A Glimpse into Babel: An Analysis of Multilinguality in Wikidata

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

Alessandro Piscopo
University of Southampton
Southampton, UK
A.Piscopo@soton.ac.uk

Pavlos Vougiouklis
University of Southampton
Southampton, UK
pv1e13@ecs.soton.ac.uk

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

Leslie Carr
University of Southampton
Southampton, UK
lac@ecs.soton.ac.uk

Lydia Pintscher
Wikimedia Deutschland
Berlin, Germany
lydia.pintscher@wikimedia.de

ABSTRACT

Multilinguality is an important topic for knowledge bases, especially Wikidata, that was build to serve the multilingual requirements of an international community. Its labels are the way for humans to interact with the data. In this paper, we explore the state of languages in Wikidata as of now, especially in regard to its ontology, and the relationship to Wikipedia. Furthermore, we set the multilinguality of Wikidata in the context of the real world by comparing it to the distribution of native speakers. We find an existing language maldistribution, which is less urgent in the ontology, and promising results for future improvements.

Author Keywords

Multilinguality; Wikidata; Community-driven knowledge base; Linked Data

INTRODUCTION

Wikidata is a community-driven knowledge base created as a central knowledge store for Wikimedia projects such as Wikipedia. It is now widely used in third-party applications as well. It contains linked data in the RDF format and can be queried via a SPARQL endpoint. Items in Wikidata are concepts such as people, places, or events and usually subjects in triples. Every data item in Wikidata is language-independent and connected to labels in multiple languages. Multilingual labels are important for linked data since they make the data human-accessible and therefore reusable. One example of this interaction with multilingual data are Question Answering systems that allow users an easy interaction with complex datasets. Language barriers and a lack of language diversity prevent whole communities from accessing information and contributing to more diverse knowledge online.

In this paper we look at the languages covered by Wikidata. Wikidata has been designed as an inherently multilingual project. This feature was intended to allow users to express different points of view and foster the expression of knowledge diversity. Understanding Wikidata's language coverage is crucial to understand whether this goal has been met.
Our research questions are as follows:

RQ1 *What is the state of Wikidata with regard to multilinguality?*
RQ2 *Is there a difference in the multilinguality of the ontology, compared to the overall multilinguality of the knowledge base?*
RQ3 *How does Wikidata's label distribution relate to the real world and Wikipedia's language distribution?*

In order to answer these questions, we investigate Wikidata labels and analyze their distribution. To improve language diversity on the Semantic Web, including Wikidata, we must first understand the current state. In the following section we introduction some background and related work about language diversity in the Semantic Web and on Wikidata.

BACKGROUND AND RELATED WORK

Multilinguality of structured data is a very important topic for the Semantic Web, because labels are the access point for humans to interact with the data [5]. The use cases for multilingual data are diverse: for humans to understand the information, natural language multilingual data is necessary. Consequently, there is a strong relationship between Semantic Web and natural language processing [4]. A well established and comprehensive knowledge base that includes labels in a large number of languages might serve as a base for applications interacting with humans via natural language, including chat bots [16].
Ell *et al.* in [5] evaluate labels on the Web of Data, using metrics to measure their completeness, efficient accessibility, unambiguity, and multilinguality. They draw the conclusion that there is still a big lack of multilingual information on the

OpenSym '17 August 23–25, 2017, Galway, Ireland

ACM ISBN 978-1-4503-5187-4/17/08.
DOI: https://doi.org/10.1145/3125433.3125465

Web of Data, which needs to be overcome. The resource the authors use is the Billion Triple Challenge[1], a static knowledge base crawled from the web. The knowledge base we are looking at is more dynamic, because it has a community that is actively contributing and changing the content. This means that suggestions on how to increase the coverage could be implemented in later work. To gain a better understanding of the mechanisms of the community and the coverage of the multilingual data, we compare the language distribution of the ontology (i.e. Wikidata properties) with that of the content overall. Furthermore, we set our work in the context outside the world of the Semantic Web to understand how well covered real world language communities are.

Similar work was done for Wikipedia: [6] analyzes 25 Wikipedia language versions in order to understand the overlap between the versions in regard to topic and content. Wikipedia's language distribution in relation to the global distribution of first and second languages spoken is analyzed in [13], and shows similar patterns to Wikidata. Wikipedia's multilinguality is used in the context of natural language processing, e.g. for topic mining in [11].

A possible use of multilingual data is Question Answering over Linked Data, as in [1, 7]. [14] investigates how multilingual ontologies can facilitate question answering. Similarly, [3, 9] look at how to enable multilinguality for ontologies.

MULTILINGUALITY IN WIKIDATA

Wikidata provides a language-independent ontology. That is, each concept in Wikidata, called an *item*, has a unique identifier consisting of a letter and a number; e.g. `Q12345`. The same is done for properties, which are marked with a `P` as in `P123`. Each item or property (*entity*) can be connected to a natural language label. The predicate used is `rdfs:label`; in accordance with W3C standards[2], the object's language is marked with the tag `@<language code>`. The triple to express a label looks like this: `Q12345 rdfs:label "Count von Count"@en`. The community contributes to every part of the data set, including natural language labels. Labels of items are often imported from Wikipedia or added via assistive tools. An example for such a tool can be bots, user written scripts that usually perform monotonous tasks or the *Wikidata Terminator*[3], which encourages users to translate the most frequently used items. Wikidata maintains the links between different language versions of Wikipedia and other Wikimedia projects, which means that many items are connected to a Wikipedia article. Titles of the connected articles are often imported as labels for the respective Wikidata item. Properties have a special position in Wikidata's ontology. Classes are not structurally distinguishable from other items, only implicit by their content. Properties on the other hand are easily distinguishable by their distinct identifiers. The community process to create a new property is also more complex than that for creating a new item. However, translating a property is just as easy as translating an item's label. Properties are predicates in triples, so they are frequently used and highly visible to the community. Additionally, there are currently only 3,386 properties, while there are close to 26 million items. Consequently, it could be hypothesized that language coverage should be higher in properties.

Wherever Wikidata's data is used in Wikipedia, it displays the label of the entity. One example are infoboxes, a summary of information on an article in Wikipedia. Those can reuse the data of Wikidata. Another example of Wikidata language information used in Wikipedia are ArticlePlaceholders [8], which generate an overview on a topic with data provided by Wikidata. Therefore there is a strong interest in improving the coverage of languages given its impact in, among others, Wikipedia.

METHODS

Wikidata's multilinguality is one of the core values of the project. In order to understand the state of languages in Wikidata as of now (**RQ1**), we looked at its entity labels. Thus we analyzed a database dump of Wikidata in turtle format from March 2017, to count all labels that are noted to be in a certain language (e.g. `@en` tag in the case of English). We analyzed a total of 134M labels of Wikidata's 26M entities plus three thousand properties. A large amount of information is redundant in Wikidata's turtle format to cover multiple ontologies such as *skos* and *schema*. Therefore we only considered `rdfs:label` and left out other representations such as `skos:prefLabel` and `schema:name`. We also exclude other strings that are in natural language, such as the value or object for statements containing `P1477` (*birth name*). `rdfs:label` is used to label both items and properties. Based on these numbers, we calculated percentages for each language, and looked at the distribution over all available languages.

To understand the language distribution of the ontology, as in **RQ2**, we assessed property labels via a SPARQL query to the Wikidata query endpoint[4]. Due to the structure of Wikidata, and classes being implicit, we focused on properties as an indicator for the state of the translation of the ontology.

To understand how diverse the language distribution of Wikidata is, we compared it with the native speakers in the world and Wikipedia, the widely used online encyclopedia in order to understand potential relationships between the distributions of native speaker and Wikidata's languages as well as Wikidata's and Wikipedia's biggest languages. We first noted the Wikipedia language codes for the top 100 languages by number of native speakers as of [12]. The distribution of native language speakers is compared with the labels in Wikidata to see how well the language communities are covered by human-readable knowledge in Wikidata in order to answer **RQ3**. Additionally, we used the ranking of Wikipedias based on the numbers of articles for each language version. We compared this to the ranking by label count, to see whether we can find a similar pattern of languages and get an insight into the relation between Wikipedia's and Wikidata's multilingual information.

DISCUSSION OF THE RESULTS

We analyzed the distribution of languages in Wikidata to answer the question of the state of multilinguality in Wikidata.

[1] http://km.aifb.kit.edu/projects/btc-2010/
[2] https://www.w3.org/2007/02/turtle/primer/
[3] https://tools.wmflabs.org/wikidata-terminator/
[4] http://tinyurl.com/khxu5x7

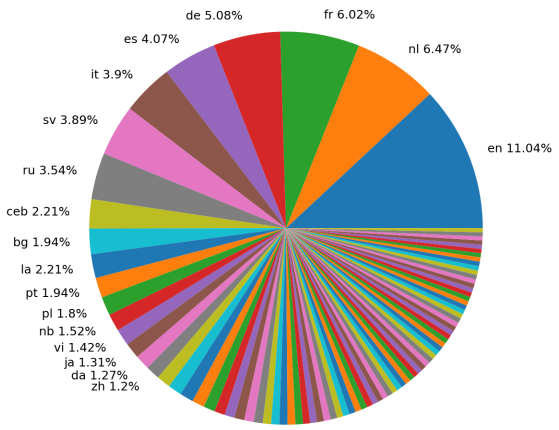

Figure 1. Percentage of all labels per language in Wikidata

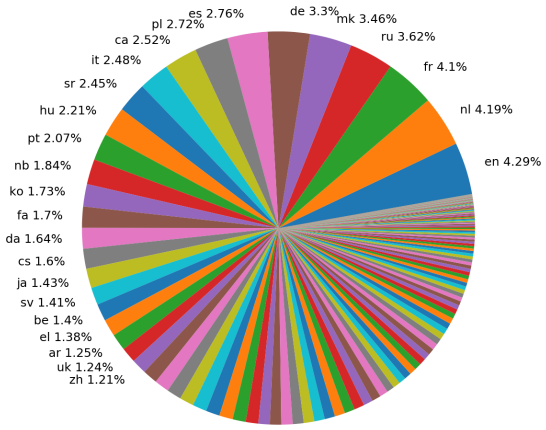

Figure 2. Distribution of languages for properties in Wikidata

RQ1 *What is the state of Wikidata with regards to multilinguality?*

As an orientation, we looked into the state of languages on the web in general. English is the language of around 51.9% of all websites[5] even though t is spoken only by 25% of the world population. Chinese is the second largest language in terms of users on the web, but only 2% of the content on the web is in Chinese [10]. Only 11 languages hold almost 50% of all language knowledge in Wikidata, as evident in Figure 1, indicating the Semantic Web has similar problems of language maldistribution. A few languages hold most of the content, while the vast majority have relatively few labels. The language holding most content is English. However, the distribution is not as extremely homogeneous in linked data as on the web in general. Therefore, linked data could offer part of the solution to the language inequality on the web.

RQ2 *Is there a difference in the multilinguality of the ontology, compared to the overall multilinguality of the knowledge base?*

The distribution of translated labels of properties in Figure 2 supports the idea that the state of languages online could be improved by linked data. It is especially important that the ontology is translated. Therefore, the more diverse distribution for properties is promising. Properties are used widely across Wikidata; therefore, it is more likely for missing translations to be detected by the community. Consequently, the distribution of languages is much less extreme; while English is still the leading language, the margin is very narrow. English has a share of 4.29% in the distribution of property labels in all of Wikidata's languages, followed by Dutch with 4.19%. The comparison of all of Wikidata's labels to only the labels of properties in Figure 3 make the distribution of labels over languages even more evident. The median for property labels is lower but the interquartile range is wider, indicating a higher variance of the data.

[5] https://w3techs.com/technologies/overview/content_language/all

RQ3 *How does Wikidata's label distribution relate to the real world and Wikipedia's language distribution?*

The comparison of Wikidata's label distribution with the distribution of first language speakers in the world, in Figure 4, shows that there is still a lot of work to be done to cover all language communities. Most evident here is the case of Chinese. There are multiple Chinese versions and there are issues in interpreting the data due to the censorship of Wikipedia in China [2], which result in a big share of the edits being made from outside China[6]. However, the biggest Chinese version, with the language code `zh`, is still very under-served, especially given the number of people speaking the language. Examples such as Dutch (`nl`) or Cebuano (`ceb`) show that it is not strictly necessary for a language to be spoken by many people to have good coverage in a knowledge base.

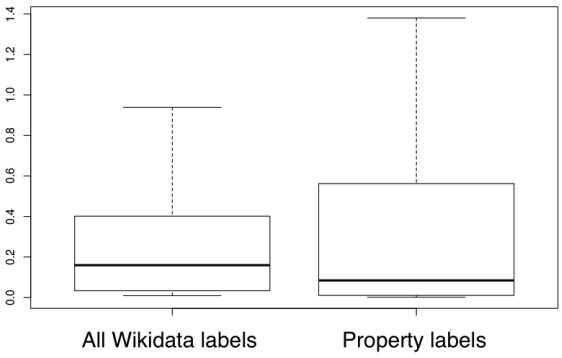

Figure 3. Boxplot of labels of all entities and labels of only the properties and their percental distribution over different languages

Dutch and Cebuano are especially interesting with regards to the relationship of Wikipedia and Wikidata. As found by [15], many members of the Wikidata community derive from

[6] https://en.wikipedia.org/wiki/Chinese_Wikipedia#Origin_of_edits

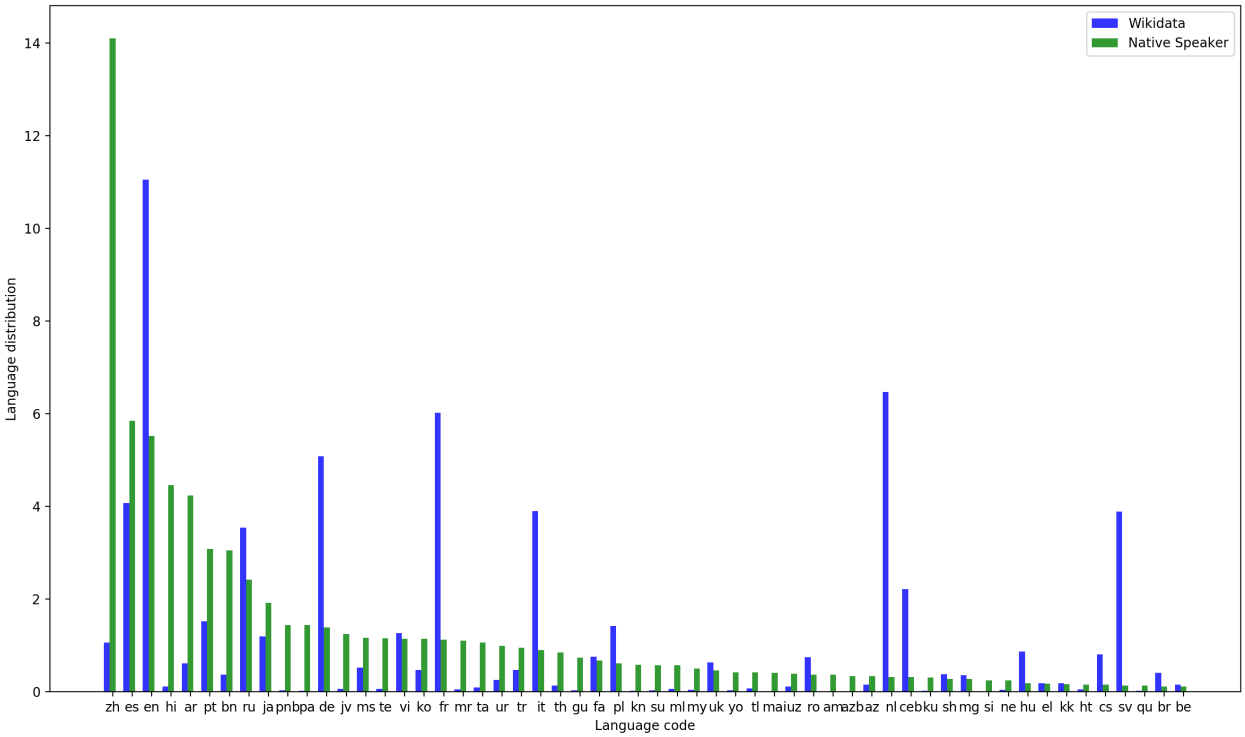

Figure 4. Comparison of distribution of languages in Wikidata and first language speakers in the world

the Wikipedia community. Dutch can thus be assumed to have very active communities on both Wikipedia and Wikidata. However, both Cebuano and Swedish (`sv`) have one contributor, who works on a bot called `lsvbot`; this is an automated tool which adds new articles to these Wikipedias. With or without such special circumstances, a connection in the ranking of biggest language versions is apparent, in terms of articles and labels for Wikipedia and Wikidata respectively. This is illustrated in Table 1. It reflects the import of Wikipedia article titles as entity labels in Wikidata, as well as how the communities are intertwined. The fact that many titles are imported can be seen also in the comparison to Wikidata property labels of Table 1. Swedish is only ranked 20th, while Cebuano does not appear in the top 25 anymore at all. Since there are no Wikipedia articles linked to properties, those cannot be imported and have to be translated by the community of either project on Wikidata. Therefore we can assume this represents the language knowledge of editors and their bots on Wikidata quite accurately.

CONCLUSION

To gain a better understanding of the coverage of language communities on the Semantic Web, we analyzed data on natural language labels from Wikidata. This is an important topic that will influence the further direction of the Semantic Web and its use: adding one property label makes thousands of statements more useful, and can be further on used to give various language communities access to knowledge they would not have been able to access before.

There is still much room for improvement on the current state; as with most of the web, Wikidata's knowledge is mostly available in a few languages, while most languages have close to no coverage. Even languages spoken by large parts of the world population are not necessarily well covered. The languages that are covered the most are similar to Wikipedia, which we assume to be due to two factors: imports of Wikipedia article titles and an overlap of communities.

A few promising observations can be made, however. Languages do not necessarily have to be spoken by many people to achieve a higher level of completeness; suitable tools can greatly accelerate the process. At the same time, there are more comprehensive translations for data that is used more, as shown with properties in Wikidata. This should be the direction of future work: supporting the variety of languages on the Semantic Web, with tools that support the editor community in various languages and import more data in different languages to achieve greater coverage and with that more reuse.

ACKNOWLEDGMENTS

This project is supported by funding received from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 642795 (WDAqua ITN)

The authors gratefully acknowledge the input given by the Wikidata team at Wikimedia Deutschland and Matthew Flaschen (Wikimedia Foundation).

Rank	Wikipedia	Wikidata	WD properties
1	en	en	en
2	ceb	nl	nl
3	sv	fr	fr
4	de	de	ru
5	nl	es	mk
6	fr	it	de
7	ru	sv	es
8	it	ru	pl
9	es	ceb	ca
10	war	bg	it
11	pl	la	sr
12	vi	pt	hu
13	ja	pl	pt
14	pt	nb	nb
15	zh	vi	ko
16	uk	ja	fa
17	ca	da	da
18	fa	zh	cs
19	ar	war	ja
20	no	nn	sv
21	sh	fi	be
22	fi	ca	el
23	hu	hu	ar
24	id	cs	uk
25	cs	fa	zh

Table 1. Ranking of number of Wikipedia articles by language, all labels in Wikidata, and labels for properties in Wikidata

REFERENCES

1. Nitish Aggarwal, Tamara Polajnar, and Paul Buitelaar. 2013. *Cross-Lingual Natural Language Querying over the Web of Data*. Springer Berlin Heidelberg, Berlin, Heidelberg, 152–163. DOI: http://dx.doi.org/10.1007/978-3-642-38824-8_13
2. David Bamman, Brendan O'Connor, and Noah A. Smith. 2012. Censorship and deletion practices in Chinese social media. *First Monday* 17, 3 (2012), 429–440. DOI: http://dx.doi.org/10.5210/fm.v17i3.3943
3. Marcirio Chaves and Cássia Trojahn. 2010. Towards a Multilingual Ontology for Ontology-driven Content Mining in Social Web Sites. (Nov. 2010). http://repositorio-cientifico.uatlantica.pt/handle/10884/305
4. Maud Ehrmann, Francesco Cecconi, Daniele Vannella, John Philip McCrae, Philipp Cimiano, and Roberto Navigli. 2014. Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0.. In *LREC*. 401–408.
5. Basil Ell, Denny Vrandečić, and Elena Simperl. 2011. Labels in the Web of Data. *The Semantic Web–ISWC 2011* (2011), 162–176.
6. Brent 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 SIGCHI conference on human factors in computing systems*. ACM, 291–300.
7. Konrad Höffner, Sebastian Walter, Edgard Marx, Jens Lehmann, Axel-Cyrille Ngonga Ngomo, and Ricardo Usbeck. 2016. Overcoming challenges of semantic question answering in the semantic web. *Semantic Web Journal* (2016).
8. Lucie-Aimée Kaffee. *Generating Article Placeholders from Wikidata for Wikipedia: Increasing Access to Free and Open Knowledge*. Bachelor's Thesis.
9. Elena Montiel-Ponsoda, Jorge Gracia, Guadalupe Aguado-de Cea, and Asunción Gómez-Pérez. 2011. Representing translations on the semantic web. In *Proceedings of the 2nd International Conference on Multilingual Semantic Web-Volume 775*. CEUR-WS. org, 25–37.
10. Mozilla. 2017. Internet Health Report v.0.1 2017. (2017). https://internethealthreport.org/v01/
11. Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining Multilingual Topics from Wikipedia. In *Proceedings of the 18th International Conference on World Wide Web (WWW '09)*. ACM, New York, NY, USA, 1155–1156. DOI: http://dx.doi.org/10.1145/1526709.1526904
12. Mikael Parkvall. 2007. Världens 100 största språk 2007. *The World's* 100 (2007).
13. Molly Jackman Pat Wu. 2016. State of connectivity 2015: A report on global internet access. (Feb. 2016). http://newsroom.fb.com/news/2016/02/state-of-connectivity-2015-a-report-on-global-internet-access/
14. Maria Teresa Pazienza, Armando Stellato, Lina Henriksen, Patrizia Paggio, and Fabio Massimo Zanzotto. 2005. Ontology mapping to support multilingual ontology-based question answering. In *Proceedings of the Fourth International Semantic Web Conference (ISWC), Galway, Ireland*.
15. Alessandro Piscopo, Christopher Phethean, and Elena Simperl. 2017. Wikidatians are Born: Paths to Full Participation in a Collaborative Structured Knowledge Base. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.
16. Pavlos Vougiouklis, Jonathon Hare, and Elena Simperl. 2016. A neural network approach for knowledge-driven response generation. (2016).