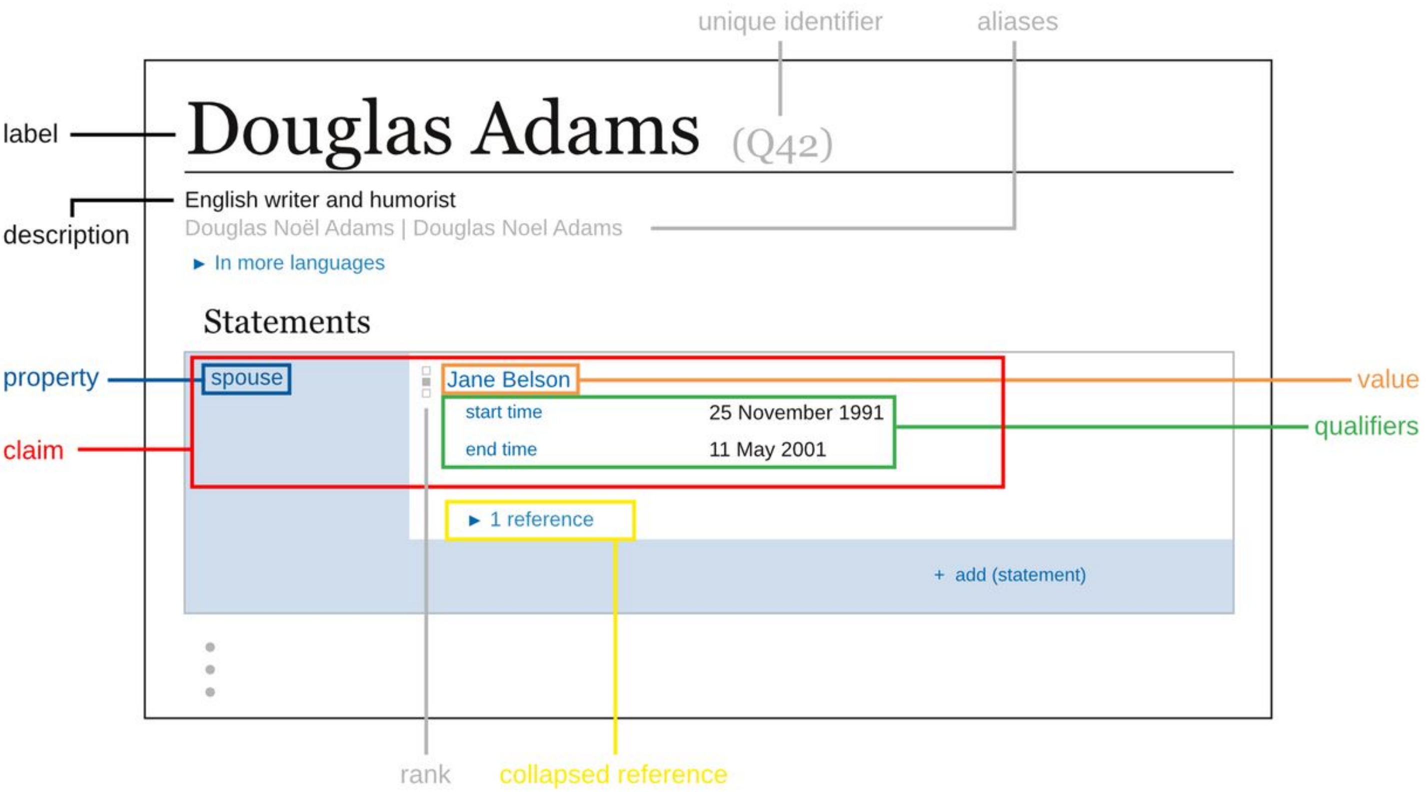
A Glimpse into Babel: An Analysis of Multilinguality in Wikidata

Lucie-Aimée Kaffee* (kaffee@soton.ac.uk)
Alessandro Piscopo*, Pavlos Vougiouklis*,
Elena Simperl*, Leslie Carr*, Lydia Pintscher**
*ECS, University of Southampton
**Wikimedia Deutschland



Douglas Adams (Q42)

English writer and humorist

Douglas Noël Adams | Douglas Noel Adams

▸ In more languages

Statements

spouse

Jane Belson

start time	25 November 1991
end time	11 May 2001

▸ 1 reference

+ add (statement)

An Item in Wikidata

CC BY-SA Incabell

https://commons.wikimedia.org/wiki/File:WikidataTermDescriptorDiagram.png

Wikidata

- Wikidata turtle dump of March 2017
- 26M entities
- 3,386 properties
- 134M labels
- rdfs:label

Q12345 rdfs:label “Count von Count”@en
Q12345 rdfs:label “Graf Zahl”@de
Q12345 rdfs:label “Граф фон Знак”@ru

Multilinguality in Wikidata - Why do we care?

- Labels are the access point for humans
- Give language communities access to existing knowledge
- Central storage for translations for (under resourced) languages
- Semantic Web in NLP
- Translations, Question Answering, Chat Bots, ...

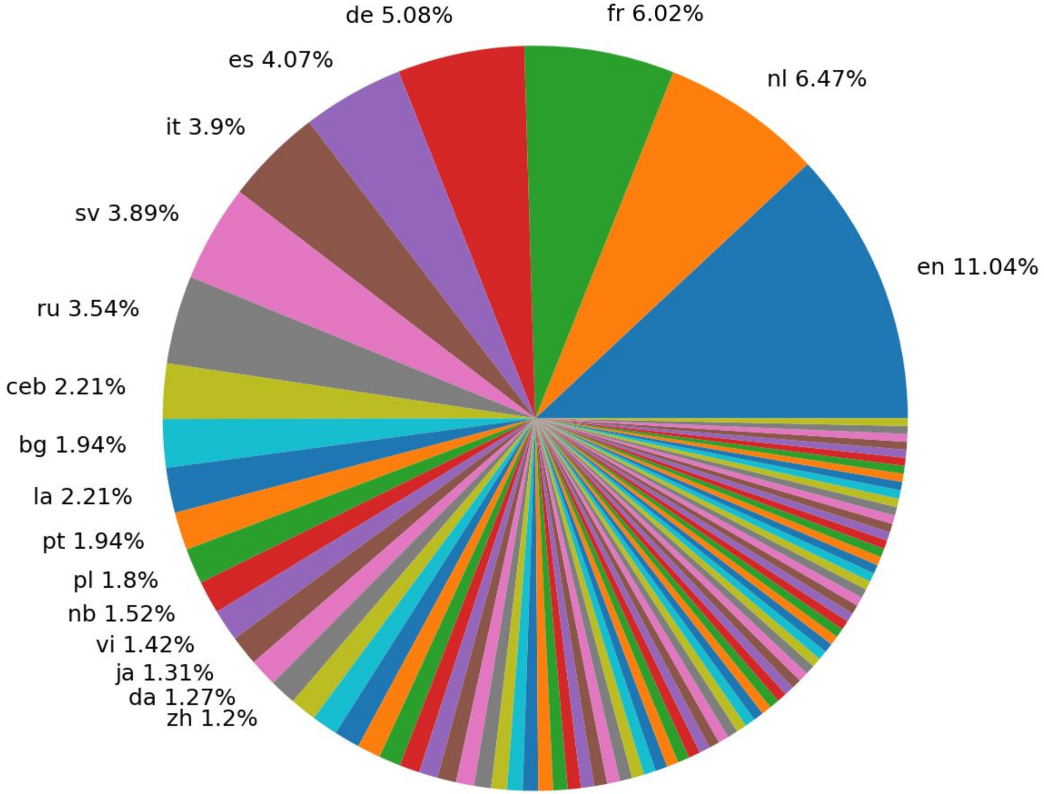
Research Questions

- **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?

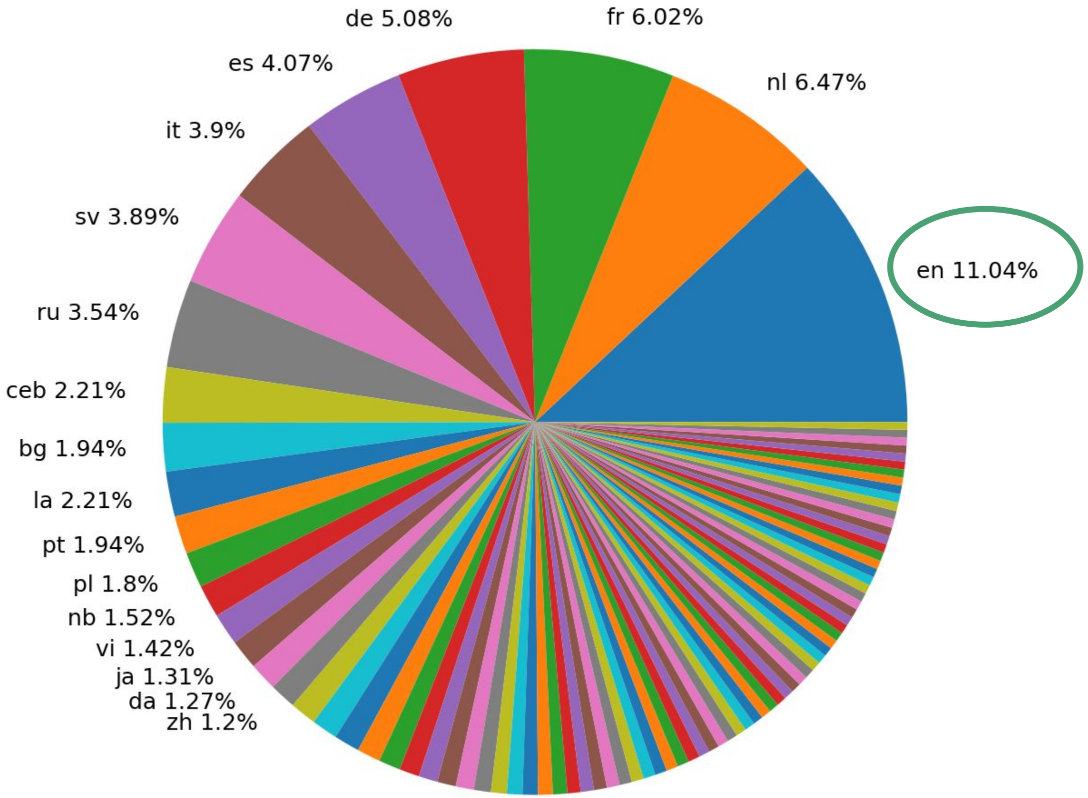
RQ1 What is the state of Wikidata with regard to multilinguality?

The Web

- > 50% English
- 25% of the world population English-speaking
- 2nd biggest language of users on the Web: Chinese
- Only 2% of web content Chinese

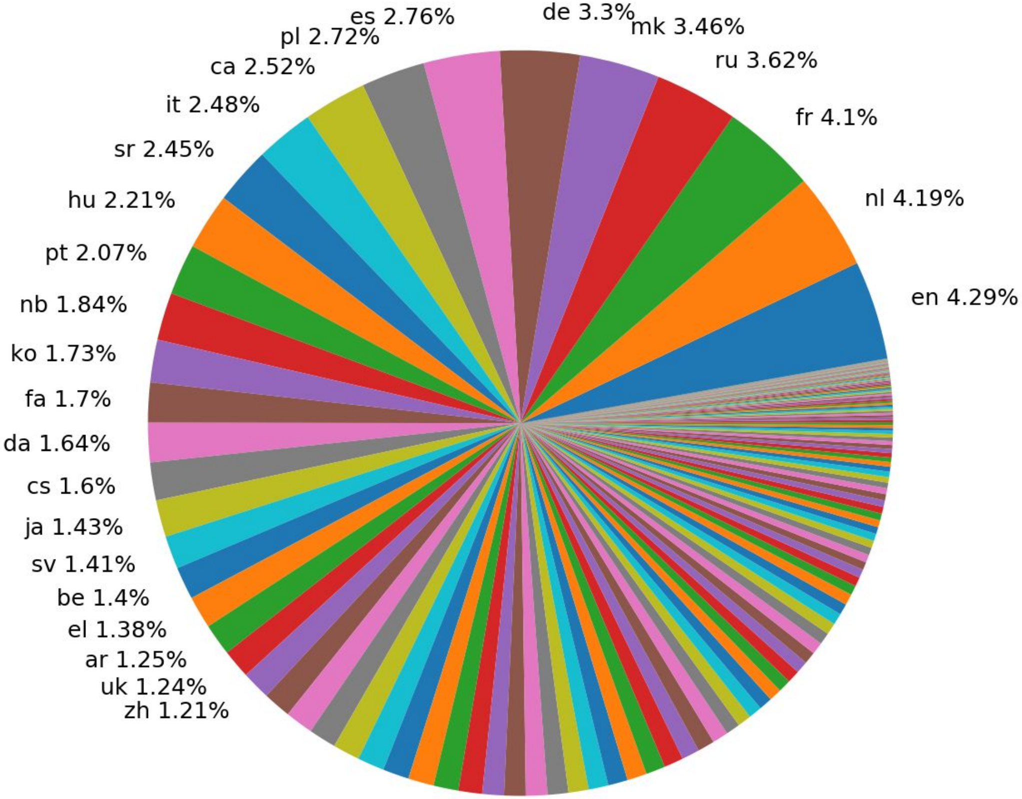


Percentage of all labels per language in Wikidata

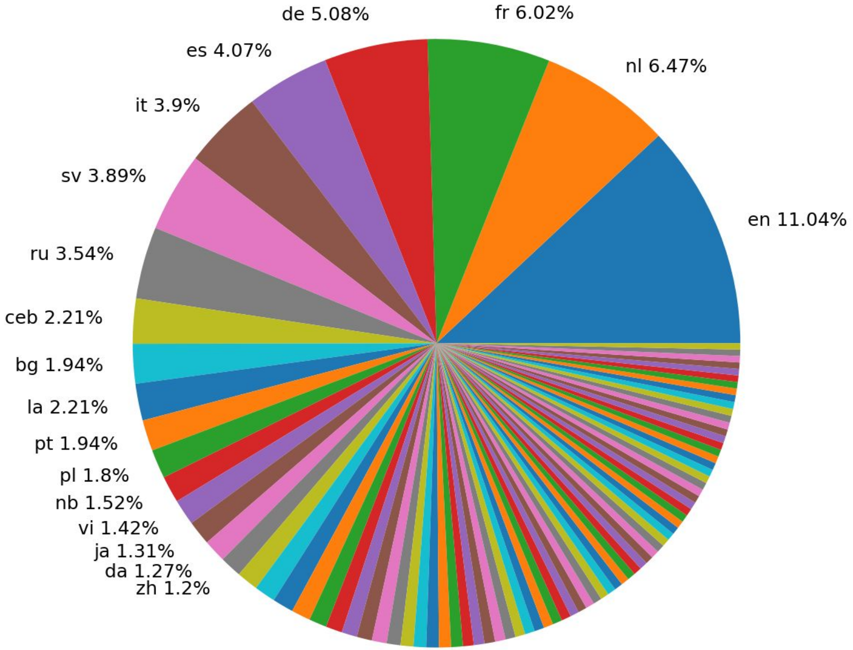


Percentage of all labels per language in Wikidata

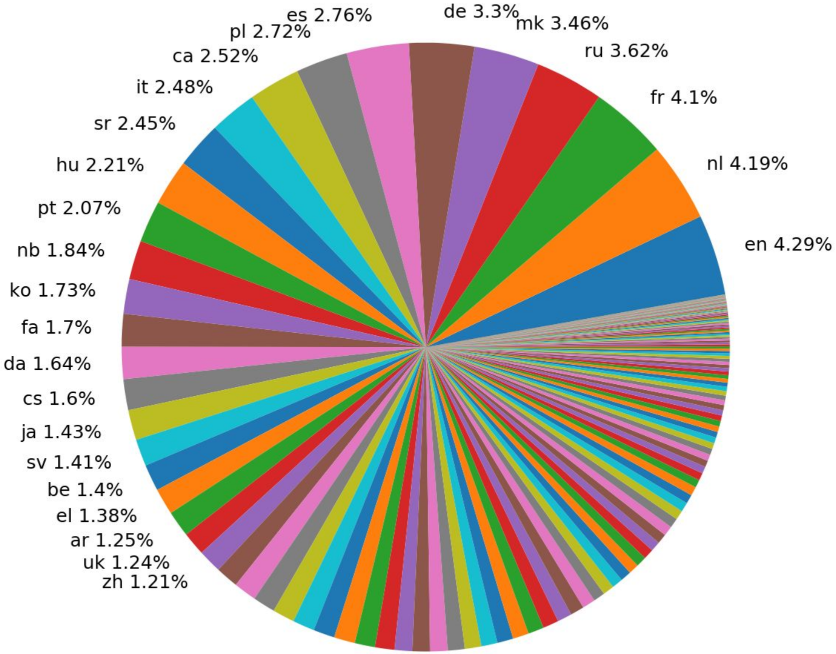
RQ2 Is there a difference in the multilinguality of the ontology, compared to the overall multilinguality of the knowledge base?



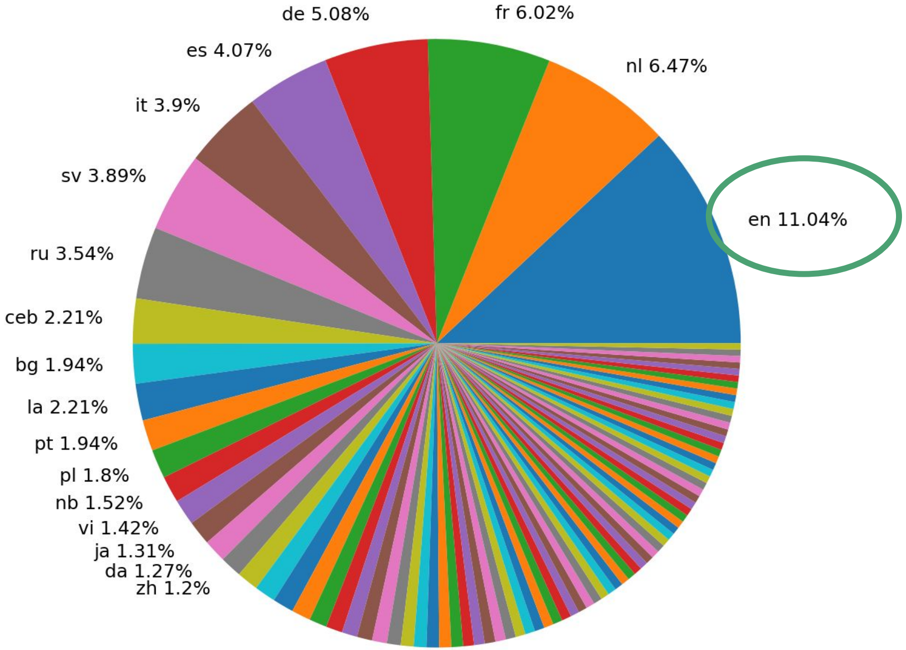
Distribution of languages for properties in Wikidata



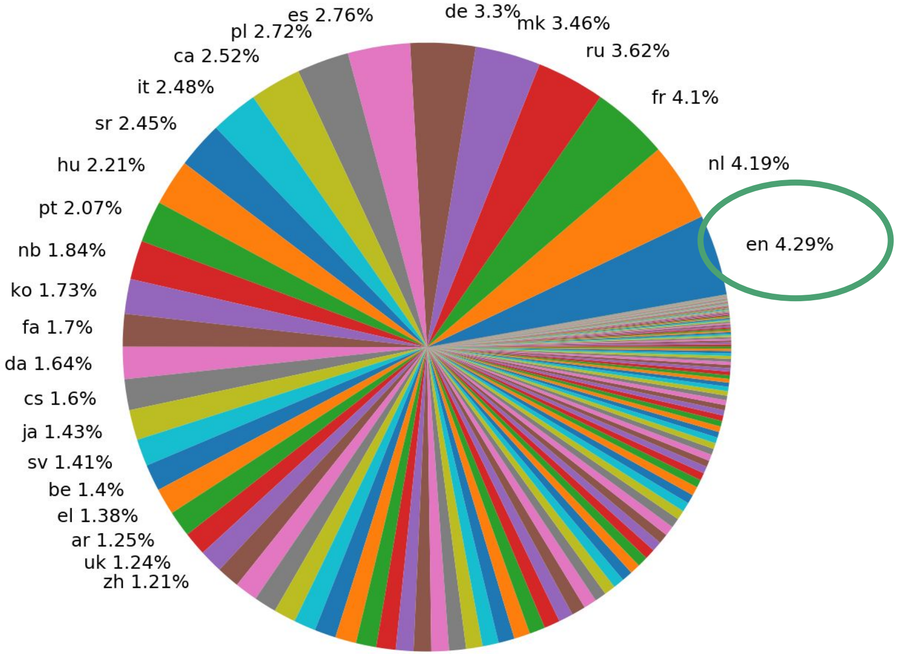
Percentage of all labels per language



Distribution of languages for properties



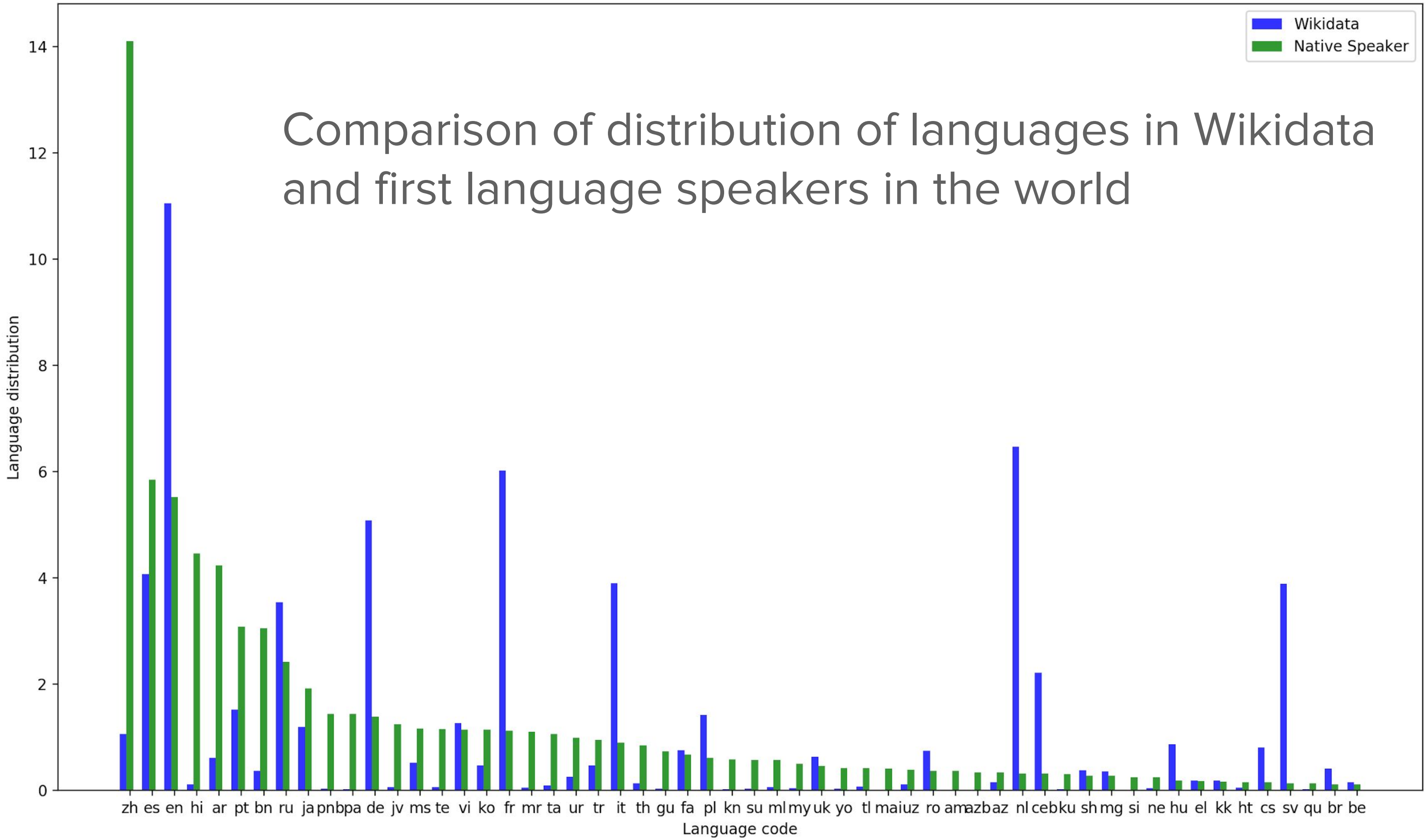
Percentage of all labels per language



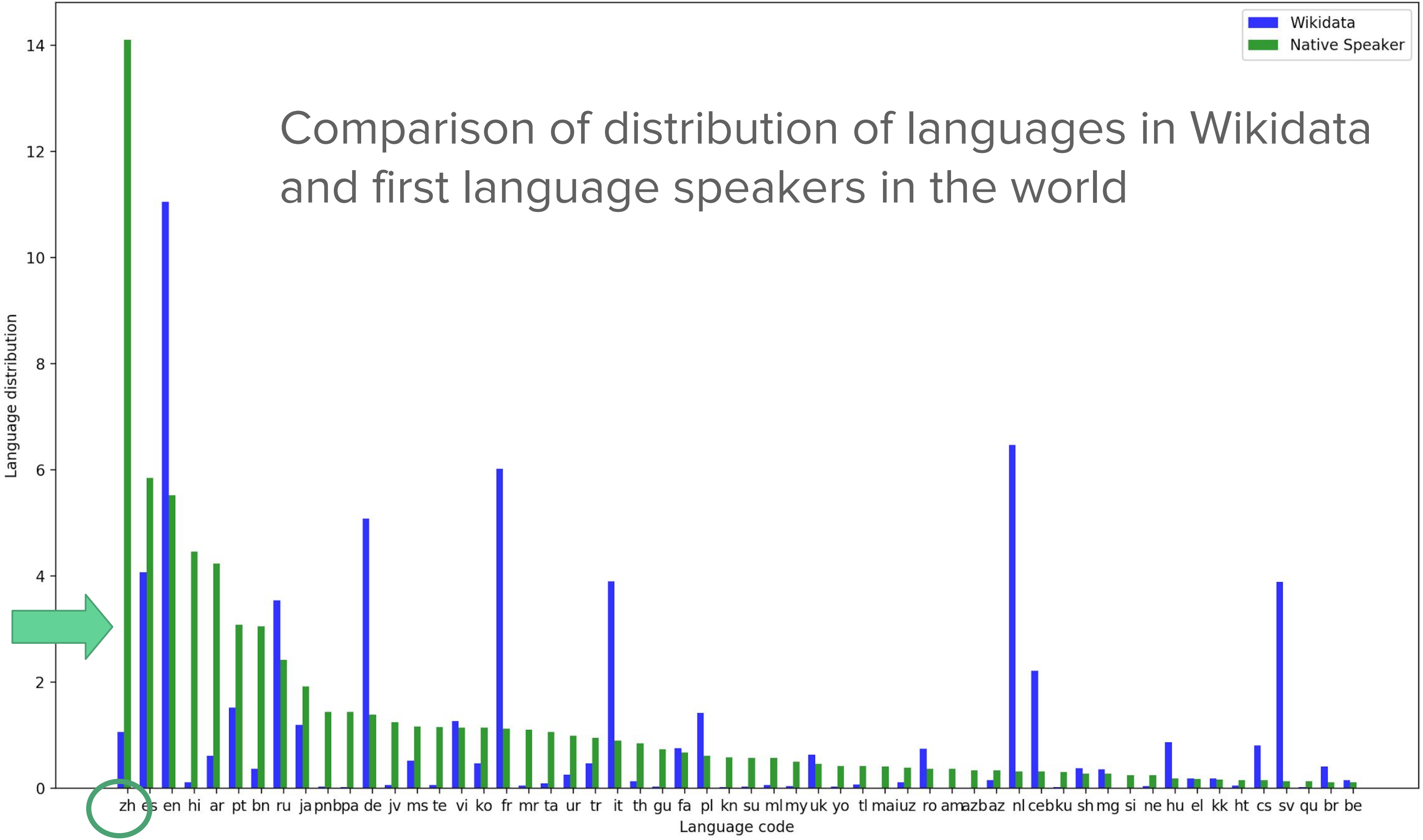
Distribution of languages for properties

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

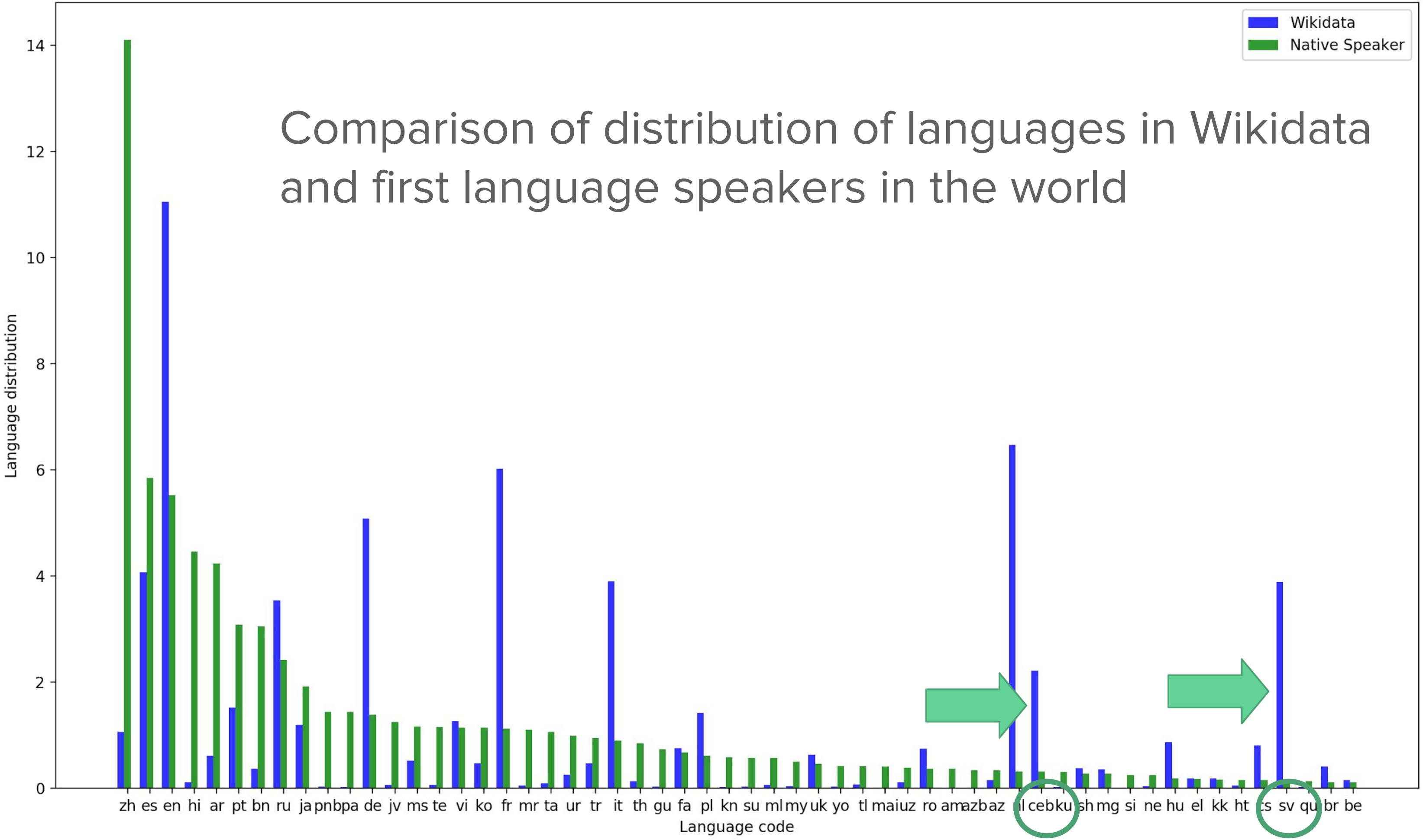
Comparison of distribution of languages in Wikidata and first language speakers in the world



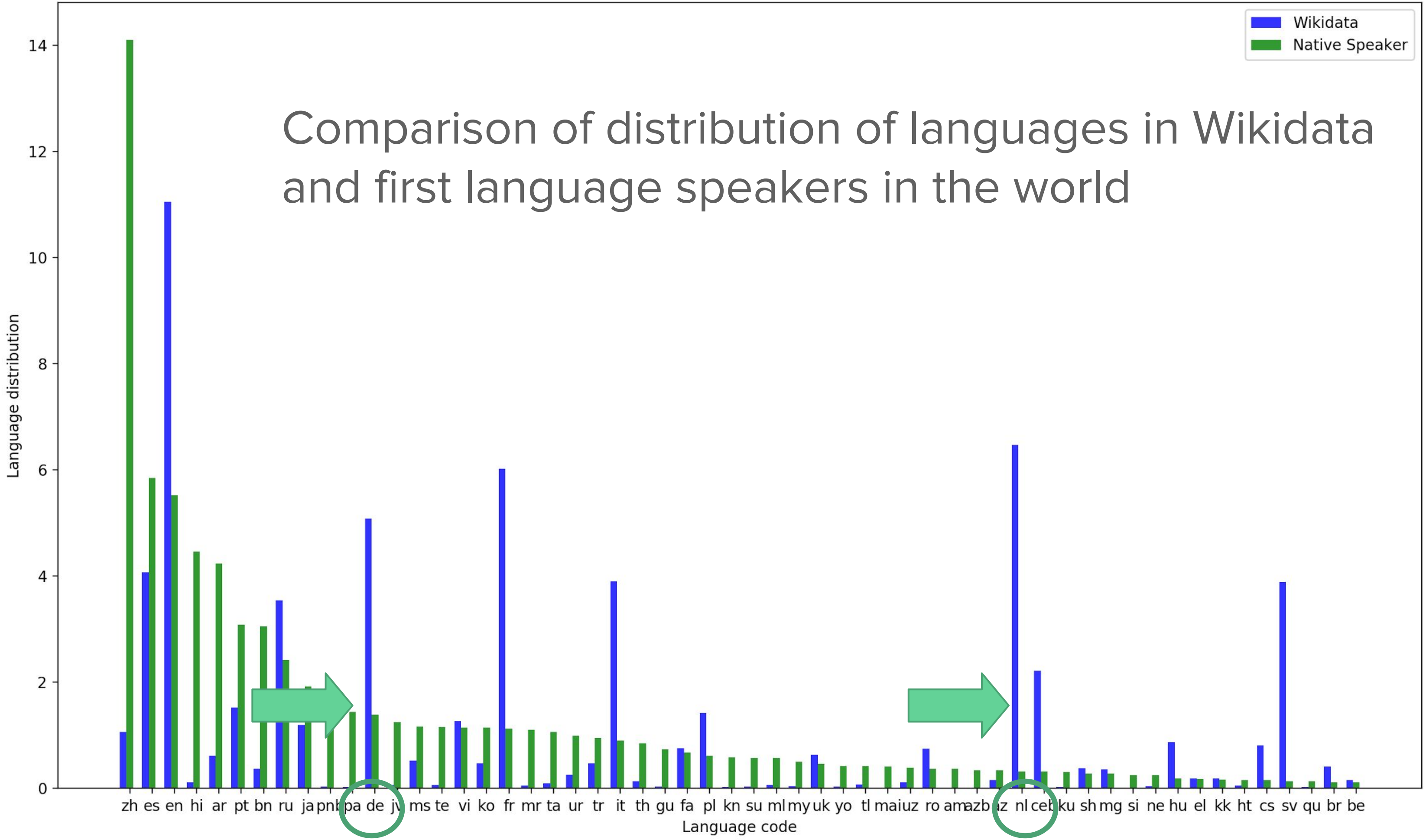
Comparison of distribution of languages in Wikidata and first language speakers in the world



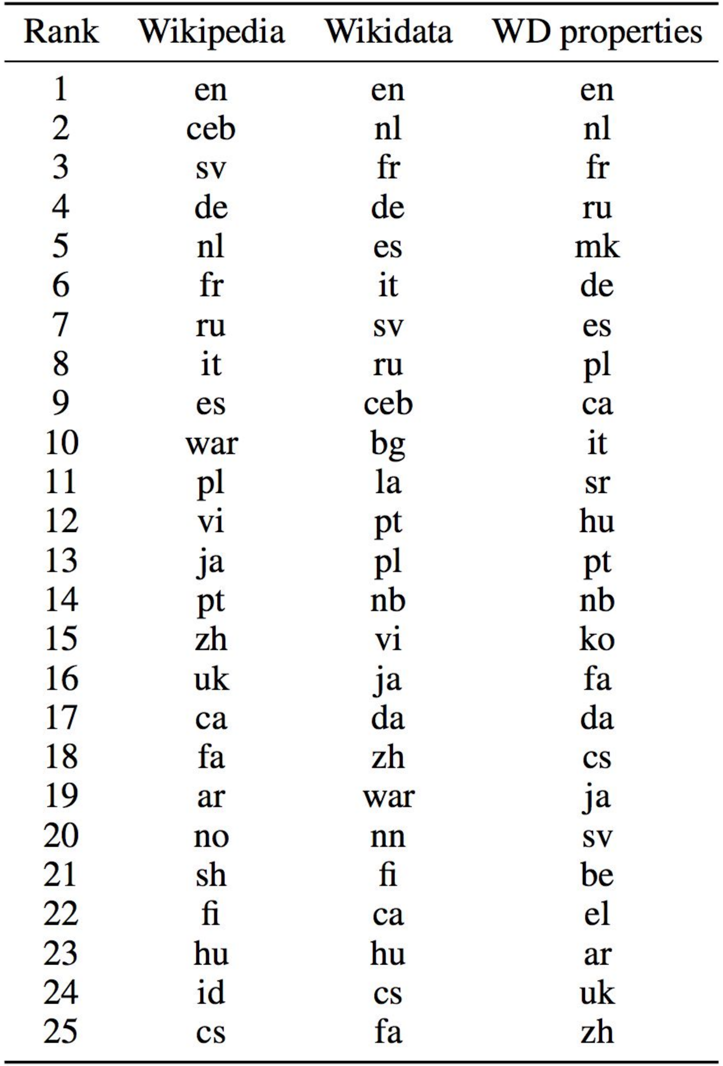
Comparison of distribution of languages in Wikidata and first language speakers in the world



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

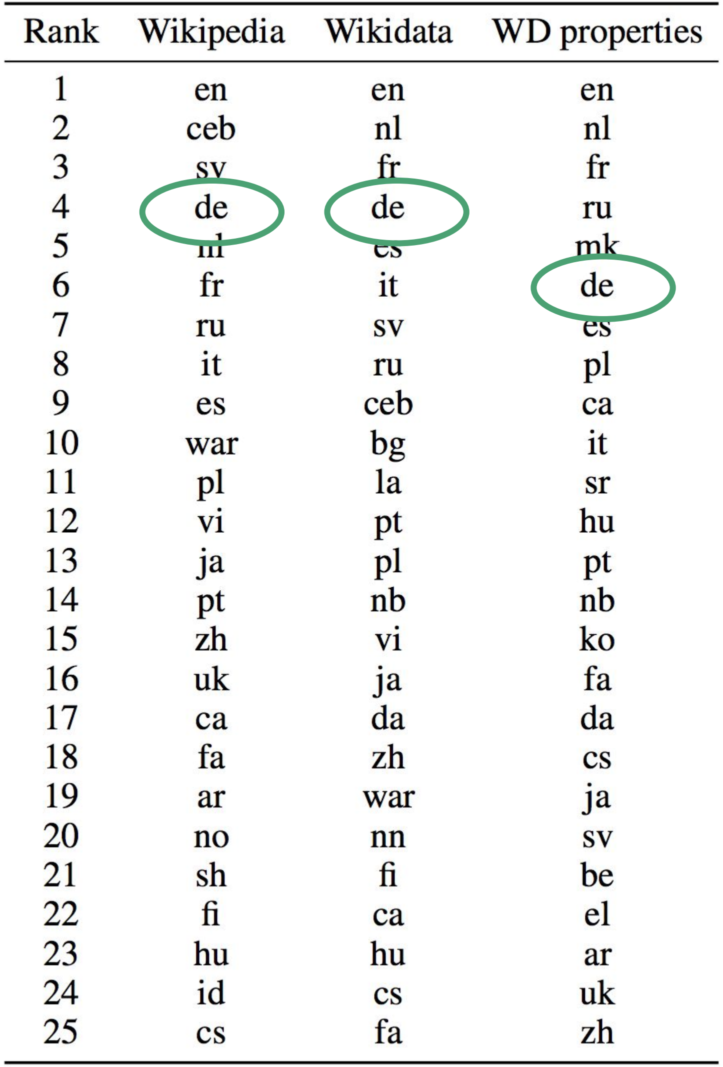


Ranking of number of Wikipedia articles by language, all labels in Wikidata, and labels for properties in Wikidata



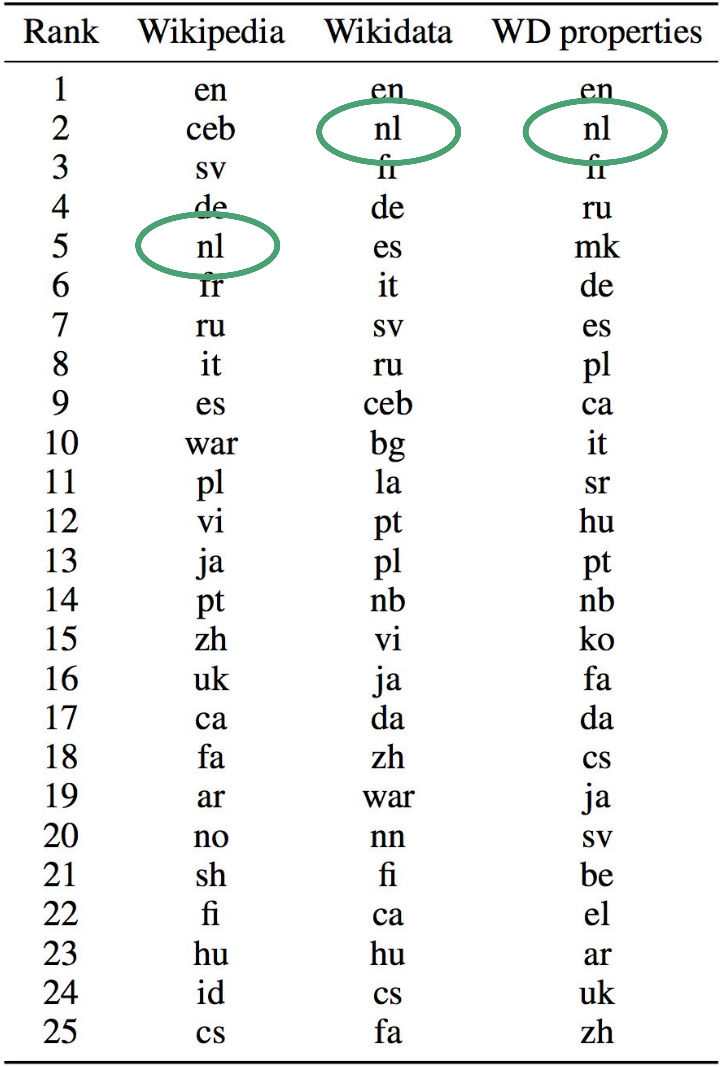
Rank	Wikipedia	Wikidata	WD properties
1	en	en	en
2	ceb	nl	nl
3	sv	fr	fr
4	de	de	ru
5	nl	es	mk
6	fr	it	de
7	ru	sv	es
8	it	ru	pl
9	es	ceb	ca
10	war	bg	it
11	pl	la	sr
12	vi	pt	hu
13	ja	pl	pt
14	pt	nb	nb
15	zh	vi	ko
16	uk	ja	fa
17	ca	da	da
18	fa	zh	cs
19	ar	war	ja
20	no	nn	sv
21	sh	fi	be
22	fi	ca	el
23	hu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

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



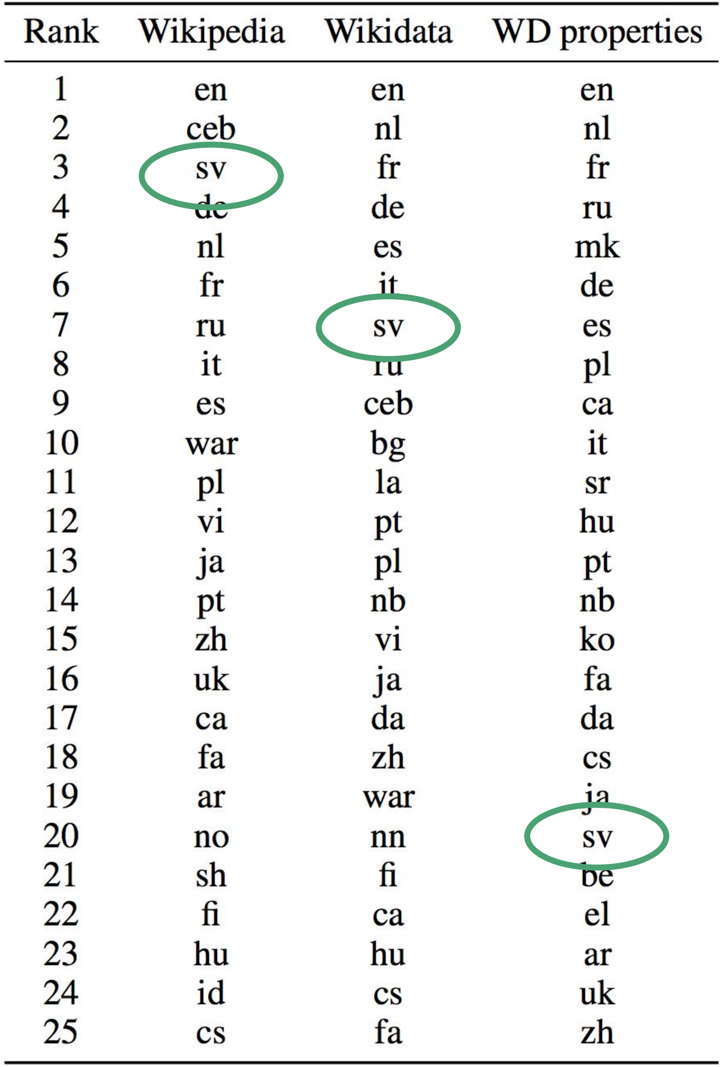
Rank	Wikipedia	Wikidata	WD properties
1	en	en	en
2	ceb	nl	nl
3	sv	fr	fr
4	de	de	ru
5	nl	es	mk
6	fr	it	de
7	ru	sv	es
8	it	ru	pl
9	es	ceb	ca
10	war	bg	it
11	pl	la	sr
12	vi	pt	hu
13	ja	pl	pt
14	pt	nb	nb
15	zh	vi	ko
16	uk	ja	fa
17	ca	da	da
18	fa	zh	cs
19	ar	war	ja
20	no	nn	sv
21	sh	fi	be
22	fi	ca	el
23	hu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

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



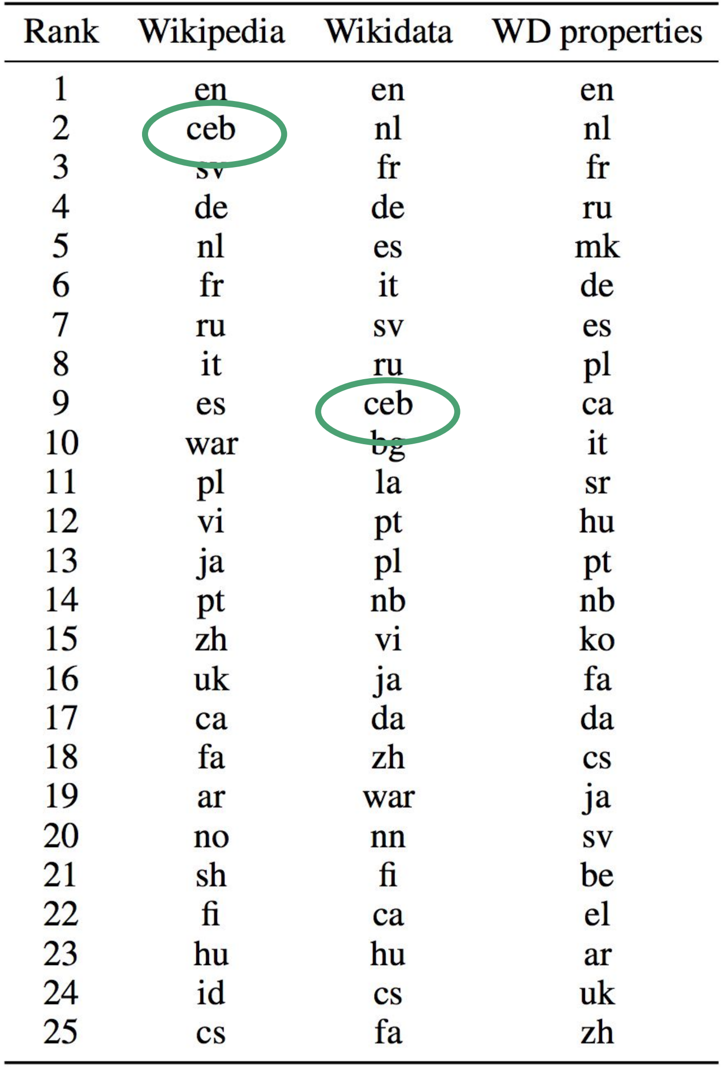
Rank	Wikipedia	Wikidata	WD properties
1	en	en	en
2	ceb	nl	nl
3	sv	fi	fi
4	de	de	ru
5	nl	es	mk
6	fr	it	de
7	ru	sv	es
8	it	ru	pl
9	es	ceb	ca
10	war	bg	it
11	pl	la	sr
12	vi	pt	hu
13	ja	pl	pt
14	pt	nb	nb
15	zh	vi	ko
16	uk	ja	fa
17	ca	da	da
18	fa	zh	cs
19	ar	war	ja
20	no	nn	sv
21	sh	fi	be
22	fi	ca	el
23	hu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

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



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

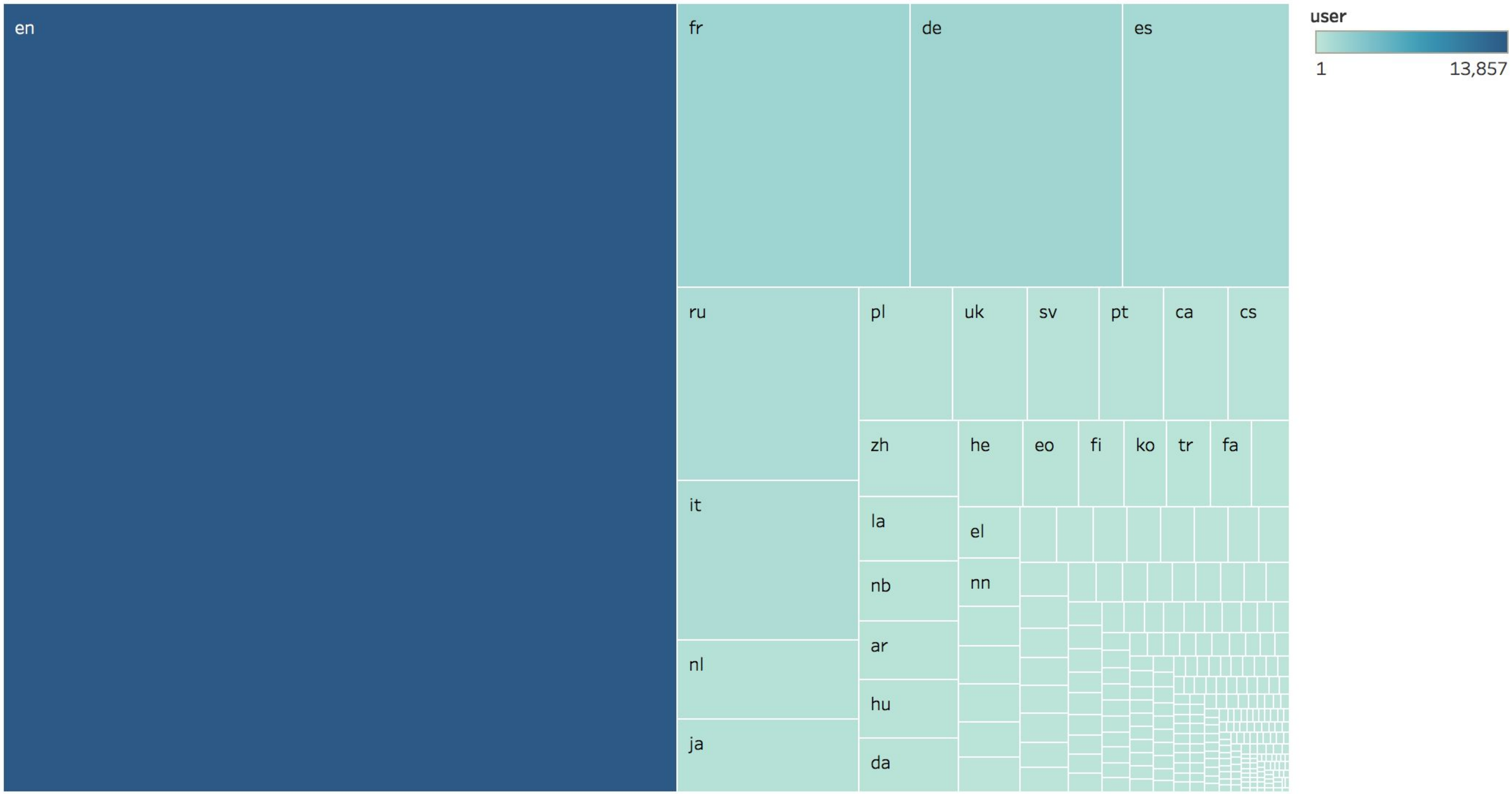
Ranking of number of Wikipedia articles by language, all labels in Wikidata, and labels for properties in Wikidata



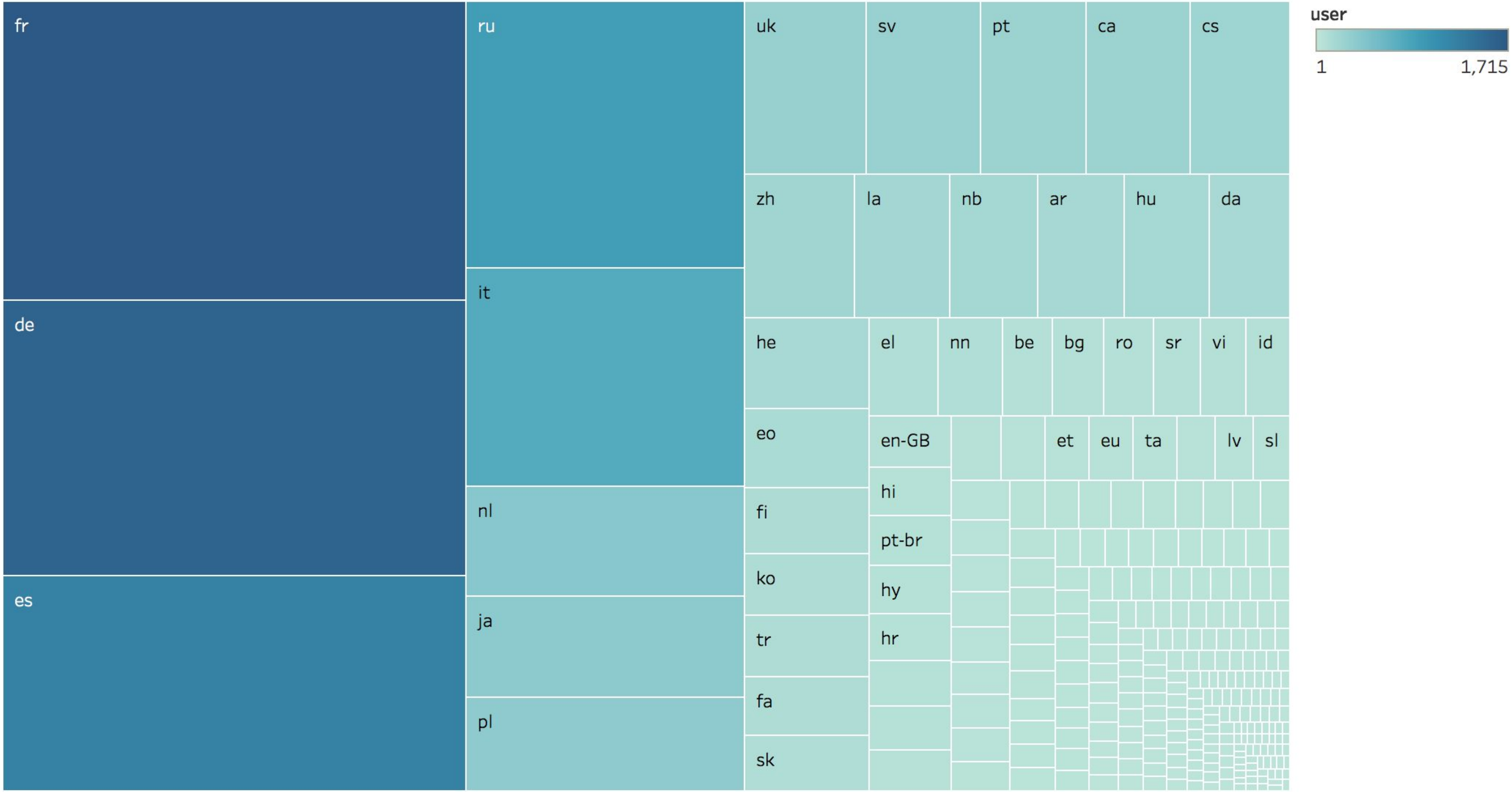
Rank	Wikipedia	Wikidata	WD properties
1	en	en	en
2	ceb	nl	nl
3	sv	fr	fr
4	de	de	ru
5	nl	es	mk
6	fr	it	de
7	ru	sv	es
8	it	ru	pl
9	es	ceb	ca
10	war	bg	it
11	pl	la	sr
12	vi	pt	hu
13	ja	pl	pt
14	pt	nb	nb
15	zh	vi	ko
16	uk	ja	fa
17	ca	da	da
18	fa	zh	cs
19	ar	war	ja
20	no	nn	sv
21	sh	fi	be
22	fi	ca	el
23	hu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

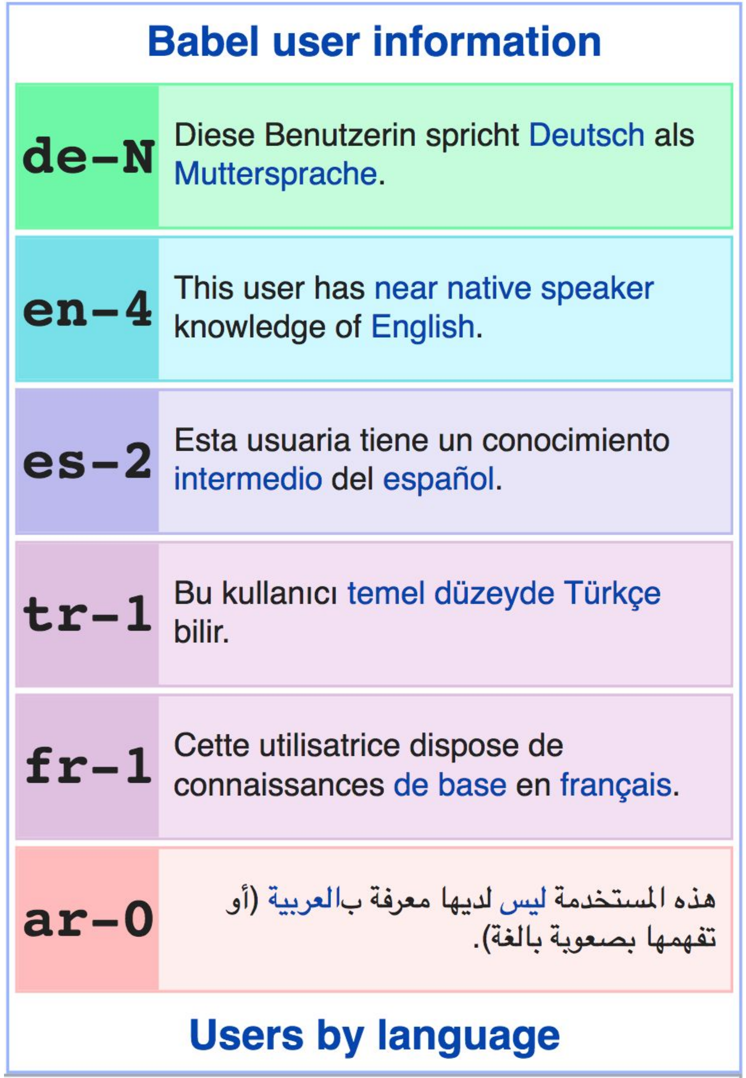
User in Wikidata

User Language Wikidata



User Language Wikidata (without English)



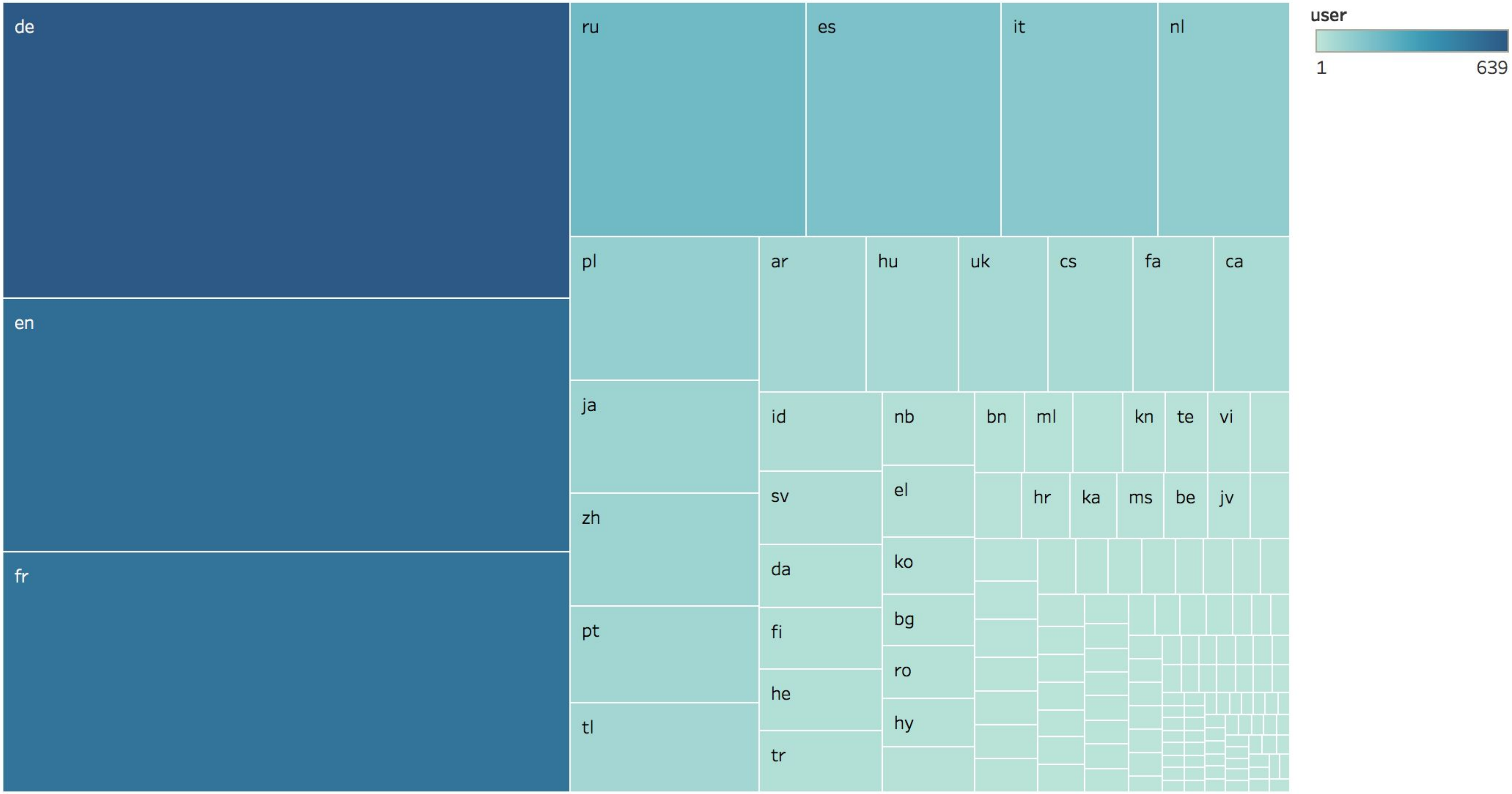


Babel user information

de-N	Diese Benutzerin spricht Deutsch als Muttersprache.
en-4	This user has near native speaker knowledge of English.
es-2	Esta usuaria tiene un conocimiento intermedio del español.
tr-1	Bu kullanıcı temel düzeyde Türkçe bilir.
fr-1	Cette utilisatrice dispose de connaissances de base en français.
ar-0	هذه المستخدمة ليس لديها معرفة بالعربية (أو تفهمها بصعوبة بالغة).

Users by language

BabelBox User Language



Conclusion

- Even languages spoken by large part of the world are not well-covered
- Community can have a big impact on language data
- More variety possible given the languages community knows
- Importing of labels/Bots can have a big impact of distribution of languages
- Still a long way to a truly multilingual (semantic) web

A Glimpse into Babel:
An Analysis of Multilinguality in Wikidata

Lucie-Aimée Kaffee
kaffee@soton.ac.uk

Paper at https://eprints.soton.ac.uk/413433/
Q37859976

UNIVERSITY OF
Southampton