those paired paragraphs, we examine the edit paragraphs in four ways as follows:

- **Pre-edit:** Edit paragraph (text) before revision

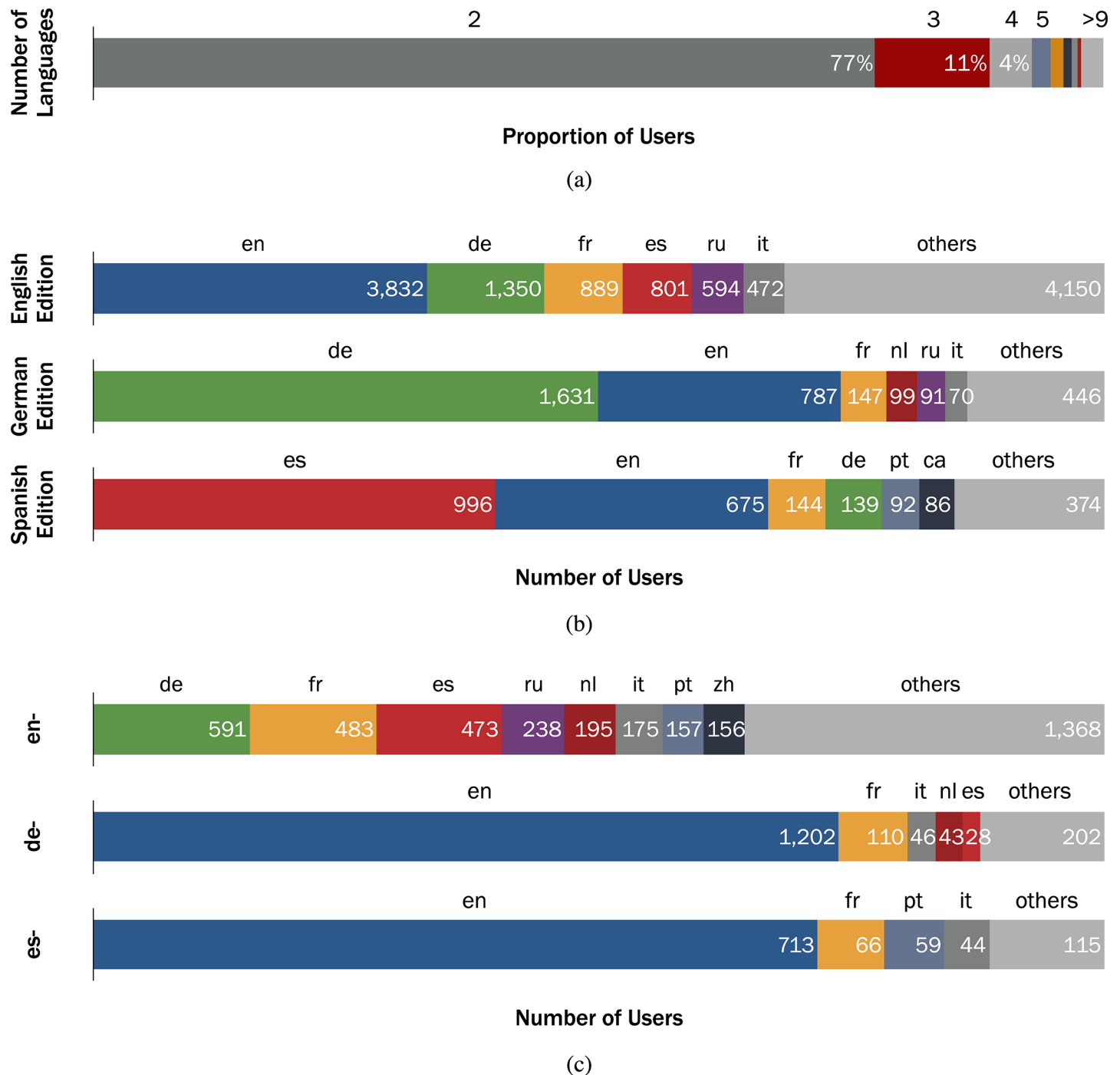


Fig 1. (a) Number of languages multilinguals edit in Wikipedia. Regarding the number of languages they edit, 77.3% of multilingual editors are bilingual, followed by 11.4% trilingual and 4.1% and quadrilingual editors. We discard editors who edited more than 10 languages (these editors account for 2.3% of all multilingual editors). **(b) Primary languages of multilingual editors for three of the largest language editions.** The English edition has the largest, yet the most varied, number of multilingual editors by primary language. 32.9% of the multilingual editors who edited the English edition primarily edited English. In comparison, 49.9% of the multilingual editors in the German edition primarily edited the German edition. Multilingual editors who primarily edit English are the second-largest proportion of multilingual editors in the Spanish and German editions. **(c) Second most used languages for the three primary editor groups.** English primary editors have much more diverse language usage, compared to German and Spanish primary editors. Most of German and Spanish primary editors contribute to English edition as their second most edited version.

doi:10.1371/journal.pone.0155305.g001

Table 1. Number of editors, article edit sessions, edits, and edit paragraphs. There is more activity in the English edition (en) than in either the German (de) or Spanish (es) edition. In all three language editions there are more primary editors (p) than non-primary editors (np) and primary editors are more active than non-primary editors.

	en-p	en-np	de-p	de-np	es-p	es-np
# Multilingual Editors	3,832	7,784	1,631	1,640	996	1,510
# Article Edit Sessions	200,883	36,959	112,788	7,334	63,947	5,609
# Edits	298,868	51,665	151,014	9,111	104,341	7,757
# Edited paragraphs	1,447,692	230,893	816,647	27,656	554,762	25,340

doi:10.1371/journal.pone.0155305.t001



Fig 2. Two example diffs of edits to a Wikipedia article 2015 FIFA Women's World Cup. (a) Edit paragraph containing a visible edit. (b) Edit paragraph containing a non-visible edit. We define an edit paragraph as one line of Wikipedia markup, utilizing the diff from Wikipedia. *Pre-edit* text and *post-edit* text are shown on the left and right sides, respectively, with text differences highlighted in yellow (deleted text) and blue (inserted text).

doi:10.1371/journal.pone.0155305.g002

- **Post-edit:** Edit paragraph (text) after revision
- **Diff:** The actual revision including both inserted and deleted text in an edit
- **Delta:** The difference between a *Pre-Edit* paragraph and its corresponding *Post-Edit* paragraph for each computed measure

The *Diff* includes both inserted and deleted text and is used to measure an editor's direct contributions. The difference between computed measures for *Pre-edits* and *Post-edits*, defined as *Delta*, is an indirect approach that considers the context of an edit.

Comparing Editor Behaviors within Languages. Rather than comparing each editor's behavior across languages, we compare editing behavior of primary versus non-primary editors within each language, as defined in the previous section. That means, for example, we take all editors in our dataset who contributed to any articles in the English edition, divide them into two groups: 1) those who contributed to the English edition more than any other language edition, and 2) those who contributed to some other language edition more than the English

+ EN-P × EN-NP + DE-P × DE-NP + ES-P × ES-NP

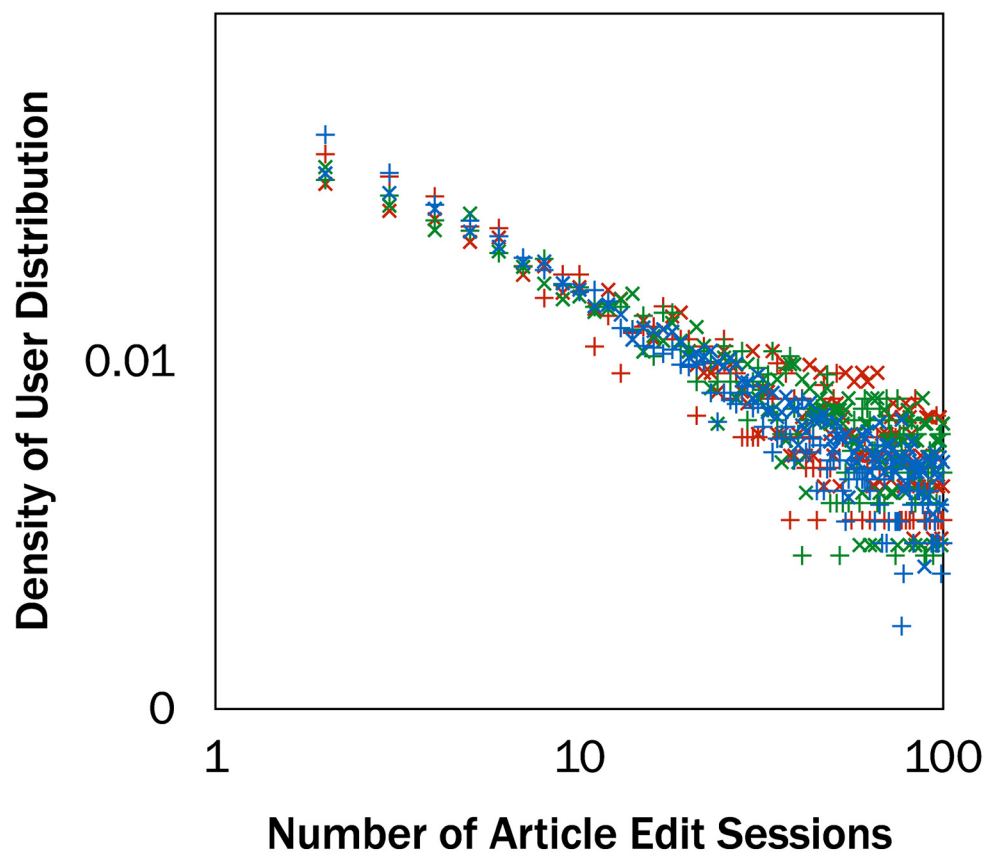


Fig 3. Distribution of editors and article edit sessions for the three language editions—English (EN), German (DE), and Spanish (ES)—for primary (P) and non-primary (NP) multilingual editors. We display the number of article edit sessions up to 100. Dots in the same color denote the same language edition. The plot, on a log-log scale, shows that the distributions for primary and non-primary users in all three editions is heavy-tailed; most users perform only a few edits while a few users perform many edits.

doi:10.1371/journal.pone.0155305.g003

edition, and then we compare the behaviors of those two groups. This way, we do not need to worry about the inherent differences among languages (e.g., German tends to have longer words than English).

Editor Engagement. Since we categorized multilingual editors into two groups (primary and non-primary) on the basis of the number of article edit sessions they had across language editions, we measure the level of quantitative engagement of editors within an article edit session, such that the measure becomes independent of the number of article edit sessions.

We measure the amount of engagement based on four metrics. First, we count the number of revisions committed within an article edit session, only including the sessions containing more than 1 edit. Second, we measure the session length in minutes by the difference of time-stamps between the first and the last revision in each article edit session. Third, we measure the amount of text added and deleted in terms of characters, words, and sentences for all the *Delta*

Table 2. Measures of engagement, interest, and language proficiency.

Engagement
- Number of edits
- Session length in minutes
- Number of edited characters, words, sentences (for <i>Delta</i>)
- Fraction of non-visible article edit sessions.
Interest
- Normalized frequency of edits for each topic
Language Proficiency
- Lexical diversity measures (for <i>Pre-Edit</i> and <i>Delta</i>)
- Entropy of unigram, bigram, trigram frequencies
- Syntactic complexity measures (for <i>Pre-Edit</i> and <i>Delta</i>)
- Entropy of POS unigram, bigram, trigram frequencies
- Article usage measures (for <i>Diff</i>)
- Fraction of articles in added tokens

doi:10.1371/journal.pone.0155305.t002

in a session. Last, we count the number of edit sessions that only include non-visible edits indicating the changes the editor made did not affect the text of the article (i.e., modifying URL link, images, etc.). By computing this, we can examine whether an editor's engagement results in article content changes. These four metrics are summarized in [Table 2](#).

For each metric, we compute the mean over all article edit sessions for each editor, and then again compute the mean of primary and non-primary editors to represent the level of engagement for each groups. We perform two-tailed independent samples t-test to examine if the differences between the means of the two groups are significant.

Editor Interest. Assuming that the topics of the edited articles represent editors' interests in specific fields, we measure the interests of multilingual editors using a Bayesian topic model. We first determine the main topic category for each article a multilingual user edited and assign each article a single topic label. We then compute the proportion of interest from primary and non-primary editors for each topic. By comparing the distributions, we can compare the interests of primary and non-primary editors.

- **Topic Modeling.** We use a Bayesian topic model as an automatic method to categorize Wikipedia articles into different topics. Although Wikipedia already provides "categories", there are many cases where one Wikipedia article is assigned to multiple categories. Moreover, Wikipedia categories form a complex network in which the relationships between categories are often unclear [28]. Instead of using these Wikipedia-defined categories, we develop a more consistent and replicable methodology. We first model the topics using a widely-used topic model, the latent Dirichlet allocation (LDA) [29] using the Python Gensim [30] library with the online variational Bayes algorithm. We set the number of topics (k) to 100 and use the default values for the hyperparameters.
- **Clustering and Labeling Articles.** With the 100-dimensional topic proportion vectors generated by LDA from the step above, we cluster the articles using DBSCAN [31] in the Python Scikit-learn [32] package. The generated clusters show consistency in the topics of the articles; hence, we consider each cluster as representing a single topic. We manually examine the articles in each cluster, assign a topic label to the cluster, and assign the same topic label to each article in the cluster. This process results in 20 clusters and their topic labels for each

language edition with all articles labeled with one of the 20 topic labels. To verify the clustering process, we simply compare the inter-cluster and intra-cluster distances using Euclidean distance. In Eq (2), the numerator is the average of all pairwise distances of the articles in a cluster, representing the intra-class distance. The denominator is the mean of the pairwise distances of the cluster medoids for all clusters, representing the inter-class distance. For both the numerator and the denominator, c_i denotes cluster i in the set of all clusters C , and x denotes each article in c_i . The resulting I_{c_i} represents the ratio of intra-class to inter-class distance; so, we would expect this ratio to be significantly below one if the clustering is done well. The average value of I_{c_i} for i th cluster is 0.59 with a standard deviation 0.22.

$$\text{medoid}(c_i) = \arg \min_m \sum_{x \in c_i} \text{dist}(x, m) \quad (1)$$

$$I_{c_i} = \frac{\frac{1}{|c_i|} \sum_{x \in c_i} \text{dist}(x, \text{medoid}(c_i))}{\frac{1}{|C| - 1} \sum_{c_j \in C, c_j \neq c_i} \text{dist}(\text{medoid}(c_i), \text{medoid}(c_j))} \quad (2)$$

- **Comparing Levels of Interest.** The previous step provides topic labels for each article. With those labels, we can define, for each multilingual editor, the level of interest on a topic as the proportion of the article edit sessions for articles in that topic over all article edit sessions by the editor. Then, we average the interest levels of editors in the primary and non-primary groups. We define this as the “normalized frequency of edits for each topic” (also listed in Table 2). To compare the interest levels between the two groups, we perform a two-tailed independent samples t-test for each topic.

Language Proficiency. In order to measure the language proficiency of multilingual editors, we focus on three aspects of their edits: (1) *Pre-Edits*, (2) *Delta*, and (3) *Diff*.

Using *Pre-Edits*, we analyze the language complexity of the paragraphs of the articles that multilingual editors choose to revise. Using *Delta* and *Diff* we quantify the editors’ contributions in terms proficiency to estimate editors’ linguistic abilities. With *Pre-Edits* and *Delta*, the textual contents are sufficient in length to perform both lexical diversity and syntactic complexity measures, which have previously been used as estimates of language proficiency in language acquisition research [33]. Meanwhile, for *Diff*, which often consist of just a few words, we focus on the definite and indefinite articles edited (e.g., ‘a’, ‘an’, and ‘the’ in English). These have also been shown by previous research to be a good proxy for language proficiency in English, German, and Spanish [34–36].

- **Lexical Diversity Measures.** We first compute the entropy of unigram, bigram, and trigram frequencies as complexity measures [37]. These measures are used to quantify the richness of word usage in Wikipedia articles [18]. To calculate the entropy for each edit paragraph, we average the metrics over all unigrams, bigrams, and trigrams that appear in the paragraph.
- **Syntactic Complexity Measures.** We compute the entropy of parts-of-speech (POS) frequency. Analyzing the sequence of POS has advantages over analyzing the sequence of words. POS entropy can ignore extremely trivial edits (e.g., correcting typos) as well as meaningless bot-produced edits [18]. In addition, looking for diverse combinations of POS is a good approach for detecting complex syntactic structure [38]. We count the occurrence of POS unigrams, bigrams, and trigrams in each edit paragraph and then compute an entropy measure based on the frequency distributions of POS that captures the amount of

Table 3. The definite and indefinite articles tracked in English, German, and Spanish.

	English	German	Spanish
Definite	the	des, die, den, der, dem, das	el, la, los, las
Indefinite	a, an	eine, eines, einer, einem, einen, ein	un, una, unos, unas

doi:10.1371/journal.pone.0155305.t003

information and thus indicates the diverse use of POS in the edits.

For automatic POS tagging on edits, we employ a maximum entropy POS tagger [39] trained on the Penn Treebank corpus for English edits with Penn Treebank tagset [40]. For German edits, we use the Stanford log-linear POS tagger [41] trained on the NEGRA corpus with Stuttgart-Tübingen Tagset (STTS) [42]. For Spanish edits, we use the tagger trained on the AnCora corpus with its tagset [43].

- **Usage of Articles.** A number of previous studies investigate the difficulty of learning the definite and indefinite articles for language learners. For example, Butler [34] finds that children who are learning English as their second language (L2) have more difficulty in understanding articles compared to children who are acquiring English as their first language (L1). They show that L1 learners make lower frequency of errors than L2 learners. In addition, Japanese-speaking English learners have the most difficulty in understanding the article system and even proficient English learners only record about a 90% of success rate when given a task to choose the proper article [44]. In addition, the difficulty of the article system in Spanish and German are comparable to English [35, 36].

The difficulty comes from the complex usage of articles: i.e., that there is not a one-to-one correspondence between languages. Such difficulty imposes a challenge for language learners of a second language [34]. Since it is computationally challenging to automatically test the appropriateness of the articles editors use, we approximate the level of understanding of the article system in each language by measuring the frequencies with which definite and indefinite articles are used. We count the number of definite and indefinite articles in each *Diff* by primary and non-primary users. The definite and indefinite articles we count in each language are shown in Table 3. The number of articles in a *Diff* is divided by total number of words of the *Diff*, and we refer to this quantity as the “fraction of articles in added tokens.”

- **Computing Language Complexity Measures for Each Editor.** The language proficiency measures were summarized with a *maximum* value for each editor rather than the mean value. We compute the maximum value of the proficiency measures for all edits belonging to the same editor as a representation of that editor’s highest displayed level of linguistic ability to produce complex edits. The maximum value is a fairer estimate than the mean or the minimum value as not all possible revisions to an article always require an editor’s maximum linguistic ability. For this reason, widely used central tendency measures (e.g., mean) would not reflect editors’ true proficiencies properly. Summarizing each editor’s edits with the maximum value assumes that each editor shows his/her maximum linguistic ability to edit a paragraph at least once, which is more reasonable.

Since the maximum value may be merely affected by the number of edits, we control the number across editors to three. We uniformly sample three edits within the entire edits an editor made, repeating 100 times for editors and average the evaluated 100 maximum values.

All of the engagement, interest, and proficiency measures described in this section are summarized in Table 2.

Results

In this section, we report our findings on engagement, interests, and language proficiency of primary and non-primary multilingual editors for the English, German, and Spanish editions of Wikipedia.

Editor Engagement

We show in [Fig 4](#) that primary editors commit more edits than non-primary editors within an article edit session, for all three language editions. Likewise, the article edit sessions of primary editors are longer than those of non-primary editors. These results indicate that primary editors are more engaged. They make more edits and spend more time revising each article.

We also find that the number of tokens added per edit session by primary editors is higher than the number added by non-primary editors in all three language editions. This also holds whether tokens are measured by characters, words, or sentences. These findings on the amount of edited content align with the previous findings on the number of edits and the overall time spent indicating that in general primary editors are more engaged in revising the text of articles than non-primary editors.

Finally, we find that non-primary editors make more non-visible edits, such as adding/removing hyperlinks or applying stylistic changes, in all three languages. This tendency indicates that editors may be making different types of edits in their primary and non-primary languages. Similar results have been shown in qualitative research (e.g., [\[17\]](#)) at a much smaller scale.

Editor Interests

The twenty topics discovered for the articles edited by multilingual users in the English edition are as follows: *Science, Football, Film, Middle East Geography, American Sports, Songs & Albums, Musicians, Cities, Global Sports, TV Shows, Politics, History, Military, Transportation, Computer, Education, Geographical Locations, Descriptive, Olympics, and Animals & Plants*. [Fig 5](#) shows the number of articles in each topic in the English edition. The titles of the representative articles for each topic are presented in [S1](#), [S2](#) and [S3](#) Tables.

Similarly, the twenty topics discovered for Spanish articles are as follows: *Art, Descriptive, Soccer, Film, Animal, Global Sports, History, Plants, Politicians, Natural Science, Social Science, Music, Cities, Geographical Locations, Olympics, Literature, Musicians, Politics, Entertainment, and Tennis*.

Finally, the twenty topics discovered for German articles are as follows: *Computer, Natural Science, Descriptive, Geographical Locations: U.S., Names, Geographical Locations: Europe, History, Academic, Celebrities, Soccer, Cultural Heritage, Musicians, Natural Topography, Land Transport, Politicians, Entertainment, Air transport, Global Sports, Authors, and Military*.

The overall difference of interest between primary and non-primary editors is not significant ($\chi^2_{en}(19) = 0.005$, $\chi^2_{es}(19) = 0.031$, $\chi^2_{de}(19) = 0.019$, *ns*). Indeed, as shown in [Fig 5](#) bottom, we observe that primary and non-primary editors of the English edition have similar interests. This tendency is also shown in the other language editions.

However, when looking at each topic, we observe notable differences of the level of interests between primary and non-primary editors for a few topics in each language. In the English edition, we observe two topics with significantly different levels of interest between primary and non-primary editors. Primary editors show significantly higher level of interest in the topic of *Computer*, while non-primary editors show higher level of interest for *Cities*. In the German edition, primary editors show more interest in *Computer* and *Natural Science*, while non-

Editor Engagement

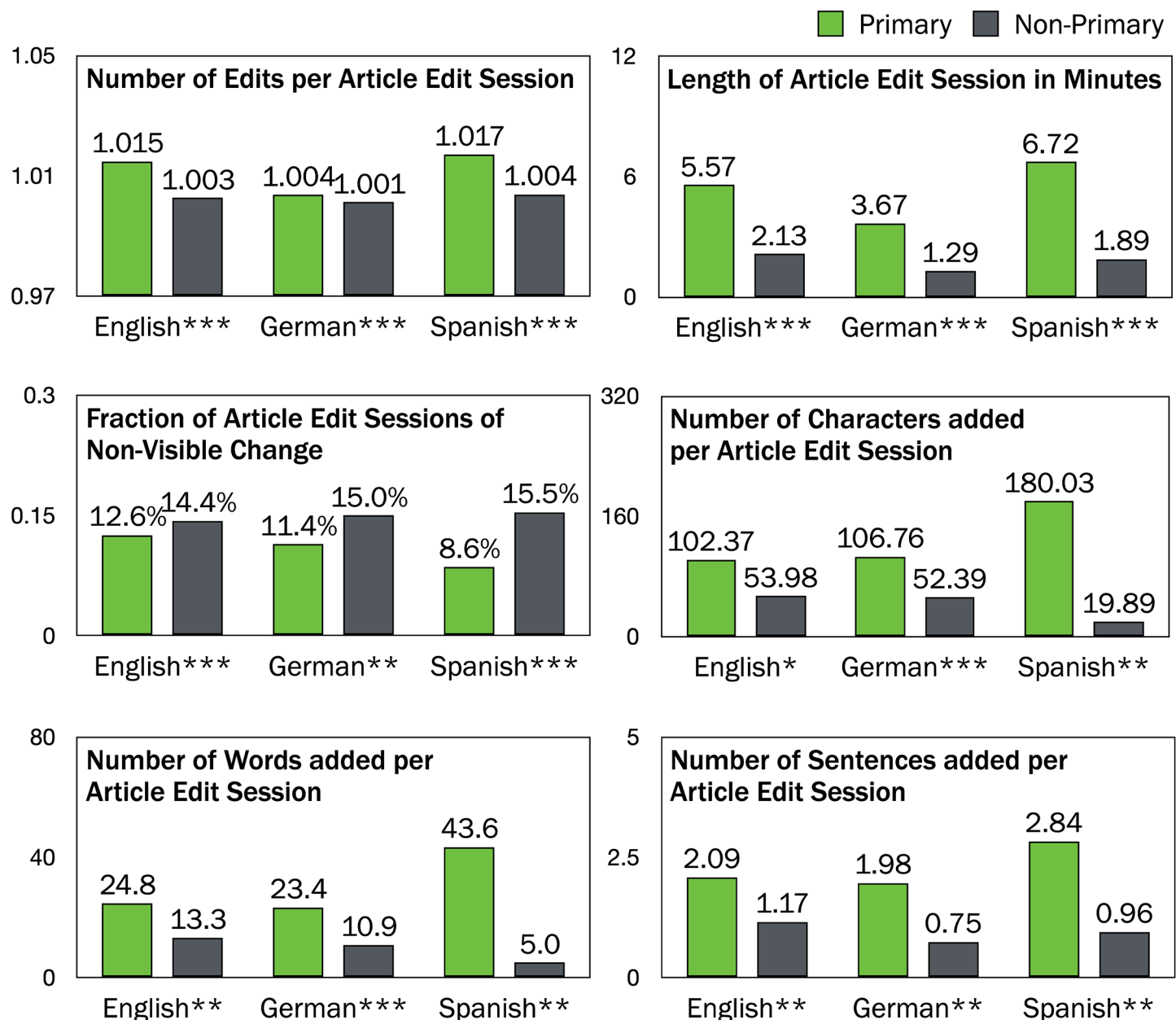


Fig 4. The evaluated engagement measures for primary and non-primary editor groups. For all metrics and languages, primary and non-primary editors are showing significantly different behavior: primary editors tend to be more engaged than non-primary editors. (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.)

doi:10.1371/journal.pone.0155305.g004

primary editors show more interest in *Soccer* and *Global Sports*. In the Spanish editions, primary editors show more interest in *Politicians*, *Social Science*, and *Entertainment*, while non-primary editors show more interest in *Plants* and *Geographical Locations*.

In the German and the Spanish editions, we also observe an interesting pattern where topics with more interest from primary users have higher syntactic complexity compared to topics with more interest from non-primary users (see [Table 4](#)). For example, in the German edition,

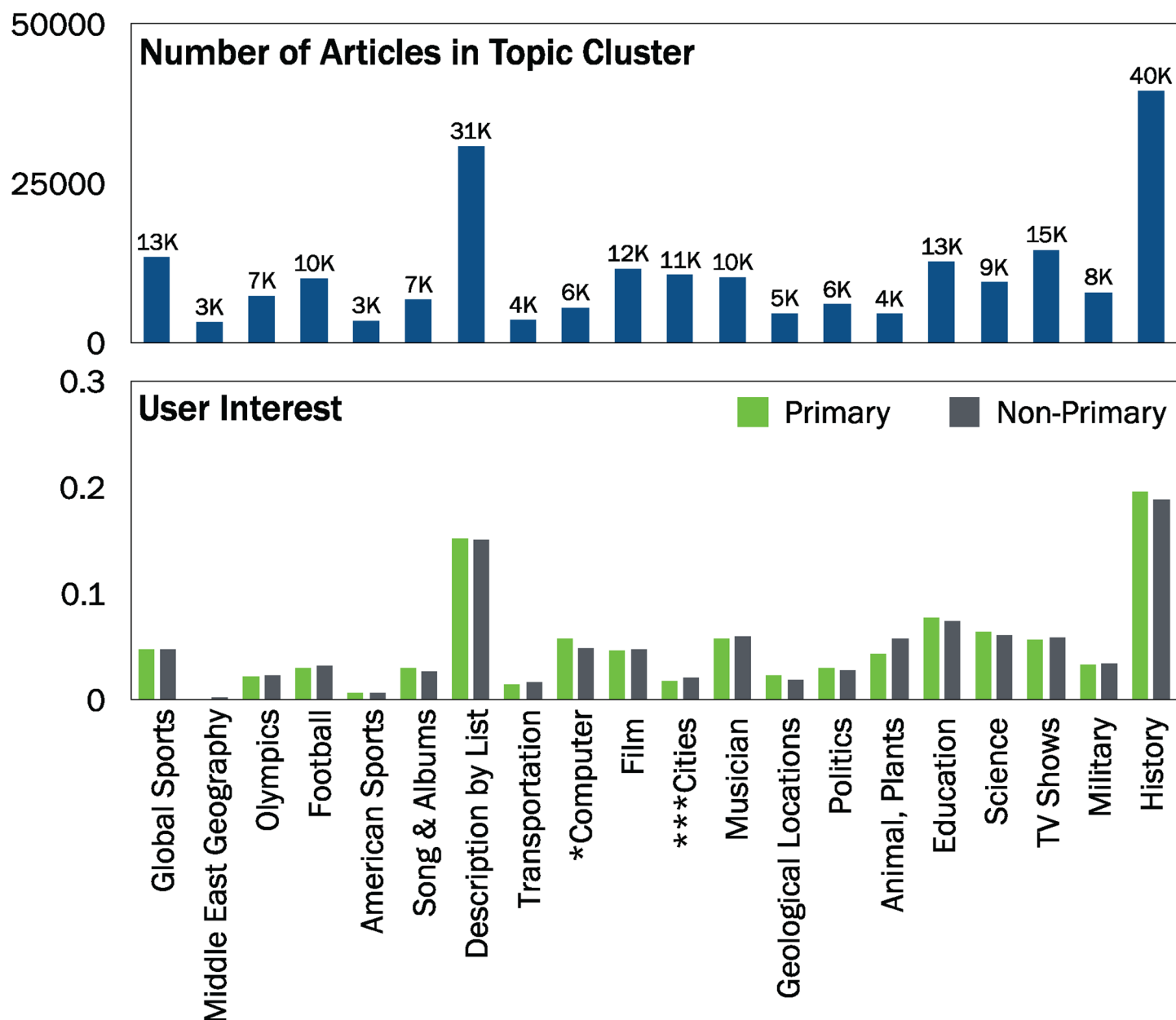


Fig 5. Top: Number of English Wikipedia articles in each topic cluster. Articles related to Descriptive and History make up a large proportion of all articles while Middle East Geography, American Sports, and Transportation articles are relatively small in proportion. **Bottom: Differences in interest in the English edition.** Primary and non-primary editors show similar levels of interest for most topics. (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.)

doi:10.1371/journal.pone.0155305.g005

Table 4. Topics having significant difference of interest levels between primary and non-primary users in each language edition. Number next to each topic represents the average entropy for Part-of-Speech trigrams for the articles within each topic.

	English	German	Spanish
Primary	Computer (2.58)	Computer (3.21), Natural Science (3.52)	Politicians (3.19), Social Science (3.01), Entertainment (2.97)
Non-Primary	Cities (2.63)	Soccer (2.76), Global Sports (2.52)	Descriptive (2.80), Plants (2.37), Geographical Locations (2.53)

doi:10.1371/journal.pone.0155305.t004

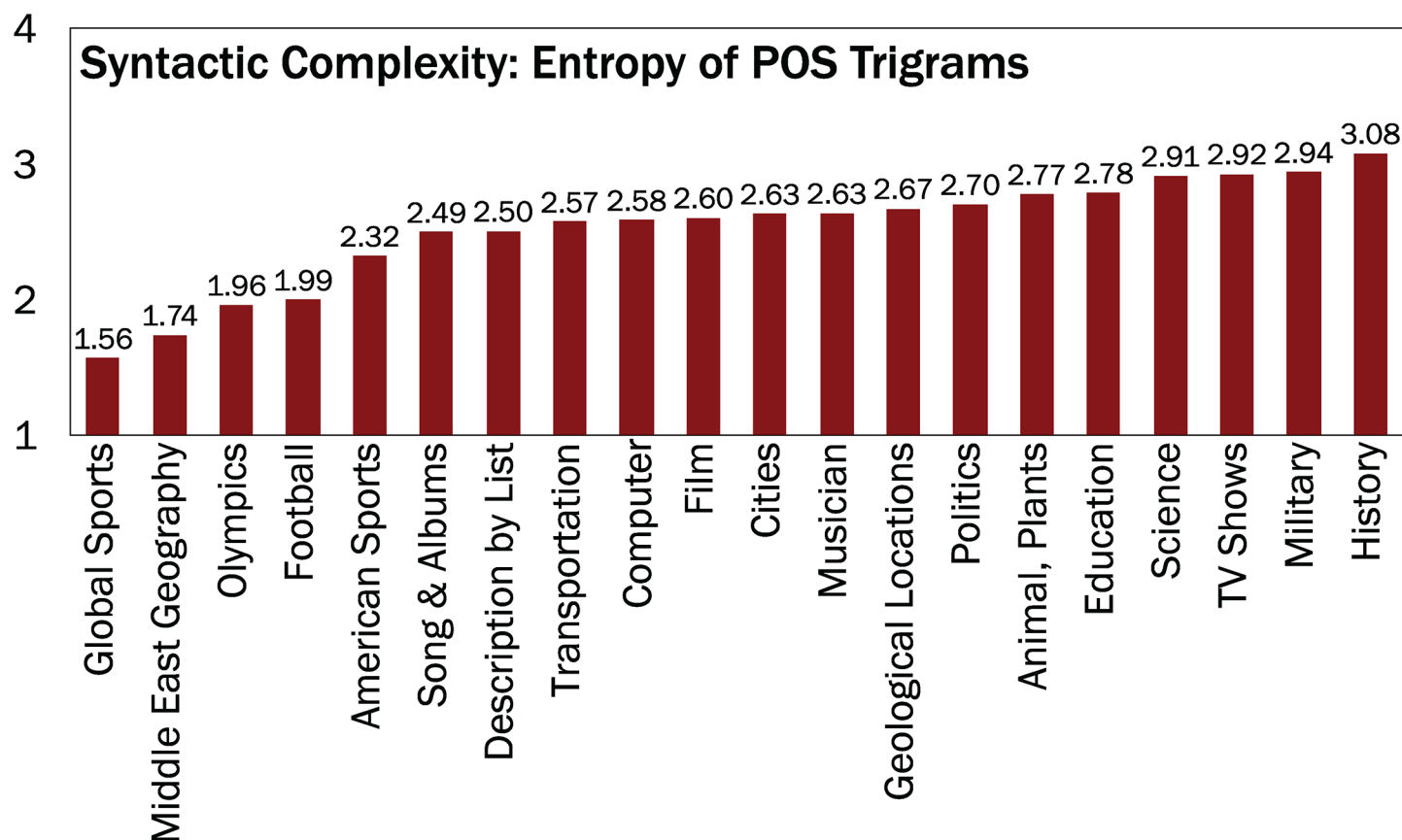


Fig 6. Language complexity varies by topic. Language complexity as measured by the entropy of POS trigrams varies by topic. Thus, we control for topic in order to measure language complexity more accurately.

doi:10.1371/journal.pone.0155305.g006

Natural Science has an average entropy value of 3.52 compared to 2.52 of *Soccer*. We do not observe such a pattern in the English edition. The results for other language editions are presented in [S4 Table](#).

Language Proficiency

Prior to computing the language complexity measures for edits, we control for the topics of the articles presented in the previous section, as prior research found that the language used in conceptual articles tend to be more complex than the language used in biographical and factual articles [18]. [Fig 6](#) shows that the complexity of language differs greatly by topic. Sports-related and other fact-oriented articles (e.g. *Football*, *Middle East Geography*, and *Olympics*) show lower language complexity than conceptual articles (e.g., *History*, *Education*, and *Science*). To control for topic, we calculate each of language complexity measures within a topic (intra-topic) and then average them over all topics (inter-topic).

Pre-Edit. We first examine the complexity of *Pre-Edits* to understand the text which multilingual editors choose to edit in their primary and non-primary languages.

[Fig 7a](#) shows that the entropy of unigrams, bigrams, and trigrams is always higher for the edits of multilingual users writing in English or Spanish as a primary language compared to those writing in English or Spanish as a non-primary language. The same is true for unigram

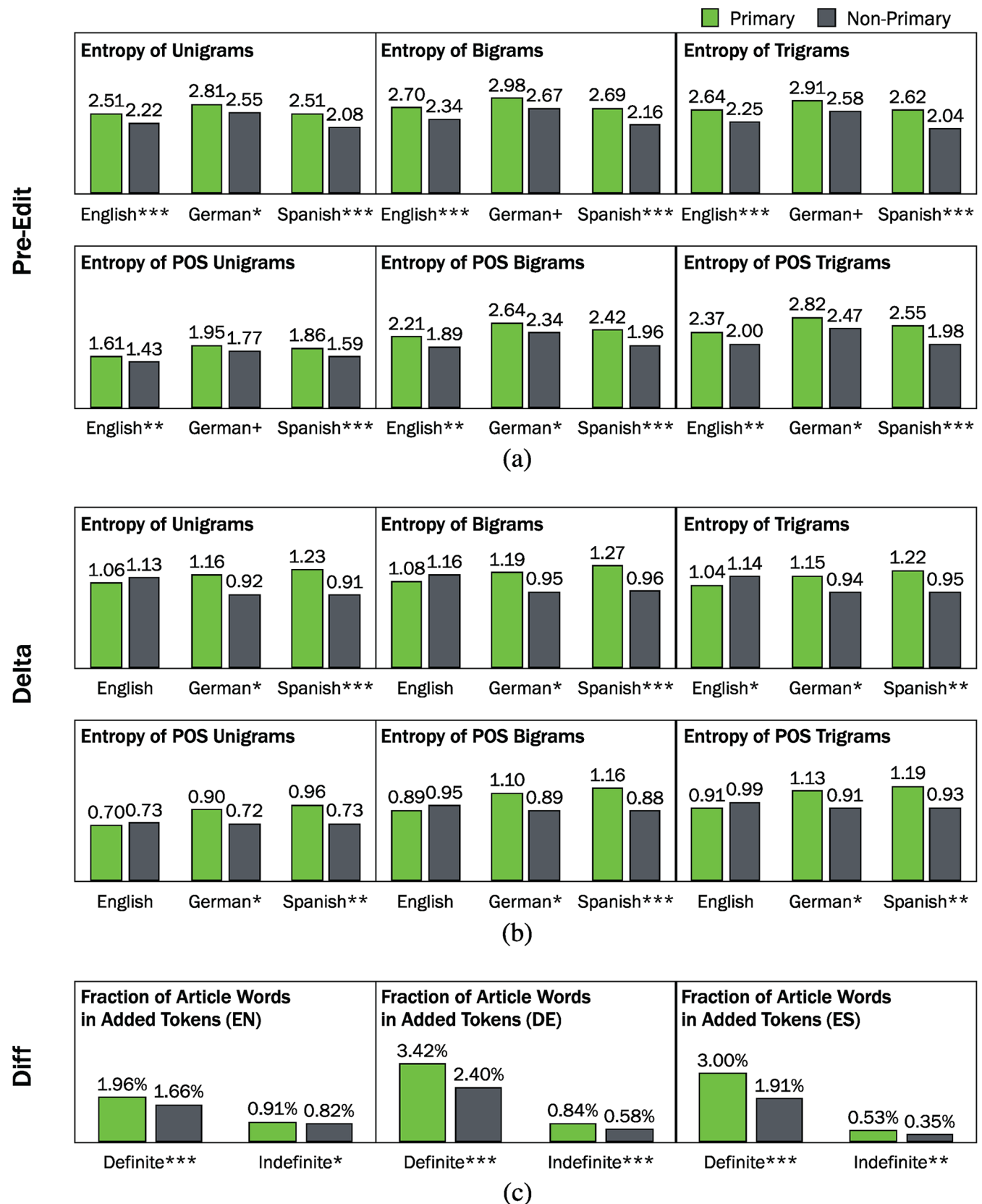


Fig 7. (a) Language complexity of pre-edits for primary and non-primary editors. Both lexical diversity measures and syntactic diversity measures show primary users edit more complex articles. **(b) The increase of lexical and syntactic diversity per paragraph per edit (delta).** The higher delta complexity scores for primary users indicate that multilinguals have higher linguistic abilities and make more complex edits in their primary languages. **(c) Fraction of article words in added tokens.** Primary editors use more article words, both definite and indefinite articles, than non-primary editors. (* $p < 0.08$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.).

doi:10.1371/journal.pone.0155305.g007

entropy in German, while bigram and trigram entropy show a similar pattern although the difference is not significant. In short, we observe a difference between primary and non-primary editors for all conditions, indicating that primary editors use more diverse terms while editing.

Moreover, we observe that the entropy of part-of-speech unigrams, bigrams, and trigrams is significantly higher for edits by primary editors in English and Spanish. Likewise, the POS entropy of edits is always higher for primary than non-primary editors in the German edition. The difference for bigrams and trigrams is significant while the difference in unigram POS entropy is not. This indicates that syntactic complexity also differs between primary and non-primary editors in all three languages in general. Thus, we conclude that primary editors edit more complex parts of articles compared to non-primary editors. In addition, the result implies that multilingual editors writing in a non-primary language may face a complexity barrier whereby they shy away from editing more complex sections of articles.

Delta. Delta complexity measures the difference in complexity before and after each article edit session and thus provides a measure of how the edits by the multilingual user changed the complexity of the article.

[Fig 7b](#) shows that there are significant differences between primary and non-primary groups, but that these differences are not consistent across the three language editions. Specifically, for the German and Spanish editions, we find the average increase in entropy between article edit sessions is higher from primary than non-primary editors. Surprisingly, however, we do not observe a significant difference between editors writing in English as a primary or non-primary language for the lexical diversity and syntactic complexity. Opposite to the other language editions, all entropy measures are slightly higher for non-primary editors of the English edition compared to primary editors.

These results implies that primary editors in German and Spanish editions possess higher linguistic proficiency, in terms of writing ability at least, than non-primary editors. This is consistent to the findings from the analysis of user interest. However, this does not apply to English, where there is no evidence that primary and non-primary editors of English possess different levels of linguistic proficiency.

Diff. [Fig 7c](#) shows that there are significant differences in the use of articles between by primary and non-primary editors. For all three language editions, primary editors use more articles—both definite and indefinite—than non-primary editors. Given that choosing the proper article is a difficult task—even proficient learners only have an accuracy of $\sim 90\%$ [\[44\]](#)—the difference in the use of articles between primary and non-primary users likely stems from different levels of proficiency. It is possible that non-primary editors simply skip using article words when confused as to which article to use.

Discussion

We looked at the editing behaviors of multilingual editors in Wikipedia, the world's largest general reference work. Other recent papers have sourced Wikipedia as a unique dataset of revision history that can be used to predict various collective opinions and actions [\[45–49\]](#).

That body of work is concentrated in a single language edition of Wikipedia, and they simply analyzed the numbers and patterns of revisions rather than the contents of the revisions. Our contributions from this perspective are that (i) we analyzed the revision history data in multiple language editions of Wikipedia, and (ii) we used text mining techniques with the contents of the revisions to conduct word-level syntactic and topical analysis.

This paper offers a first in-depth look at the behaviors of multilingual Wikipedia editors who are invaluable in delivering information across the language barrier. Previous research shows that multilingual editors' cultural background and language influence their perspective—how they view, interpret, and document world events [50, 51]. Moreover, there is selection bias (gatekeeping) in multilingual editors when choosing the articles to contribute, as well as different standards as to what is noteworthy or significant enough to have an article in different languages [52, 53]. Hence, it is important to analyze and understand their behaviors from multiple facets—including their topical interests and linguistic patterns. Surprisingly, we find that the overall distribution of interests is consistent between primary and non-primary users in each language edition with only a few exceptions. From German and Spanish Wikipedia editions, we observed that the topics with higher interest from primary users are relatively more complex than the topics with higher interest from non-primary users. However, we did not observe the same pattern in English.

With respect to linguistic complexity, we found that primary and non-primary editors differ in their behavior. Within the Spanish and German editions, we found that (i) primary users choose to edit more complex text than non-primary users (*pre-edit*), (ii) the edits of primary users result in a larger increase in the complexity of the articles than the edits of non-primary users (*delta*), and (iii) the content of the edits by primary users show greater language proficiency compared to the edits by non-primary users (*diff*, particularly the use of articles). The English edition of Wikipedia shows similar findings for *pre-edit* and *diff*. However, the results for the English edition in *delta* are markedly different from the Spanish and German editions. In this case, we found almost no difference in *delta* between primary and non-primary editors. The one exception was that the edits of non-primary users raised the complexity of articles slightly more than the edits of primary users.

These findings reinforce how strong of a barrier language is on online collaboration platforms. Even multilingual users who edit multiple editions of Wikipedia devote most of their efforts to editing the edition of their primary language—making more edits and spending more time within article edit sessions in their primary languages than in their non-primary languages. When multilingual users do edit in their non-primary languages, they often face a language complexity barrier. Users editing a non-primary language edition targeted their edits toward the grammatically simpler parts of the articles. In addition, with the exception of English, the edits that they made did not raise the linguistic complexity of the articles as much as the edits by multilingual users who primarily edited the language. This accords with linguistics research that multilinguals have differing levels of competency in their languages and that such differences are often related to how much they use each language [25, 54]. However, the contribution of editors in the English edition is a unique and noteworthy exception to the general pattern. This adds to research about the unique role English has online as a bridge between speakers of different languages [7, 12, 13, 15]. Non-primary editors of the English edition were able to contribute edits that were of equal complexity as those of primary editors. This ability to overcome the language barrier for English may help to partially mitigate the information asymmetries between English and other language editions as users contribute unique knowledge and information from their primary languages into the English edition. As the English edition is very central in the network of cross-language editing activity on Wikipedia [7], information may further flow from the English edition into other language editions [55].

A limitation of our study is that the metrics we used are language dependent (e.g., linguistic complexity based on POS tag entropy). Given this limitation, we compared primary and non-primary editors within each language, rather than across languages. If we can find better, language-independent metrics, it will be possible to compare editors across languages as well. Further, because our study relies on NLP tools, we were limited to the three languages, English, German, and Spanish, with the most accurate NLP tools. With advances and availability of NLP tools for other languages, it will be possible to expand this study and examine a larger variety of languages in the future. Another limitation in this study is that a multilingual editor's primary language is not always their native language. To quantify the degree of discrepancy, we checked the alignment between editors' primary languages and their self-declared native languages for the English edition by constructing another dataset of 221,162 editors who contributed to one or more articles in the English edition of Wikipedia appearing within the category "Wikipedia Controversial Topics." Among these editors, there are 18,962 users who disclosed their native languages on their user pages. Since the editors are crawled from the edit histories of the English edition of Wikipedia, we retain only the editors whose primary language or disclosed native language is English. We find that 94.4% of these editors (1,604 of 1,699) have the same primary language as the disclosed native language, indicating that the primary language (i.e., the language of the most edited edition) of a multilingual Wikipedia editor is a good proxy for the user's native language.

There remains a range of questions to tackle. There are many other online collaboration platforms beyond Wikipedia. In addition to the specific editing behaviors we studied in this paper, there are additional behavioral patterns that could be examined to further study the nature of cross-lingual knowledge diffusion and the contributions of multilingual users. Furthermore, to fully understand the behavior and the roles of multilingual users, the contributions of multilingual users should also be compared with those of monolingual users.

Supporting Information

S1 Table. Topic Clusters from the English Edition of Wikipedia. Examples of English Wikipedia article titles for discovered topics.
(PDF)

S2 Table. Topic Clusters from the Spanish Edition of Wikipedia. Examples of Spanish Wikipedia article titles for discovered topics (presented in English).
(PDF)

S3 Table. Topic Clusters from the German Edition of Wikipedia. Examples of German Wikipedia article titles for discovered topics (presented in English).
(PDF)

S4 Table. Syntactic Complexity for each Topic by Language Edition. Syntactic complexity for 20 English, Spanish and German topics and independent t-test results comparing contributions by primary and non-primary editors.
(PDF)

Acknowledgments

We thank Wikimedia foundation for providing access to the database on the Wikimedia Tool-server. We also thank the PLOS ONE reviewers and the associate editor for the detailed and insightful comments.

Author Contributions

Conceived and designed the experiments: Suin Kim SP Sooyoung Kim JB AHO. Performed the experiments: Suin Kim SP Sooyoung Kim JB. Analyzed the data: Suin Kim SP. Wrote the paper: Suin Kim SP SAH AHO.

References

1. Lewandowski D, Spree U. Ranking of Wikipedia articles in search engines revisited: Fair ranking for reasonable quality? *Journal of the American Society for Information Science and technology*. 2011; 62(1):117–132. doi: [10.1002/asi.21423](https://doi.org/10.1002/asi.21423)
2. WikiStats—Mediawiki Statistics; 2015.
3. Hecht B, Gergle D. The Tower of Babel meets Web 2.0: User-generated content and its applications in a multilingual context. In: *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM; 2010. p. 291–300.
4. Sen SW, Ford H, Musicant DR, Graham M, Keyes OS, Hecht B. Barriers to the localness of volunteered geographic information. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI 2015. ACM; 2015.
5. History of Wikipedia; 2015.
6. Danet B, Herring SC. *The multilingual Internet: Language, culture, and communication online*. Oxford University Press; 2007.
7. Hale SA. Multilinguals and Wikipedia Editing. In: *WebSci' 14*. ACM; 2014.
8. Ronen S, Gonçalves B, Hu KZ, Vespignani A, Pinker S, Hidalgo CA. Links that speak: The global language network and its association with global fame. *Proceedings of the National Academy of Sciences*. 2014; 111(52):E5616–E5622. doi: [10.1073/pnas.1410931111](https://doi.org/10.1073/pnas.1410931111)
9. Hecht BJ, Gergle D. On the localness of user-generated content. In: *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM; 2010. p. 229–232.
10. Cohen N. Hungry for New Content, Google Tries to Grow Its Own in Africa; 2015.
11. Kleinz T. Wikipedia erhält staatliche Förderung; 2007.
12. Herring SC, Paolillo JC, Ramos-Vielba I, Kouper I, Wright E, Stoerger S, et al. Language Networks on LiveJournal. In: *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*. HICSS'07. Washington, DC, USA: IEEE Computer Society; 2007.
13. Kim S, Weber I, Wei L, Oh A. Sociolinguistic analysis of twitter in multilingual societies. In: *Proceedings of the 25th ACM conference on Hypertext and social media*. ACM; 2014. p. 243–248.
14. Eleta I, Golbeck J. Bridging Languages in Social Networks: How Multilingual Users of Twitter Connect Language Communities. *Proceedings of the American Society for Information Science and Technology*. 2012; 49(1):1–4. doi: [10.1002/meet.14504901327](https://doi.org/10.1002/meet.14504901327)
15. Hale SA. Global Connectivity and Multilinguals in the Twitter Network. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI'14. New York, NY, USA: ACM; 2014. p. 833–842.
16. Hale SA. Net Increase? Cross-Lingual Linking in the Blogosphere. *Journal of Computer-Mediated Communication*. 2012; 17(2):135–151. doi: [10.1111/j.1083-6101.2011.01568.x](https://doi.org/10.1111/j.1083-6101.2011.01568.x)
17. Hale SA. Cross-language Wikipedia Editing of Okinawa, Japan. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI'15. New York, NY, USA: ACM; 2015. p. 183–192.
18. Yasseri T, Kornai A, Kertész J. A practical approach to language complexity: A Wikipedia case study. *PLoS ONE*. 2012; 7(11). doi: [10.1371/journal.pone.0048386](https://doi.org/10.1371/journal.pone.0048386)
19. Ortega F, Gonzalez-Barahona JM, Robles G. On the inequality of contributions to Wikipedia. In: *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*. IEEE; 2008. p. 304–304.
20. Iba T, Nemoto K, Peters B, Gloor PA. Analyzing the creative editing behavior of Wikipedia editors: Through dynamic social network analysis. *Procedia-Social and Behavioral Sciences*. 2010; 2(4):6441–6456. doi: [10.1016/j.sbspro.2010.04.054](https://doi.org/10.1016/j.sbspro.2010.04.054)
21. Lieberman MD, Lin J. You Are Where You Edit: Locating Wikipedia Contributors through Edit Histories. In: *ICWSM*; 2009.
22. Pavlenko A. *Emotions and Multilingualism*. Cambridge University Press; 2006.
23. Dewaele JM. *Emotions in multiple languages*. Palgrave Macmillan; 2010.

24. Barron-Hauwaert S. Bilingual: life and reality. *International Journal of Bilingual Education and Bilingualism*. 2011; 14(1):107–110. doi: [10.1080/13670050.2010.538192](https://doi.org/10.1080/13670050.2010.538192)
25. Haugen E. *The Norwegian Language in America: A Study in Bilingual Behavior*. Indiana University Press; 1969.
26. Geiger RS, Halfaker A. Using edit sessions to measure participation in Wikipedia. In: *CSCW 2013*; 2013.
27. *Understanding Editing Behaviors in Multilingual Wikipedia*; 2016.
28. Kittur A, Chi EH, Suh B. What's in Wikipedia?: Mapping Topics and Conflict Using Socially Annotated Category Structure. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI'09; 2009. p. 1509–1512.
29. Blei DM, Ng AY, Jordan MI, Lafferty J. Latent dirichlet allocation. *Journal of Machine Learning Research*. 2003; 3.
30. Sojka P. Software framework for topic modelling with large corpora. In: *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer; 2010.
31. Ester M, Krieger HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Kdd*. vol. 96; 1996. p. 226–231.
32. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011; 12:2825–2830.
33. Bulté B, Housen A. Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*. 2014; 26:42–65. doi: [10.1016/j.jslw.2014.09.005](https://doi.org/10.1016/j.jslw.2014.09.005)
34. Butler YG. Second Language Learners' Theories on the Use of English Articles. *Studies in second language acquisition*. 2002; 24(03):451–480. doi: [10.1017/S0272263102003042](https://doi.org/10.1017/S0272263102003042)
35. Ibanez MdPV, Ohtani A. Annotating article errors in Spanish learner texts: design and evaluation of an annotation scheme. 2014;.
36. Jaensch C. L3 acquisition of articles in German by native Japanese speakers. In: *Proceedings of the 9th Generative Approaches to Second Language Acquisition Conference (GASLA 2007)*. Somerville, MA: Cascadia Proceedings Project. vol. 8189; 2008.
37. Manning CD, Schütze H. *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press; 1999.
38. Brian Roark KH Margaret Mitchell. Syntactic complexity measures for detecting Mild Cognitive Impairment. *BioNLP'07*. 2007; p. 1–8.
39. Ratnaparkhi A, et al. A maximum entropy model for part-of-speech tagging. In: *Proceedings of the conference on empirical methods in natural language processing*. vol. 1; 1996. p. 133–142.
40. Marcus MP, Santorini B, Marcinkiewicz MA. *Building a Large Annotated Corpus of English: The Penn Treebank*. Computational Linguistics. 1993;.
41. Toutanova K, Klein D, Manning CD, Singer Y. Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network. *NAACL'03*; 2003.
42. Skut W, Brants T, Krenn B, Uszkoreit H. A linguistically interpreted corpus of German newspaper text. In: *the ESSLLI Workshop on Recent Advances in Corpus Annotation*; 1998.
43. Taulé M, Martí MA, Recasens M, Computació CDLI. Ancora: Multilevel annotated corpora for Catalan and Spanish. In: *6th International Conference on Language Resources and Evaluation*; 2008.
44. McEnery T, Xiao R, Tono Y. *Corpus-based language studies: An advanced resource book*. Taylor & Francis; 2006.
45. Mestyán M, Yasserli T, Kertész J. Early prediction of movie box office success based on Wikipedia activity big data. *PloS one*. 2013; 8(8). doi: [10.1371/journal.pone.0071226](https://doi.org/10.1371/journal.pone.0071226) PMID: [23990938](https://pubmed.ncbi.nlm.nih.gov/23990938/)
46. Moat HS, Curme C, Avakian A, Kenett DY, Stanley HE, Preis T. Quantifying Wikipedia usage patterns before stock market moves. *Scientific reports*. 2013; 3. doi: [10.1038/srep01801](https://doi.org/10.1038/srep01801)
47. Curme C, Preis T, Stanley HE, Moat HS. Quantifying the semantics of search behavior before stock market moves. *Proceedings of the National Academy of Sciences*. 2014; 111(32):11600–11605. doi: [10.1073/pnas.1324054111](https://doi.org/10.1073/pnas.1324054111)
48. Masucci AP, Kalampokis A, Eguíluz VM, Hernández-García E. Wikipedia information flow analysis reveals the scale-free architecture of the semantic space. *PloS one*. 2011; 6(2). doi: [10.1371/journal.pone.0017333](https://doi.org/10.1371/journal.pone.0017333)
49. Keegan B, Gergle D, Contractor N. Hot off the wiki: Structures and dynamics of Wikipedia's coverage of breaking news events. *American Behavioral Scientist*. 2013; doi: [10.1177/0002764212469367](https://doi.org/10.1177/0002764212469367)
50. Fichman P, Hara N. *Global Wikipedia: International and cross-cultural issues in online collaboration*. Rowman & Littlefield Publishers, Inc.; 2014.

51. Kolbitsch J, Maurer HA. The Transformation of the Web: How Emerging Communities Shape the Information we Consume. *J UCS*. 2006; 12(2):187–213.
52. Callahan ES, Herring SC. Cultural bias in Wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology*. 2011; 62(10):1899–1915. doi: [10.1002/asi.21577](https://doi.org/10.1002/asi.21577)
53. Hautasaari A, Ishida T. Analysis of Discussion Contributions in Translated Wikipedia Articles. In: *Proceedings of the 4th International Conference on Intercultural Collaboration*. ICIC'12; 2012. p. 57–66.
54. Grosjean F. *Bilingual: Life and Reality*. Harvard University Press; 2010.
55. Warncke-Wang M, Uduwage A, Dong Z, Riedl J. In Search of the ur-Wikipedia: Universality, Similarity, and Translation in the Wikipedia Inter-language Link Network. In: *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*. WikiSym'12. New York, NY, USA; 2012. p. 20:1–20:10.