Wikipedias in African languages

An overview based on Wikimedia statistics

Ingo Koll a.k.a. mtumiaji:kipala on sw.wikipedia.org

(Tip of the hat to Eric Zachte , ayatollah of wikimedia statistics !)

African Language Wikipedias (ALWs)

1. Some remarks on statistics and rankings
2. Which A.L.W.s do we have?
3. What is in it? A look at content
4. Who does it? A look at editors
5. Who reads it? A look at users
6. Closing observations and assumptions

"Never
believe
any statistic
as long as you
didn't forge it
yourself."

Alleged quote by
Winston
Churchill

probably
invented
and ascribed to
him by
Josef Goebbels

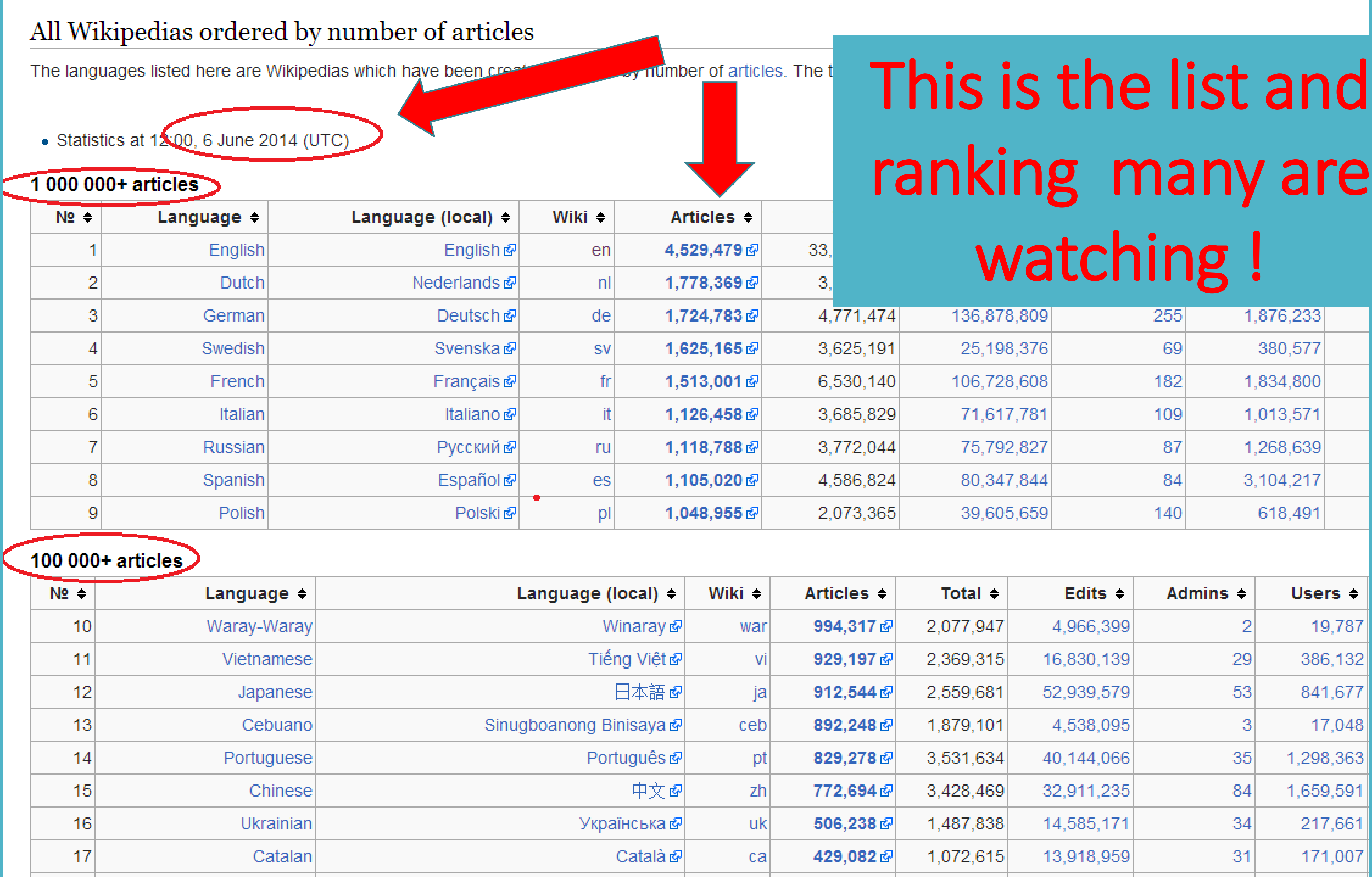
The motto for ANY statistic

This is not just a joke!

Wikimedia statistics are competetive for many editors

Debates are hot

Manipulations occur



All Wikipedias ordered by number of articles

The languages listed here are Wikipedias which have been cre by number of articles. The t

- Statistics at 12:00, 6 June 2014 (UTC)

1 000 000+ articles

№	Language	Language (local)	Wiki	Articles				
1	English	English	en	**4,529,479**	33,			
2	Dutch	Nederlands	nl	**1,778,369**	3,			
3	German	Deutsch	de	**1,724,783**	4,771,474	136,878,809	255	1,876,233
4	Swedish	Svenska	sv	**1,625,165**	3,625,191	25,198,376	69	380,577
5	French	Français	fr	**1,513,001**	6,530,140	106,728,608	182	1,834,800
6	Italian	Italiano	it	**1,126,458**	3,685,829	71,617,781	109	1,013,571
7	Russian	Русский	ru	**1,118,788**	3,772,044	75,792,827	87	1,268,639
8	Spanish	Español	es	**1,105,020**	4,586,824	80,347,844	84	3,104,217
9	Polish	Polski	pl	**1,048,955**	2,073,365	39,605,659	140	618,491

100 000+ articles

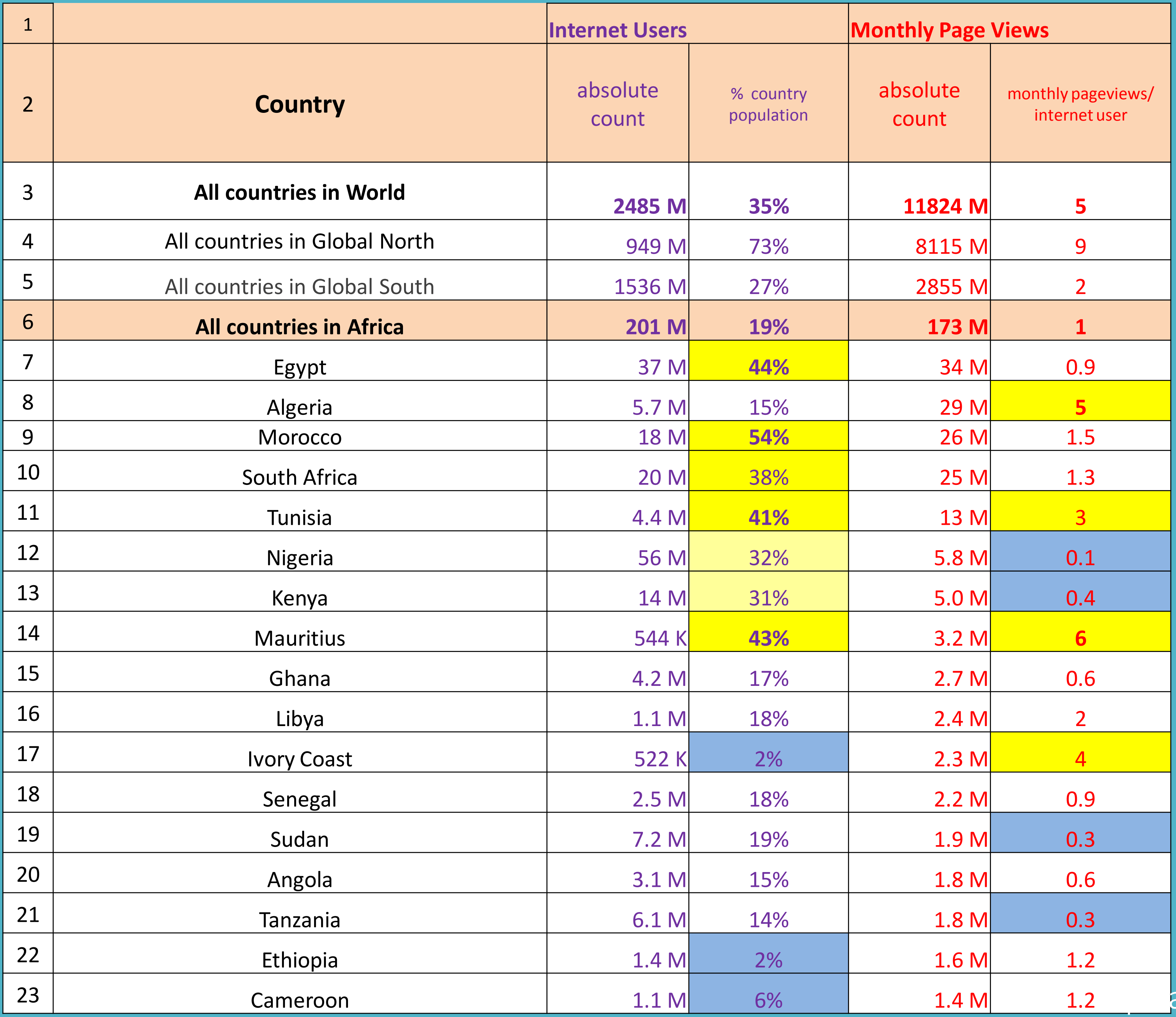
№	Language	Language (local)	Wiki	Articles	Total	Edits	Admins	Users
10	Waray-Waray	Winaray	war	**994,317**	2,077,947	4,966,399	2	19,787
11	Vietnamese	Tiếng Việt	vi	**929,197**	2,369,315	16,830,139	29	386,132
12	Japanese	日本語	ja	**912,544**	2,559,681	52,939,579	53	841,677
13	Cebuano	Sinugboanong Binisaya	ceb	**892,248**	1,879,101	4,538,095	3	17,048
14	Portuguese	Português	pt	**829,278**	3,531,634	40,144,066	35	1,298,363
15	Chinese	中文	zh	**772,694**	3,428,469	32,911,235	84	1,659,591
16	Ukrainian	Українська	uk	**506,238**	1,487,838	14,585,171	34	217,661
17	Catalan	Català	ca	**429,082**	1,072,615	13,918,959	31	171,007

This is the list and ranking many are watching !

Wikipedia in Africa

Some settings

- Low share of internet users among population
- few wikipedia looks among internet users



1		Internet Users		Monthly Page Views	
2	**Country**	absolute count	% country population	absolute count	monthly pageviews/ internet user
3	**All countries in World**	**2485 M**	**35%**	**11824 M**	**5**
4	All countries in Global North	949 M	73%	8115 M	9
5	All countries in Global South	1536 M	27%	2855 M	2
6	**All countries in Africa**	**201 M**	**19%**	**173 M**	**1**
7	Egypt	37 M	**44%**	34 M	0.9
8	Algeria	5.7 M	15%	29 M	**5**
9	Morocco	18 M	**54%**	26 M	1.5
10	South Africa	20 M	38%	25 M	1.3
11	Tunisia	4.4 M	**41%**	13 M	3
12	Nigeria	56 M	32%	5.8 M	0.1
13	Kenya	14 M	31%	5.0 M	0.4
14	Mauritius	544 K	**43%**	3.2 M	**6**
15	Ghana	4.2 M	17%	2.7 M	0.6
16	Libya	1.1 M	18%	2.4 M	2
17	Ivory Coast	522 K	2%	2.3 M	4
18	Senegal	2.5 M	18%	2.2 M	0.9
19	Sudan	7.2 M	19%	1.9 M	0.3
20	Angola	3.1 M	15%	1.8 M	0.6
21	Tanzania	6.1 M	14%	1.8 M	0.3
22	Ethiopia	1.4 M	2%	1.6 M	1.2
23	Cameroon	1.1 M	6%	1.4 M	1.2

Internet access and wikipedia use in African countries (top of the list)

Yellow marked: multiple of averages (higher)

Blue marked: fraction only of averages (lower)

(further down the list there is more blue)

a: Wikipedias in African languages, 20.06.2014

Wikipedia ranking and bots

Share of Bot-created articles

	ALL Wikipedias	**English Wp**	**Dutch Wp**	**German Wp**	**Waray Waray**
	Σ	en	nl	de	war
Σ total articles created	31.4 M	4.6 M	1.8 M	1.7 M	960 k
Σ manually created articles	24.0 M	4.4 M	728 k	1.7 M	115 k
Σ articles created by bots	**7.4 M**	**155 k**	**1.0 M**	**739**	**845 k**
Share of articles created by bots	**23.6%**	**3%**	**59%**	**0%**	**88%**

Wikipedia: 287 languages

ca. 40 African language wikipedias

A look at wikipedias by size

Huge Wikipedias (+ 1 mio articles)	> 9
Large Wikipedias (100k to 1 Mio articles)	> 43
Medium size Wikipedias (10k to less than 100k)	> 74
Small Wikipedias (1 k to less than 10 k articles)	>101
Tiny Wikipedias (100 to less than 1k articles)	> 50
“Micropedias” (less than 100 articles)	> 10

Wikipedias in African languages:

„African languages“ – languages native to Africa
= including Afrikaans + Masri
= excluding English, French, Arabic

~~Huge Wikipedias~~	~~(1mio+)~~	
~~Large Wikipedias~~	~~(100k +)~~	
Medium size Wikipedias	(10k +)	6
Small Wikipedias	(1k+)	7
Tiny Wikipedias	(100+)	23 +
“Micropedias”		4 (?)

The African language wikipedias

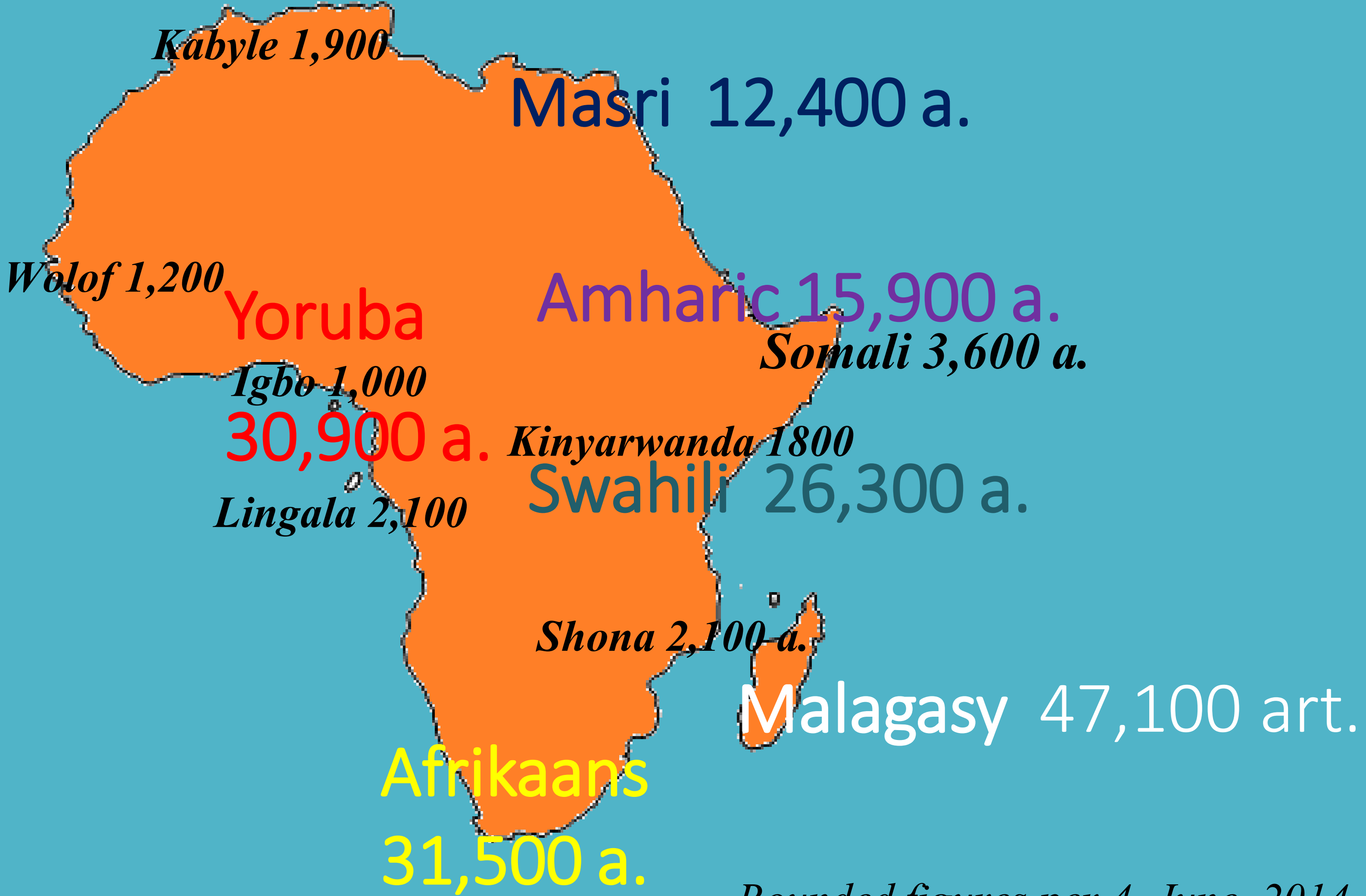
Medium and Small size African language wikipedias

TINY

Kongo	876
N. Sotho	690
Zulu	628
Hausa	586
Setswana	509
Twi	460
Bambara	399
Siswati	398
Kikuyu	358
Xhosa	329
Kirundi	313
Oromo	303
Tsonga	302
Ewe	297
Akan	284
Tigrinya	269
Sangro	215
Ganda	212
Venda	208
Fulfulde	208
Sesotho	193
Tumbuka	173
Chichewa	168

MICRO

Ndonga	21
Afar	6
Kanuri	1
Herero	0



Rounded figures per 4. June, 2014

Wikipedias in African languages: What`s in it?

Official article count		Database
Malagasy	47 k	88 MB*
Afrikaans	31 k	107 MB
Yoruba	31 k	20 MB
Swahili	26 k	48 MB
Amharic	13 k	13 MB
Masri	12 k	30 MB
Somali	4.0 k	7.2 MB
		Text, discussion, help

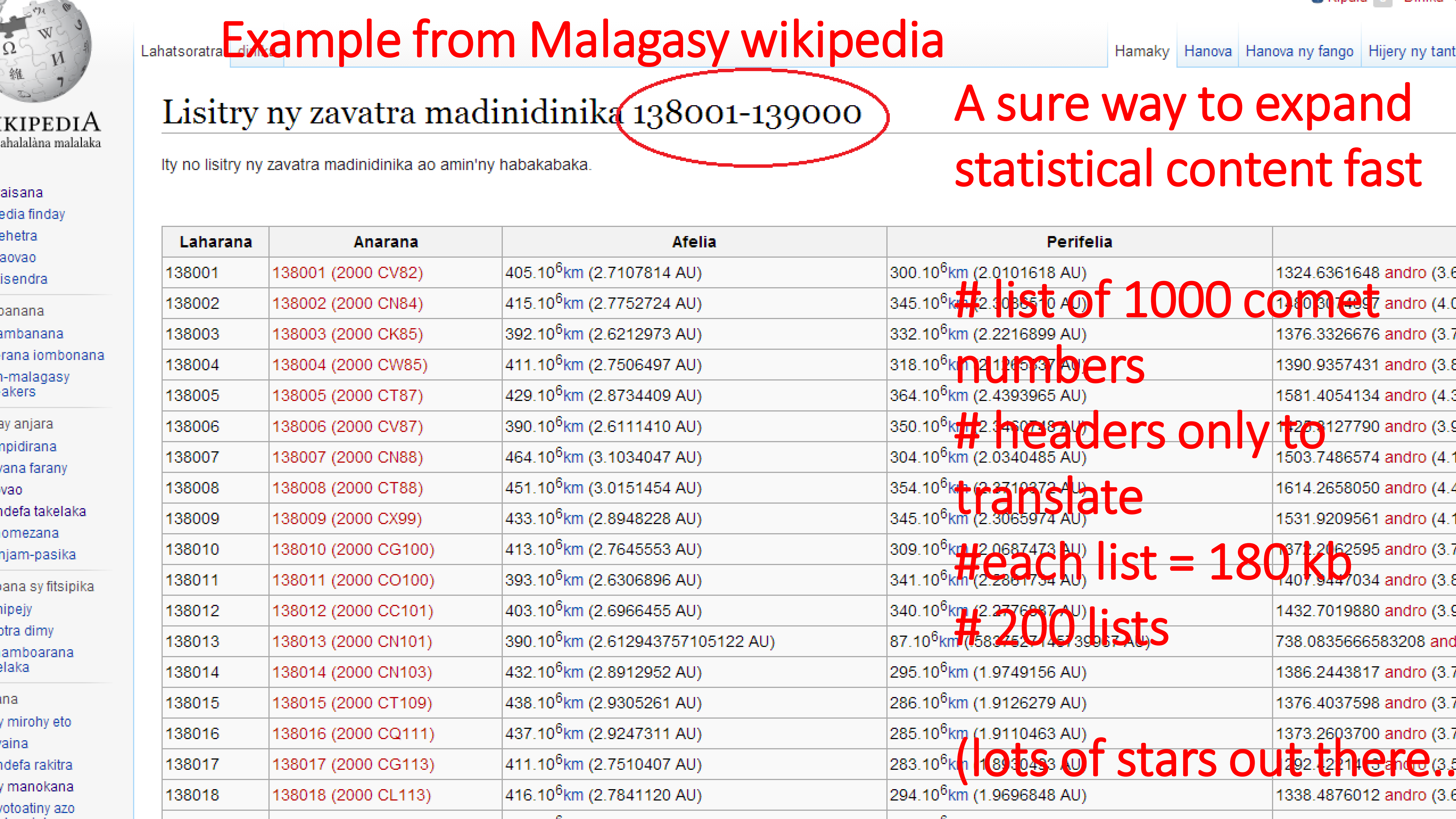
Alternative article count

(exclude less than 200 ch)

Malagasy	44 k
Afrikaans	30 k
Yoruba	5.4 k
Swahili	22 k
Amharic	4.4 k
Masri	9.7 k
Somali	2.4 k

* Malagasy database 37 MB lists only! (34 MB – thousands of comet names + few data!)

Example from Malagasy wikipedia



aisana
edia finday
ehetra
aovao
isendra

oanana
ambanana
rana iombonana
n-malagasy
akers

ay anjara
npidirana
vana farany
vao
ndefa takelaka
omezana
njam-pasika

oana sy fitsipika
nipejy
otra dimy
amboarana
elaka

na
y mirohy eto
vaina
ndefa rakitra
y manokana
votoatiny azo

Lisitry ny zavatra madinidinika 138001-139000

Ity no lisitry ny zavatra madinidinika ao amin'ny habakabaka.

Laharana	Anarana	Afelia	Perifelia	
138001	138001 (2000 CV82)	405.10^6km (2.7107814 AU)	300.10^6km (2.0101618 AU)	1324.6361648 andro (3.6
138002	138002 (2000 CN84)	415.10^6km (2.7752724 AU)	345.10^6km (2.30[illegible] AU)	[illegible] andro (4.0
138003	138003 (2000 CK85)	392.10^6km (2.6212973 AU)	332.10^6km (2.2216899 AU)	1376.3326676 andro (3.7
138004	138004 (2000 CW85)	411.10^6km (2.7506497 AU)	318.10^6km (2.12[illegible] AU)	1390.9357431 andro (3.8
138005	138005 (2000 CT87)	429.10^6km (2.8734409 AU)	364.10^6km (2.4393965 AU)	1581.4054134 andro (4.3
138006	138006 (2000 CV87)	390.10^6km (2.6111410 AU)	350.10^6km (2.3[illegible] AU)	[illegible]127790 andro (3.9
138007	138007 (2000 CN88)	464.10^6km (3.1034047 AU)	304.10^6km (2.0340485 AU)	1503.7486574 andro (4.1
138008	138008 (2000 CT88)	451.10^6km (3.0151454 AU)	354.10^6km (2.3710372 AU)	1614.2658050 andro (4.4
138009	138009 (2000 CX99)	433.10^6km (2.8948228 AU)	345.10^6km (2.3065974 AU)	1531.9209561 andro (4.1
138010	138010 (2000 CG100)	413.10^6km (2.7645553 AU)	309.10^6km (2.0687473 AU)	1372.2062595 andro (3.7
138011	138011 (2000 CO100)	393.10^6km (2.6306896 AU)	341.10^6km (2.2861734 AU)	1407.9447034 andro (3.8
138012	138012 (2000 CC101)	403.10^6km (2.6966455 AU)	340.10^6km (2.2776887 AU)	1432.7019880 andro (3.9
138013	138013 (2000 CN101)	390.10^6km (2.612943757105122 AU)	87.10^6km ([illegible] AU)	738.083566583208 and
138014	138014 (2000 CN103)	432.10^6km (2.8912952 AU)	295.10^6km (1.9749156 AU)	1386.2443817 andro (3.7
138015	138015 (2000 CT109)	438.10^6km (2.9305261 AU)	286.10^6km (1.9126279 AU)	1376.4037598 andro (3.7
138016	138016 (2000 CQ111)	437.10^6km (2.9247311 AU)	285.10^6km (1.9110463 AU)	1373.2603700 andro (3.7
138017	138017 (2000 CG113)	411.10^6km (2.7510407 AU)	283.10^6km (1.8930453 AU)	[illegible] andro (3.5
138018	138018 (2000 CL113)	416.10^6km (2.7841120 AU)	294.10^6km (1.9696848 AU)	1338.4876012 andro (3.6

A sure way to expand statistical content fast

# list of 1000 comet numbers

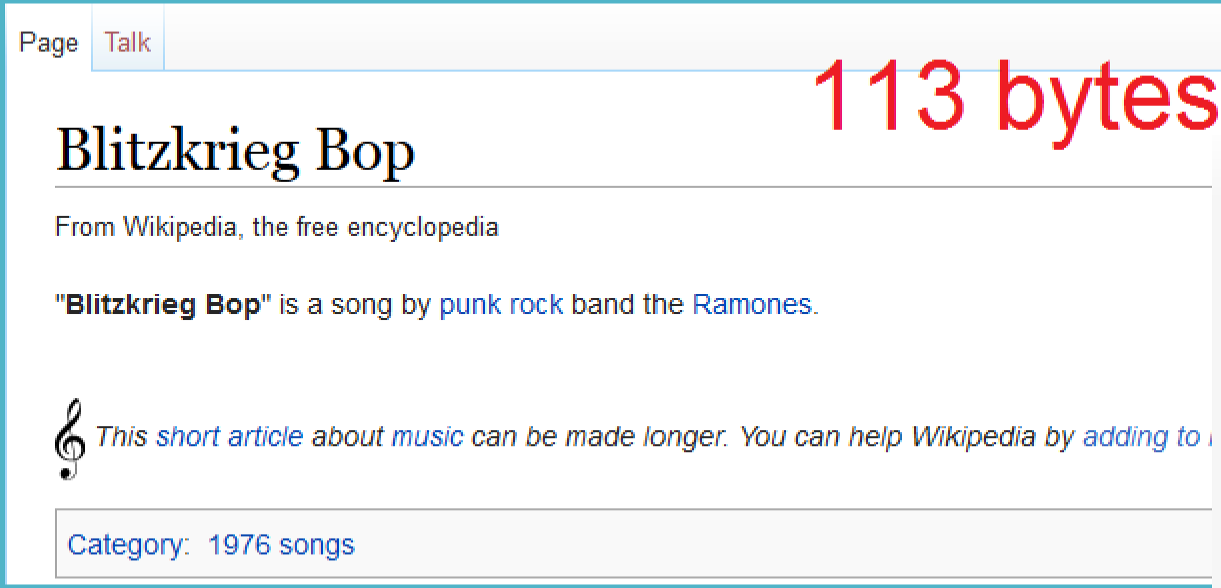
# headers only to translate

#each list = 180 kb

# 200 lists

(lots of stars out there...

2 different types of very short articles



113 bytes

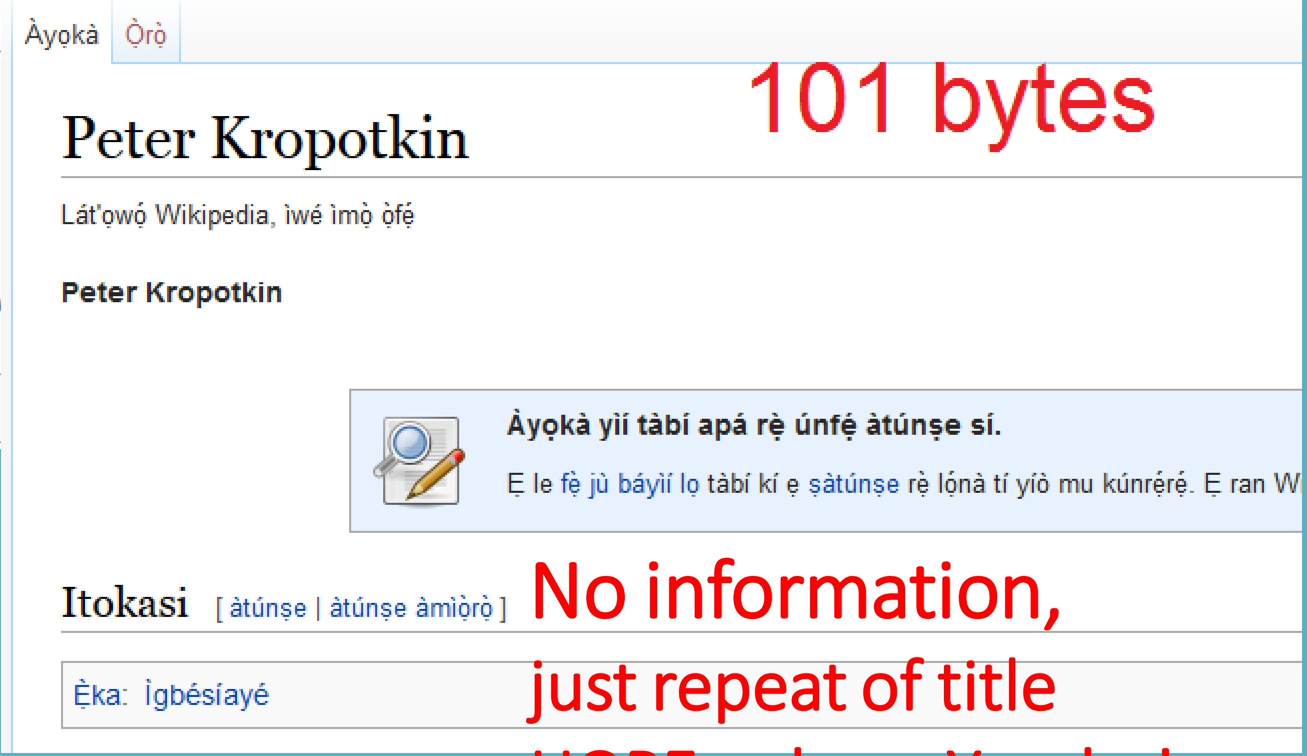
Blitzkrieg Bop

From Wikipedia, the free encyclopedia

"**Blitzkrieg Bop**" is a song by punk rock band the Ramones.

This short article about music can be made longer. You can help Wikipedia by adding to i

Category: 1976 songs



101 bytes

Peter Kropotkin

Lát'ọwọ́ Wikipedia, ìwé ìmọ̀ ọ̀fẹ́

Peter Kropotkin

Àyọkà yìí tàbí apá rẹ̀ únfẹ́ àtúnṣe sí.

Ẹ le fẹ̀ jù báyìí lọ tàbí kí ẹ ṣàtúnṣe rẹ̀ lọ́nà tí yíò mu kúnrẹ́rẹ́. Ẹ ran W

Itokasi [àtúnṣe | àtúnṣe àmìọ̀rọ̀]

Ẹ̀ka: Ìgbésíayé

<u>Basic information, 2 links</u>

>(punk rock band) >(Ramones)

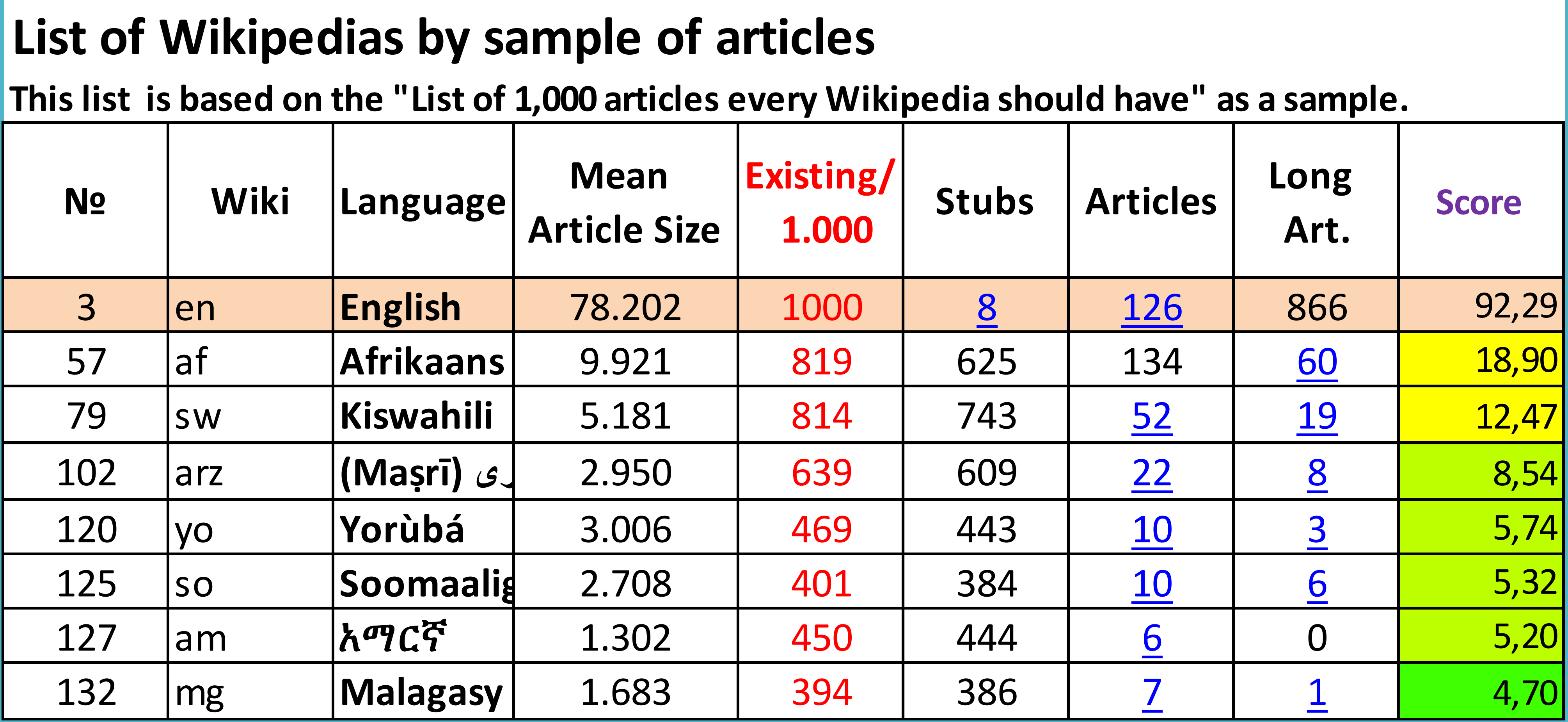
<u>+ HOPE </u>that someone adds content on simple.wikipedia

No information,

just repeat of title

HOPE only on Yoruba!

What is in it? Encyclopedial content by list of „most important articles“



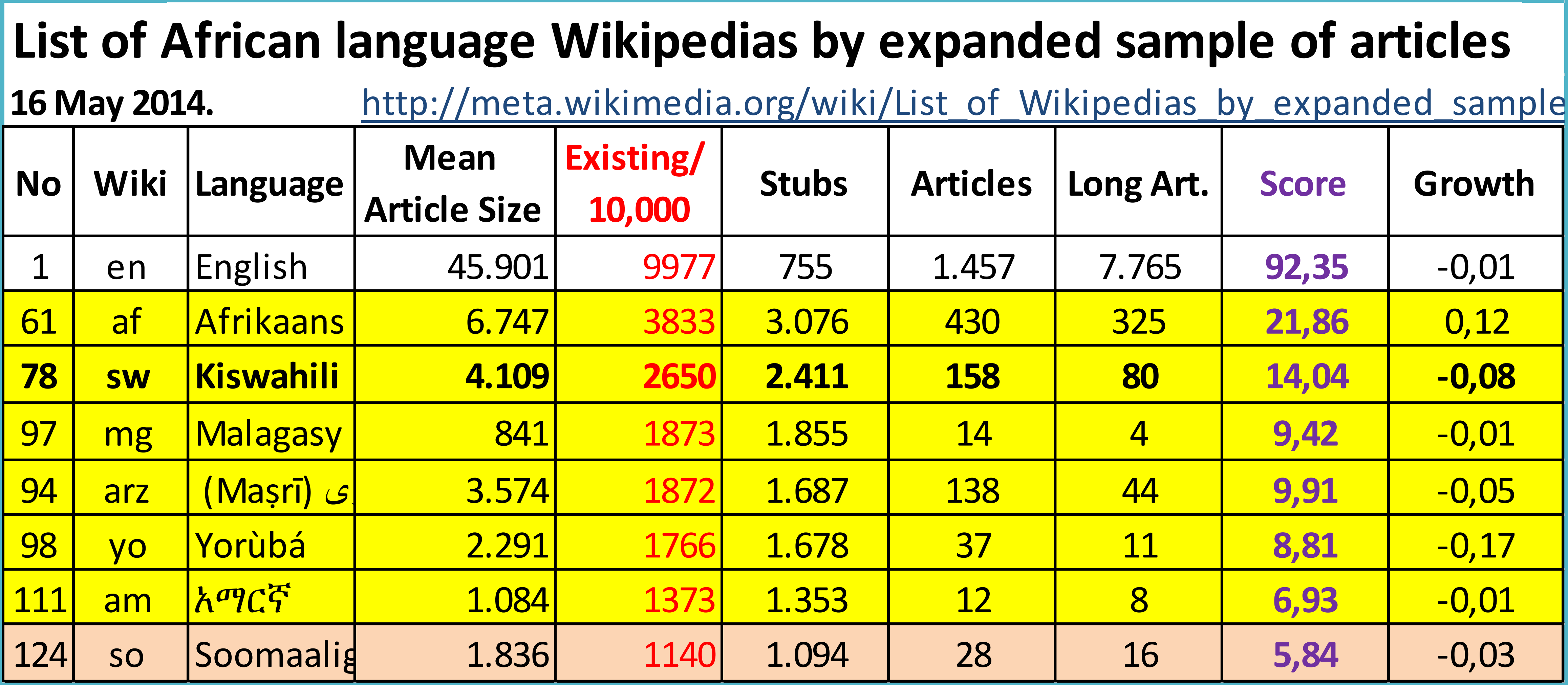
List of Wikipedias by sample of articles

This list is based on the "List of 1,000 articles every Wikipedia should have" as a sample.

№	Wiki	Language	Mean Article Size	Existing/ 1.000	Stubs	Articles	Long Art.	Score
3	en	English	78.202	1000	8	126	866	92,29
57	af	Afrikaans	9.921	819	625	134	60	18,90
79	sw	Kiswahili	5.181	814	743	52	19	12,47
102	arz	(Maṣrī) رى	2.950	639	609	22	8	8,54
120	yo	Yorùbá	3.006	469	443	10	3	5,74
125	so	Soomaalig	2.708	401	384	10	6	5,32
127	am	አማርኛ	1.302	450	444	6	0	5,20
132	mg	Malagasy	1.683	394	386	7	1	4,70

Afrikaans and Swahili are well on the way !

What is in it? Encyclopedial content by „expanded list of 10.000 most important articles“



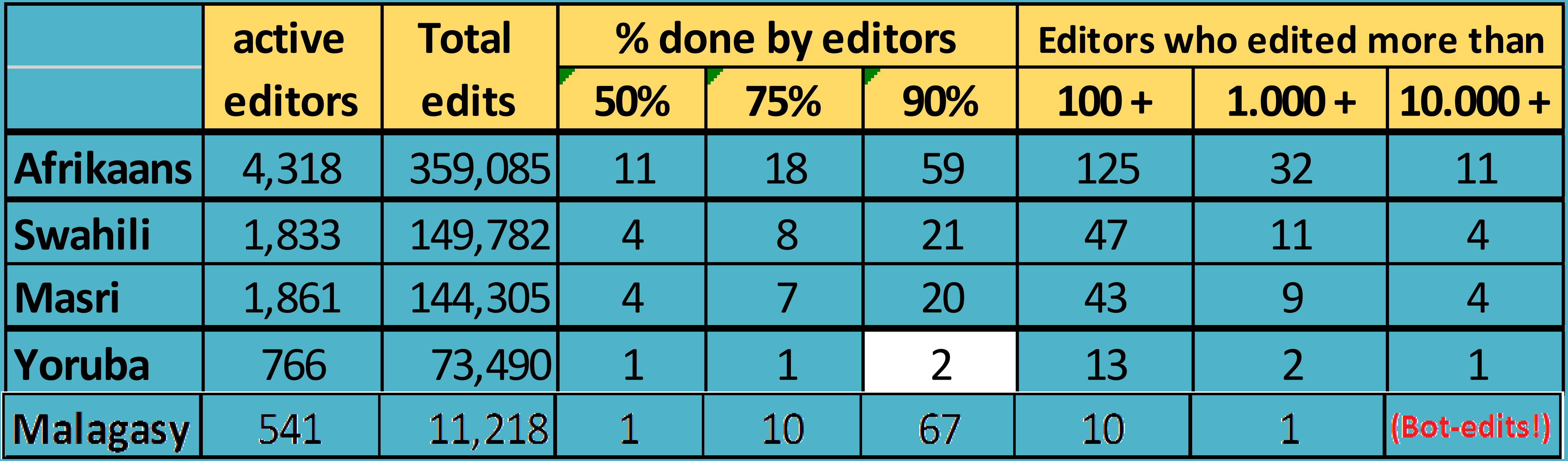
List of African language Wikipedias by expanded sample of articles

16 May 2014. http://meta.wikimedia.org/wiki/List_of_Wikipedias_by_expanded_sample

No	Wiki	Language	Mean Article Size	Existing/ 10,000	Stubs	Articles	Long Art.	Score	Growth
1	en	English	45.901	9977	755	1.457	7.765	92,35	-0,01
61	af	Afrikaans	6.747	3833	3.076	430	325	21,86	0,12
78	**sw**	**Kiswahili**	**4.109**	**2650**	**2.411**	**158**	**80**	**14,04**	**-0,08**
97	mg	Malagasy	841	1873	1.855	14	4	9,42	-0,01
94	arz	(Maṣrī) ى	3.574	1872	1.687	138	44	9,91	-0,05
98	yo	Yorùbá	2.291	1766	1.678	37	11	8,81	-0,17
111	am	አማርኛ	1.084	1373	1.353	12	8	6,93	-0,01
124	so	Soomaalig	1.836	1140	1.094	28	16	5,84	-0,03

Who is doing it ? Editors!

A look at some A.L.W. s



	active	Total	% done by editors			Editors who edited more than		
	editors	edits	50%	75%	90%	100 +	1.000 +	10.000 +
Afrikaans	4,318	359,085	11	18	59	125	32	11
Swahili	1,833	149,782	4	8	21	47	11	4
Masri	1,861	144,305	4	7	20	43	9	4
Yoruba	766	73,490	1	1	2	13	2	1
Malagasy	541	11,218	1	10	67	10	1	(Bot-edits!)

Data cover the full period of activity in 2012 and not present levels

Yoruba: partly a 1-man-show: editor „Demmy“ did more than 60,000 edits („Wikipedian of year 2012“)

Who is doing it: editors, human or bot?

Share of Bot-created articles on ALWs

	Malagasy	Yoruba	Afrikaans	Swahili	Masri	Amharic
	mg	yo	af	sw	arz	am
Σ total articles created	47 k	31 k	31 k	26 k	12 k	13 k
Σ manually created articles	2.6 k	31 k	31 k	26 k	12 k	11 k
Σ articles created by bots	**45 k**	**-**	**1**	**96**	**-**	**2.0 k**
Share of articles created by bots	**94%**	**-**	**0%**	**0%**	**-**	**15%**

- *Afrikaans: Yoruba, Masri: no bots, all manually created
- *Malagasy: nearly all bot,
- *Afrikaans, Swahili: minimal bot input

What does it say about quality?

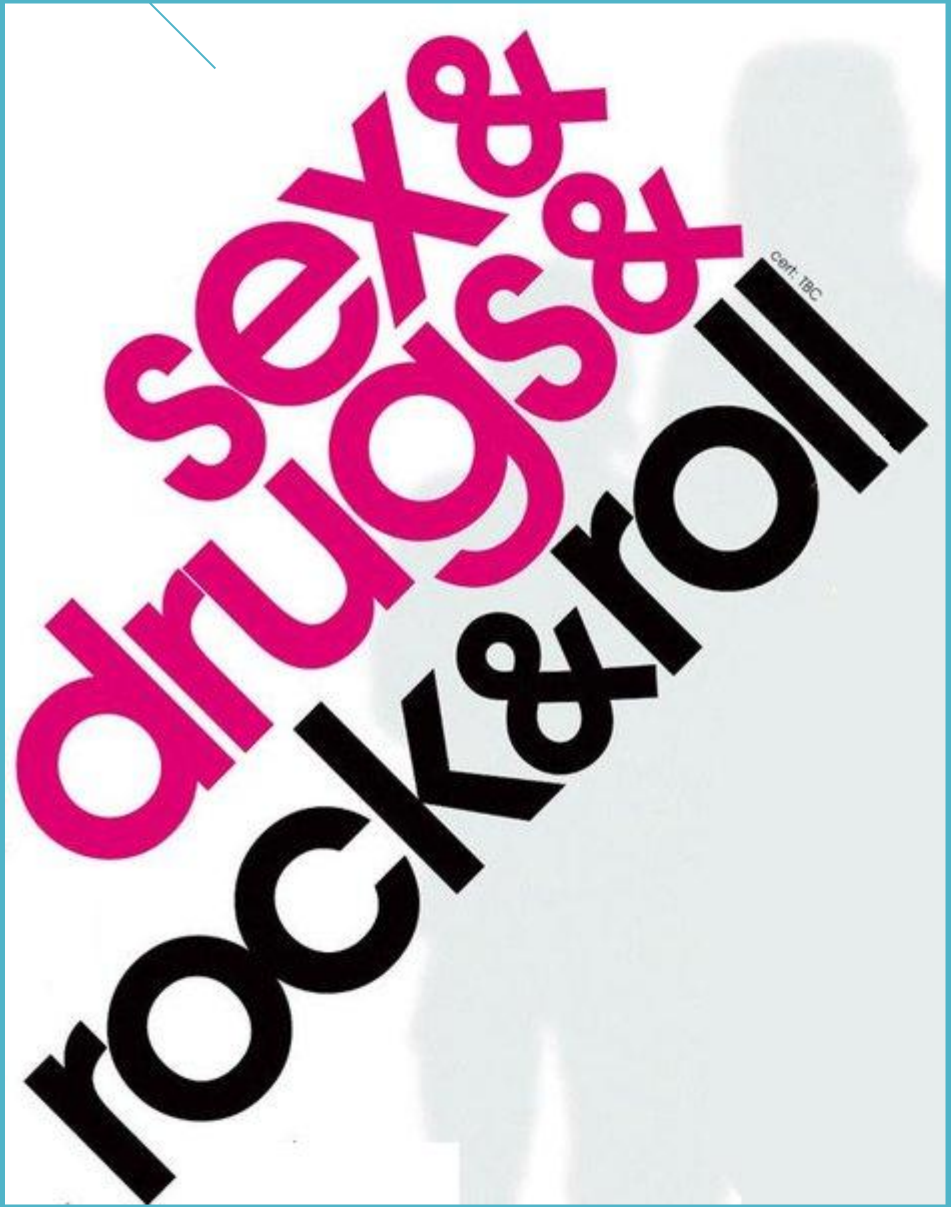
What does it say about quality?

- Bot created articles are not necessarily bad quality

- if there is a good database prepared with several informations a bot can do short articles with a respectable amount of basic information

- very short manually articles can be rather useless as shown in the example shown above

- This bandwith is also valid in the ALWs

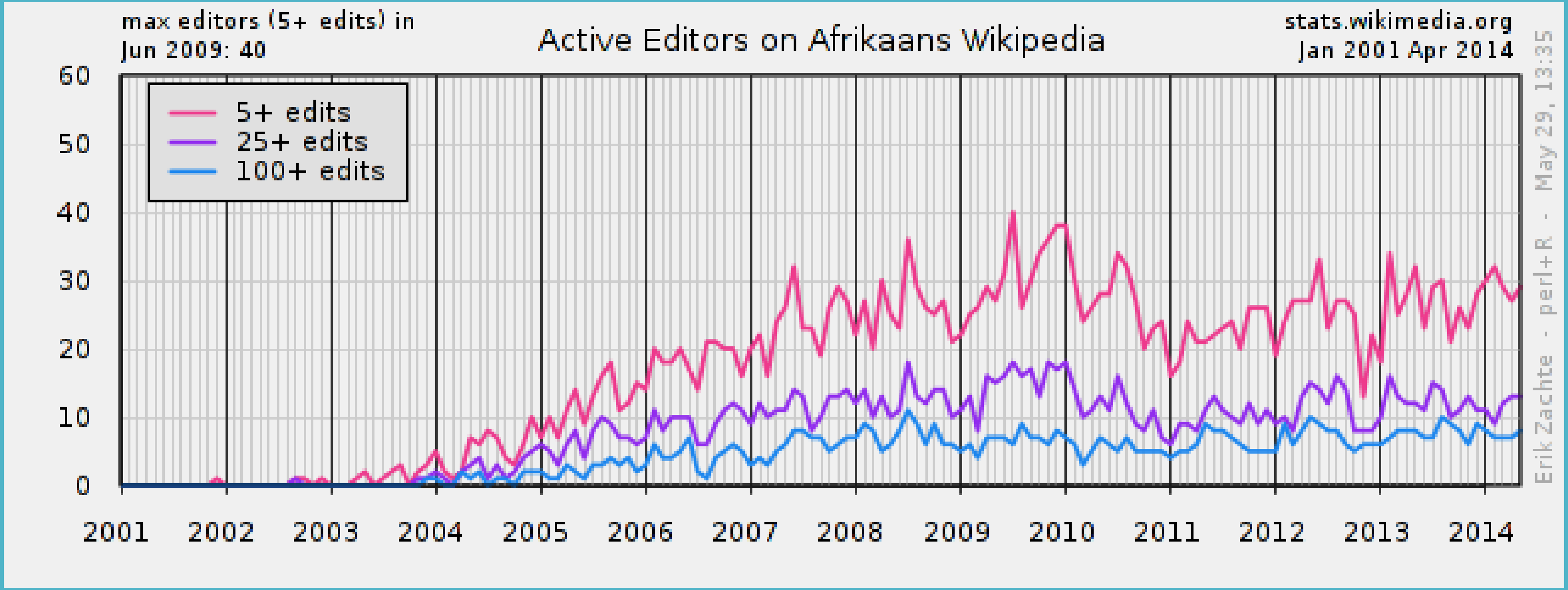


Users love these topics
On en:wikipedia: 40-50% of hits
on „**pages are related to entertainment and sexuality**”

But **Editors ?**

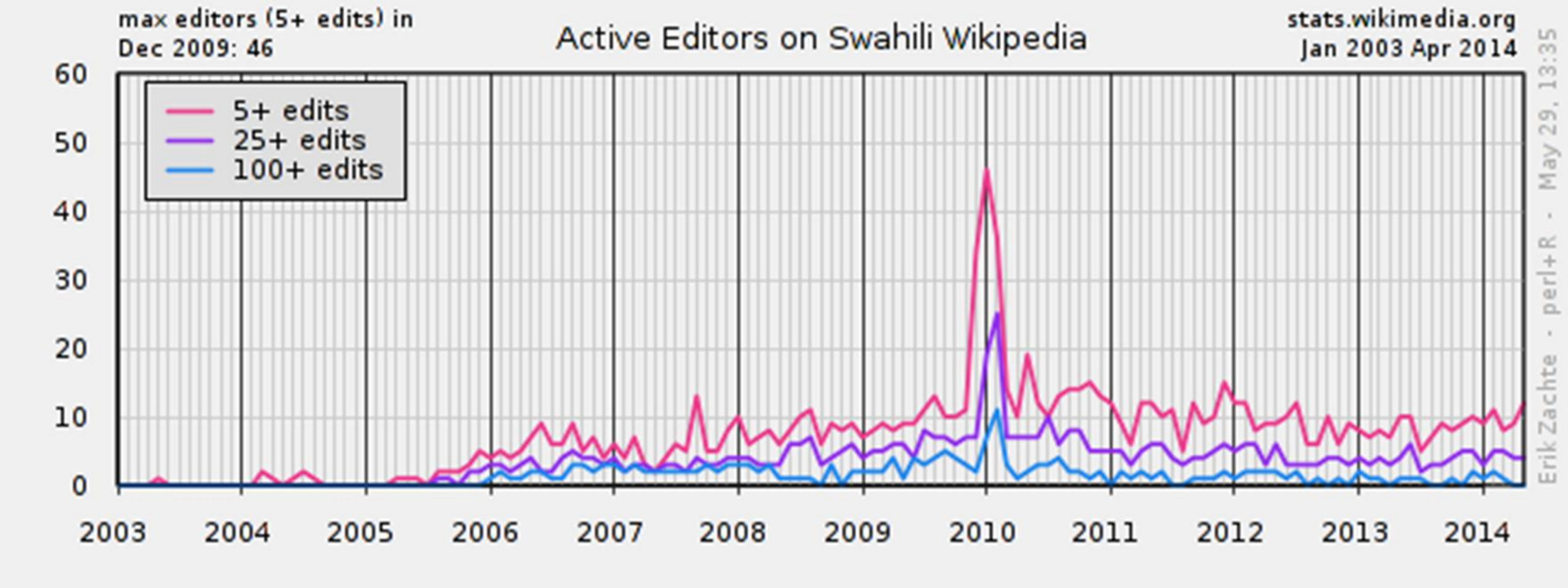
.. have no time for it when they want to bring their wikipedia forward!

Who does the job? Editor activity on af



Example of Afrikaans: edit community is large enough not to be effected by 1 or 2 editors opting out

Who does the job? Editor activity on sw



In comparison Swahili looks more fragile . . - 1 or 2 or main editors leaving could hold back further growth and maintenance considerably

Who does the job: Where are the editors?

Edits on ALWs originatinmg from different countries

Afrikaans (af)	
South Africa	73,40%
Netherlands	7,90%
New Zealand	1,90%
United States	1,80%
United Kingdom	1,20%
Other	13,80%

Swahili (sw)	
Tanzania	30,10%
Europe Unspecified	24,30%
Iran	15,30%
Canada	5,60%
Italy	4,60%
United States	3,60%
United Kingdom	3,40%
Other Kenya???	13,10%

Shona (sn) 4,742	
Canada	81,60%
United States	4,20%
Zimbabwe	4,20%
Czech Republic	3,70%
Brazil	3,20%
Other	3,10%

Homesick Africans abroad writing in home language?

Amharic (am)	
United States	47,20%
Ethiopia	18,90%
Germany	5,40%
Other	28,50%

Just globalization?

Or a sharper look from the outside at chances of language development?

Kipala: Wikipedias in African languages, 20.06.2014

Who is reading it ?

We go to Page Views, the gold of the internet!

Caveat: One million views is not 1 million viewers!

in general: 15% - 20% crawler traffic

http://stats.wikimedia.org/wikimedia/squids/SquidReportCrawlers.htm

Crawlers help wikipedia:

For example example in 1/3 of all external lookups via google (google crawler: ca. 3-4 %)

(if they would not crawl they could not refer to us!)

Pageviews for A.L.W.s I

	Language	Prim.+Sec. Speakers	Views per hour	views per day	per month
af	Afrikaans	13M	6568	157.632	**4.728.960**
arz	Masri	76M	3262	78.288	**2.348.640**
sw	Swahili	50M	2499	59.976	**1.799.280**
am	Amharic	25M	931	22.344	**670.320**
mg	Malagasy	20M	826	19.824	**594.720**
so	Somali	14M	615	14.760	**442.800**
ig	Igbo	22M	593	14.232	**426.960**
ln	Lingala	25M	279	6.696	**200.880**
sn	Shona	7M	244	5.856	**175.680**

The 3 ALWs with more than 1 mio views per month:
Afrikaans is not surprising,
Masri (12.000 + articles) only has a very large population with relatively good internet access
Swahili: overtook "larger" ALWs because of quality ?

Wikipedia Statistics
Wednesday April 30, 2014
http://stats.wikimedia.org/EN/Sitemap.htm

Pageviews for A.L.W.s II

	Language	Prim.+Sec. Speakers	Views per hour	views per day	per month
ee	Ewe	4M	152	3.648	**109.440**
ha	Hausa	39M	149	3.576	**107.280**
ts	Tsonga	3M	135	3.240	**97.200**
om	Oromo	26M	124	2.976	**89.280**
yo	Yoruba	25M	115	2.760	**82.800**
ss	Siswati	3M	115	2.760	**82.800**
tn	Setswana	4M	101	2.424	**72.720**
ff	Fulfulde	13M	99	2.376	**71.280**
tw	Twi	15M	95	2.280	**68.400**
ki	Kikuyu	5M	90	2.160	**64.800**

Yoruba is down in views in spite of 31.000 articles – problem of quality ?

Wikipedia Statistics
Wednesday April 30, 2014
http://stats.wikimedia.org
/EN/Sitemap.htm

Who is reading it ?

Afrikaans (af)	
South Africa	63,20%
United States	4,50%
Netherlands	3,60%
Other	28,70%

Swahili (sw)	
Tanzania	37,40%
United States	22,30%
Kenya	5,30%
United Kingdom	5,30%
Other	29,70%

Amharic (am)	
Ethiopia	22,90%
United States	19,30%
Australia	6,00%
Germany	3,10%
Other	48,70%

Oromo (om)	
United States	41,00%
United Kingdom	12,00%
Canada	9,00%
Ethiopia	7,00%
Other	31,00%

Bambara (bm)	
United States	41,50%
China	33,00%
Switzerland	5,70%
Other	19,80%

Somali (so)	
United States	16,10%
United Kingdom	9,60%
Sweden	8,70%
Norway	7,40%
China	5,50%
Djibouti	5,10%
Other	47,60%

ALWs views from countries

page views per language

Yoruba, Malagasy not available

http://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerLanguageBreakdown.htm#errata

Who is reading it ? – marketshares . . .

South Africa

Quarter	English	**Afrikaans**
2014 Q1		**2.1%**
2013 Q4	93.2%	2.0%
2013 Q3	92.9%	2.5%

Tanzania

Quarter	English	**Swahili**
2014 Q1		**3.2%**
2013 Q4	86.7%	5.0%
2013 Q3	87.9%	4.1%

Ethiopia		
Quarter	English	**Amharic**
2014 Q1		2.0 %
2013 Q4	91.1%	1.6%
2013 Q3	91.7%	1.8%

Namibia			
Quarter	English	German	**Afrikaans**
2014 Q1	**87%**	**4.5%**	**1.8%**
2013 Q4	86.9%	5.3%	2.1%
2013 Q3	85.5%	5.5%	1.4%

Vast majority of African wikipedia users go to English (or French) .

Somalia

Quarter	English	Arabic	**Somali**
2014 Q1			4.1%
2013 Q4	85.6%	12.2%	1.1%
2013 Q3	83.9%	8.4%	3.2%
2013 Q2	82.5%	7.7%	2.8%
2013 Q1	84.2%	5.8%	3.2%
2012 Q4	86.7%	5.3%	4.2%

Swahili is relatively (a bit) stronger in Tanzania.

http://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerCountryTrends.htm

Observations and assumptions

Perspectives for ALWs
*potential ahead, internet access will grow in Africa

- Some countries will enhance AL use in education
- More interest in AL content is foreseeable

- Can this potential be realized?
- depends on language politics in education

Challenges for ALWs

The near monopoly of European languages as medium of higher education (not discussing role of Arabic)

- Inability of most educated Africans to discuss scientific etc. topics in mother tongue & "national language"

- Little faith in ability of ALs to express complex issues
- Reduced ability to read content in ALWs

Closing remarks

*Afrikaans today – an ALW with a healthy structure

(Afrikaans is the only AL medium in higher education)

- Several ALWs have large potential if they can offer quality

- Try to build ALWs in languages used in education systems & make them useful as reference !

- Twende jamani, kazi iko! (Lets go, there’s a job to be done)