Analysis of Editors' Languages in Wikidata

Lucie-Aimée Kaffee
University of Southampton, UK
kaffee@soton.ac.uk

Elena Simperl
University of Southampton, UK
E.Simperl@soton.ac.uk

ABSTRACT

Wikidata is unique as a knowledge base as well as a community given its users contribute together to one cross-lingual project. To create a truly multilingual knowledge base, a variety of languages of contributors is needed. In this paper, we investigate the language distribution in Wikidata's editors, how it relates to Wikidata's content and the users' label editing. This gives us an insight into its community that can help supporting users working on multilingual projects.

KEYWORDS

Multilinguality, Wikidata, Community

ACM Reference format:
Lucie-Aimée Kaffee and Elena Simperl. 2018. Analysis of Editors' Languages in Wikidata. In *Proceedings of ACM OpenSym, Paris, France, 2018 (OpenSym 2018),* 5 pages.
https://doi.org/10.1145/3233391.3233965

1 INTRODUCTION

Wikidata is a collaboratively edited knowledge base [10]. In Wikidata users can edit in their respective languages in the same project. The community of Wikidata is diverse, as editing is not limited to one language. To understand how a multilingual community collaborates in the case of a knowledge base, we need to gain an understanding of the current community.

Wikidata's content is language independent. This means, each concept or *item* is identified by an ID that does not contain natural language but rather a number such as Q12345. Each item can be connected to a label in up to 410 languages. Labels are the way for humans to access the data – they express the concept in a natural language. Therefore, when we talk about multilingual content in Wikidata, we are usually talking about the labels that users edit.

OpenSym 2018, 2018, Paris, France

ACM ISBN 978-1-4503-5936-8/18/08.
https://doi.org/10.1145/3233391.3233965

While the aim is to build a global knowledge base, the language content of Wikidata is not equally distributed [6]. With the goal of a better language coverage of a collaborative knowledge base, we need to understand the editors who contribute the data. Therefore, we analyzed the languages of editors and their relationship to Wikidata's labels. If we can understand how editors with certain language skills edit Wikidata, we can build more efficient systems to support them in doing so.

In this exploratory study, we investigate three aspects of Wikidata editors in regards to languages: i) the multilinguality of Wikidata editors, ii) the correlation of editors' languages with existing labels in the language and iii) editors' editing of labels in Wikidata.

In order to gain an understanding of Wikidata's editors, we extract information from user pages and Wikidata's editing history. We find that users are likely to know multiple languages. The known languages correlate with the information already existing in Wikidata. However, Wikidata's users also edit labels in languages other than their own. These results can lead to a more in-depth understanding of the global community of a multilingual knowledge base.

All scripts used for this paper are available on GitHub: https://github.com/luciekaffee/Wikidata-User-Languages.

Figure 1: Example for a BabelBox of User:Frimelle

2 RELATED WORK

Multilingual structured data has been investigated in [3, 4, 7]. However, none of these take into account that a knowledge base can be built by a community. A community editing such a project can have a fundamental impact on its coverage and development, though.

Creating a resource collectively is an approach also used in crowdsourcing. Payed crowdsourcing to construct a multilingual resource has been investigated by [8]. However, Wikidata's users can be better compared to the editors of Wikipedia than crowdsourcing, as they have different incentives to contribute. Roles of editors have been investigated especially in the context of Wikipedia, e.g. in the form of *social roles* [11]. We focus on what users edit, similar to [2], who classify edits throughout the edit history.

Wikipedia projects only cover one language per project, and they vary widely in size and coverage of topics [5]. Similarly, we could find in [6] a bias towards English in the content of Wikidata, that needs to be addressed in order to make Wikidata a truly multilingual knowledge base. One of the main ways of address this maldistribution is through the community of editors. We use the *BabelBox* to understand the work of this multilingual community. BabelBoxes have been used in previous work to detect non-native speakers in English Wikipedia [1]. Wikidata compared to Wikipedia has a multilingual community editing collectively one cross-lingual project. Wikidata's community is unique in its multilinguality in the Wikimedia universe. An investigation will give us new insights on multilingual communities which is a good starting point for further work.

3 METHODS

In the following we introduce how we accessed users' languages and their respective edits to Wikidata labels.

User Language Setting

Every registered user of Wikidata can change the language of the interface. Not only interface elements are switched to their preferred language, but also all data displayed. This setting is called *User Language Setting*[1]. We extracted the aggregated number of user's languages via Wikimedia's Grafana installation[2]. The data can be downloaded as JSON. We use the user language setting as a starting point for our exploration to understand the language distribution in Wikidata users. However, English is the default language, skewing results towards this language in the data. As only one language can be set, it does not give any indication on the possible multilinguality of users.

BabelBox

Each registered user in Wikidata has a user page, on which they can add and edit information on themselves. One of the templates that can be added indicates languages spoken: *BabelBox*[3]. BabelBox lets a user self-assess their languages and the respective level of language from 0 (no knowledge) to N (native understanding). When we talk in the following of a *known language*, we usually refer to all levels excluding level 0 and 1. Accordingly, *unknown languages* are defined as all languages that are undeclared as we assume that the user has no knowledge of these languages. In order to access the user pages, we download the complete Wikidata dump[4] and extract user pages with BabelBoxes. As Wikimedia projects are connected, users can also have a so called global userpage[5], that we take into account. We process all user pages with a BabelBox in order to collect data. In total, 4,120 users have a BabelBox enabled on Wikidata or Meta. Wikidata has 19,333 active user (2,930,072 total registered users). Even though we treat those users as a sample of all users for our exploration of multilinguality in Wikidata users and their editing, there are clear limitations to this approach: Our sample is not truly random. We can assume that a certain type of users is more likely to enable BabelBoxes than other users. The close connection of Wikipedia and Wikidata users [9] lets us assume that there might be users from projects, that are more likely to enable the BabelBox (this could be an explanation e.g. for the high number of native German users). Furthermore, multilingual users might be more likely to enable a BabelBox than monolingual users.

Edit History

In Wikidata, similar to Wikipedia, each item contains not only the information displayed, but also a summary of all changes ever done to this item, the *edit history*. From the complete editing history provided by Wikidata, we extracted all edits to labels, marked with the keywords `wbsetlabel-set` and `wbsetlabel-add`. We build a database based on this editing history, where each user is connected to their respective edit of a label, and in which language the label was changed. For each user, we aggregated all edits and extracted the numbers of edits a user did in their known languages compared to the languages not known to them based on their BabelBox.

[1] https://www.wikidata.org/wiki/Help:Navigating_Wikidata/User_Options#Language_settings

[2] Active User Language on Grafana: https://grafana.wikimedia.org/dashboard/db/wikidata-site-stats?orgId=1

[3] https://en.wikipedia.org/wiki/Template:Babel

[4] https://dumps.wikimedia.org/wikidatawiki/

[5] https://meta.wikimedia.org/wiki/Global_user_pages

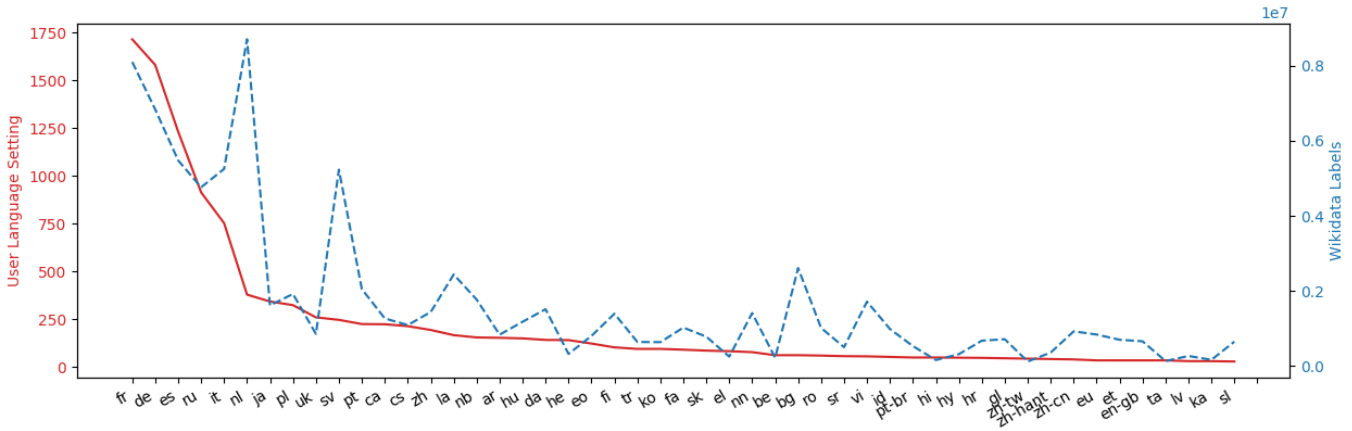

Figure 2: User Language Setting compared to Labels in Wikidata (excluding English), top 50 languages

Comparision to Wikidata's Multilingual Content

To understand whether the number of users speaking a language correlates with the number of labels in a language, we calculated the correlation coefficient using python's numpy. We excluded the languages that have a lower level than 2 in the Wikidata users. We calculated the correlation coefficient between the number of users speaking the language and the labels available in the language for the total 298 languages known to users. For the labels we use the data previously extracted in [6].

4 RESULTS & DISCUSSION

User Language Setting

The user language setting can be set by a registered user to change the language of the content displayed on the website. As the default language is English, English is set by over 50% of the user. Excluding English as in Figure 2, we get a more interesting overview of the multilinguality of Wikidata users: The most prominent language is French, followed by German, Spanish and Russian. Overlap between distributions of user languages and Wikidata content is much lower than for the BabelBox (Figure 4). It supports the assumption it is not the ideal way to measure editors languages, and a system similar to BabelBoxes gives a more comprehensive insight into editors' languages.

BabelBox

BabelBoxes are used to indicate user languages on a users page. Compared to users enabling the User Language Setting, BabelBox users are more active in editing labels. The mean of edits of babelbox users is 55 compared to 2 for non-BabelBox users with at least one edit (5440.16 vs 907.41 in average)). On Wikidata, 4120 users enabeled a BabelBox.

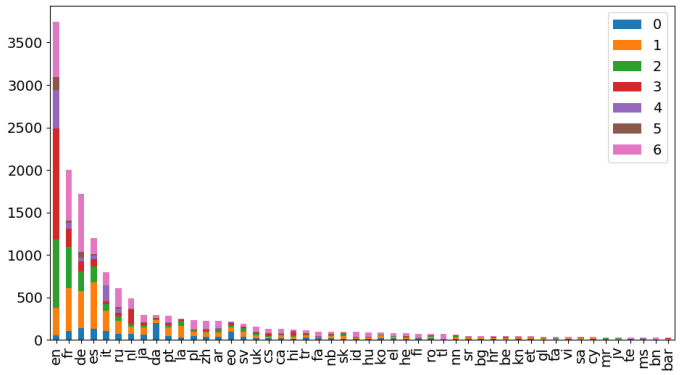

Figure 3: Top 50 languages of BabelBox users, split in the language levels, 6 is native

As can bee seen in Table 1, most user are multilingual. Users know between 1 and, for one user, 47 languages. In total, BabelBox users speak 298 different languages at least at level 2. Figure 3 shows the distribution of languages that enabled the BabelBox. By far, the most used language is English. English also has the least amount of speakers of level 0 or 1 between the four most common languages. Corresponding, to the User Language Setting, the four most prominent languages are English, German, French, and Spanish. English being the most spoken language contributes probably to the fact that currently most community discussions on Wikdiata are in English.

Learning about the languages of users and the fact that most users are multilingual is interesting as it indicates there is a community with the resources to translate data between different languages.

Wikidata Labels

To understand the relationship of editors and content, we compared BabelBox languages to distribution of languages

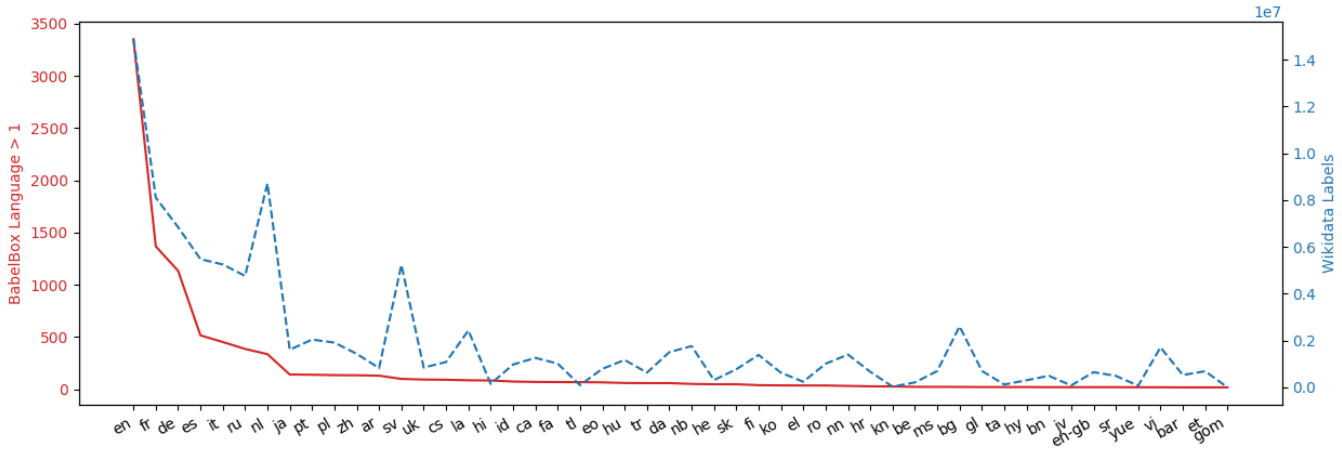

Figure 4: Language Distribution in BabelBox, excluding language levels 0 and 1, top 50 languages

Description	% User
Monolingual	11.4%
≤ 3 Languages	58.3%
≤ 5 Languages	84.2%
> 5 Languages	15.7%

Table 1: Distribution of multilingual users

in Wikidata labels. As [6] show, Wikidata labels are not equally distributed between languages, with English being the most well-covered language. As visible in Figure 4, there is an overlap of the most prominent languages regarding labels and the languages spoken by the community – English, German and French. The correlation coefficient between languages spoken by the community and Wikidata labels is **0.8979**. Languages spoken by the community also reflect in the data present on Wikidata, meaning diversification of the community can bring an increase of language coverage.

Language Editing

Based on our results from extracting BabelBoxes, we investigate the editing of labels on Wikidata by users. 1,107 users of the BabelBox users do not have label edits in Wikidata. Therefore, we are working with 3,013 users. Most of these users (3,007) are in the square we emphasize in Figure 5. In the Figure 5, each user is a point in the coordinate system, where the axis describe the nominal number of edits in a language they state to know, or that is unknown to them. We find that the majority of users edits mainly in a language they know. Generally, edits to known languages are higher in numbers than to unknown languages. However, it is common to have at least a few edits in an unknown language. Especially user with high numbers of edits in labels are likely to contribute to languages they do not know. Furthermore, there are a few interesting outliers: Editors that edit more in unknown languages than in their own languages. This indicates that editing language information in Wikidata is not very challenging – e.g. names in Latin languages can usually be transferred between different languages. Those cases should be investigated further, as such tasks can easily be automated once identified.

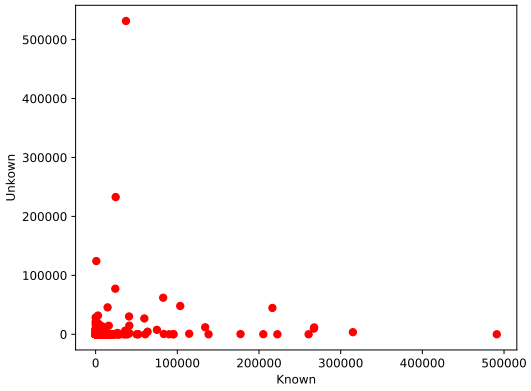

Figure 5: Users plotted according to their label edits in languages they are familiar with (known) and unknown languages, excluding most extreme outliers

5 CONCLUSION & FUTURE WORK

We analyzed Wikidata's users in regard to their languages and their label editing. This clears the way for future work on Wikidata's multilingual data with an insight into its multilingual community. We saw that the community is multilingual, language content correlates with the languages of the community and editors edit mostly in their languages, but extend their activity to unknown languages as well. Identifying these cases in detail can help with automation of tasks

and support of the editors. To support such automation, analyzing the languages and how they are related further will bring an insight of what can be done by bots.

ACKNOWLEDGEMENTS

This research is partially supported by the Answering Questions using Web Data (WDAqua) project, a Marie Skłodowska-Curie Innovative Training Network under grant agreement No 642795, part of the Horizon 2020 programme.

REFERENCES

[1] Yanging Chen, Rami Al-Rfou, and Yejin Choi. 2017. Detecting English Writing Styles For Non Native Speakers. *CoRR* abs/1704.07441 (2017). arXiv:1704.07441 http://arxiv.org/abs/1704.07441

[2] Peter Kin-Fong Fong and Robert P. Biuk-Aghai. 2010. What did they do?: deriving high-level edit histories in Wikis. In *Proceedings of the 6th International Symposium on Wikis and Open Collaboration, 2010, Gdansk, Poland, July 7-9, 2010.*

[3] Asunción Gómez-Pérez, Daniel Vila-Suero, Elena Montiel-Ponsoda, Jorge Gracia, and Guadalupe Aguado de Cea. 2013. Guidelines for multilingual linked data. In *3rd International Conference on Web Intelligence, Mining and Semantics, WIMS '13, Madrid, Spain, June 12-14, 2013.* 3.

[4] Jorge Gracia, Elena Montiel-Ponsoda, Philipp Cimiano, Asunción Gómez-Pérez, Paul Buitelaar, and John P. McCrae. 2012. Challenges for the multilingual Web of Data. *J. Web Sem.* 11 (2012), 63–71.

[5] Brent J. Hecht and Darren Gergle. 2010. The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, Atlanta, Georgia, USA, April 10-15, 2010.* 291–300.

[6] Lucie-Aimée Kaffee, Alessandro Piscopo, Pavlos Vougiouklis, Elena Simperl, Leslie Carr, and Lydia Pintscher. 2017. A Glimpse into Babel: An Analysis of Multilinguality in Wikidata. In *Proceedings of the 13th International Symposium on Open Collaboration, OpenSym 2017, Galway, Ireland, August 23-25, 2017.* 14:1–14:5.

[7] John P. McCrae, Steven Moran, Sebastian Hellmann, and Martin Brümmer. 2015. Multilingual linked data. *Semantic Web* 6, 4 (2015), 315–317.

[8] Matteo Negri, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo, and Alessandro Marchetti. 2011. Divide and Conquer: Crowdsourcing the Creation of Cross-Lingual Textual Entailment Corpora. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL.* 670–679.

[9] Alessandro Piscopo, Christopher Phethean, and Elena Simperl. 2017. Wikidatians are Born: Paths to Full Participation in a Collaborative Structured Knowledge Base. In *50th Hawaii International Conference on System Sciences, HICSS 2017, Hilton Waikoloa Village, Hawaii, USA, January 4-7, 2017.*

[10] Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

[11] Howard T. Welser, Dan Cosley, Gueorgi Kossinets, Austin Lin, Fedor Dokshin, Geri Gay, and Marc A. Smith. 2011. Finding social roles in Wikipedia. In *iConference 2011, Inspiration, Integrity, and Intrepidity, Seattle, Washington, USA, February 8-11, 2011.* 122–129.