`siganalogies`, morphological analogies on Sigmorphon 2016 and Sigmorphon 2019

January 2022

Contents

1 Morphological Analogies

An analogical proportion is defined as a 4-ary relation written $A : B :: C : D$ and which reads "A is to B as C is to D". In this dataset, we manipulate morphological analogies, *i.e.*, on analogies involving character strings, where the transformations between the objects correspond to morphological transformations of words (*e.g.*, conjugation or declension). In our dataset, A, B, C, and D are words.

The original data of each language is composed of pairs of words $\langle A, B\rangle$ (ex: $\langle$"do","doing"$\rangle$) with B the morphological transformation of A to obtain a set of morphological features F (ex: present participle). For example in our Finnish data of Sigmorphon 2016, we have $A =$"lenkkitossut", $B =$"lenkkitossuilla", and $F =$"pos=N, case=ON+ESS, num=PL" (the transformation corresponds to the nominative to essive cases of a noun for the plural). For any two pairs of words $\langle A, B\rangle, \langle A', B'\rangle$ that have the same set of features ($F = F'$), we consider $A : B :: A' : B'$ an analogical proportion. Note that for each two pairs, we only generate one analogy, *i.e.*, if we generate $A : B :: A' : B'$ we do not generate $A' : B' :: A : B$ as it will be generated by the data augmentation process. Also, analogies of the form $A : B :: A : B$ will be generated as the set of features is the same ($F = F$).

2 About Sigmorphon 2016

The Sigmorphon 2016 [1] dataset can be found at `https://github.com/ryancotterell/sigmorphon2016`. The data used here is from the task 1. More information about the dataset can be found in [1]. Japanese data has been extracted from the Japanese Bigger Analogy Test Set [2].

From the languages present in Sigmorphon 2016, 7 are available as high resource languages of Sigmorphon 2019 (Arabic, Finnish, German, Hungarian, Russian, Spanish, and Turkish) and 2 as low resource languages of Sigmorphon 2019 (Maltese and Russian). Note that Russian is among the languages with both high and low resource in Sigmorphon 2019.

Tables 1 to 3 describe the Sigmorphon 2016 data.

Language	# Analogies	# Features with analogies (% of all features)	# Words with analogies (% of vocabulary)
Arabic	373240	220 (98.65%)	13773 (99.97%)
Finnish	1342639	94 (98.95%)	22057 (99.99%)
Georgian	3553763	90 (100.00%)	14587 (100.00%)
German	994740	97 (98.98%)	17307 (99.99%)
Hungarian	3280891	85 (98.84%)	17279 (99.99%)
Maltese	104883	2419 (75.97%)	19338 (95.38%)
Navajo	502637	42 (77.78%)	4502 (99.80%)
Russian	1965533	80 (96.39%)	18793 (99.97%)
Spanish	1425838	83 (98.81%)	17145 (99.99%)
Turkish	606873	179 (95.72%)	14223 (99.94%)
Japanese	26410	20 (100.00%)	1573 (100.00%)

Table 1: Statistics of languages from Sigmorphon 2016, for the training data. Languages in bold are also present in Sigmorphon 2019.

Language	# Analogies	# Features with analogies (% of all features)	# Words with analogies (% of vocabulary)
Arabic	7671	218 (99.09%)	2638 (99.89%)
Finnish	22837	73 (84.88%)	3070 (99.16%)
Georgian	67457	48 (70.59%)	2716 (99.27%)
German	17222	97 (98.98%)	2888 (99.93%)
Hungarian	70565	76 (92.68%)	3517 (99.80%)
Maltese	3775	585 (39.24%)	2288 (67.20%)
Navajo	33976	42 (91.30%)	1578 (99.81%)
Russian	32214	77 (100.00%)	2898 (100.00%)
Spanish	25590	83 (100.00%)	2836 (100.00%)
Turkish	11518	160 (95.81%)	2691 (99.56%)

Table 2: Statistics of languages from Sigmorphon 2016, for the development data. Languages in bold are also present in Sigmorphon 2019.

Language	# Analogies	# Features with analogies (% of all features)	# Words with analogies (% of vocabulary)
Arabic	555312	220 (97.35%)	15996 (99.96%)
Finnish	4691453	95 (100.00%)	37857 (100.00%)
Georgian	8368323	90 (100.00%)	19722 (100.00%)
German	1480256	98 (98.99%)	19954 (99.99%)
Hungarian	66195	78 (95.12%)	3448 (99.88%)
Maltese	3707	597 (39.62%)	2315 (67.53%)
Navajo	4843	35 (83.33%)	618 (99.36%)
Russian	6421514	80 (96.39%)	29868 (99.99%)
Spanish	4794504	83 (100.00%)	28230 (100.00%)
Turkish	11360	161 (96.41%)	2675 (99.59%)

Table 3: Statistics of languages from Sigmorphon 2016, for the test data. Languages in bold are also present in Sigmorphon 2019.

3 About Sigmorphon 2019

The dataset can be found at `https://github.com/sigmorphon/2019`. The data used here is from the task 1.

The following languages are available as both **high** and **low** resource languages: Bengali, Czech, Greek, Irish, Latin, Portuguese, Russian, Sorani, and Swahili.

From the languages present in Sigmorphon 2016, 7 are available as high resource languages (Arabic, Finnish, German, Hungarian, Russian, Spanish, and Turkish) and 2 as low resource languages (Maltese and Russian). Note that Russian is among the languages with both high and low resource in Sigmorphon 2019.

All low resource languages have less than 25000 analogies in each set, with less than 900 analogies in the training set. Except Basque and Uzbek which have 43754 and 7312 analogies respectively, all high resource languages have at least 133000 analogies.

The set 42 languages (high resource except Basque and Uzbek) will be the one considered for our experiments.

Tables 4 to 7 describe the Sigmorphon 2019 data.

	# Analogies	# Features with analogies	# Words with analogies
Language		(% of all features)	(% of vocabulary)
Adyghe	3666973	24 (100.00%)	11155 (100.00%)
Albanian	378591	140 (100.00%)	9666 (100.00%)
Arabic	456689	196 (100.00%)	12942 (100.00%)
Armenian	391054	220 (100.00%)	14415 (100.00%)
Asturian	608932	139 (74.33%)	9122 (99.61%)
Bashkir	3912246	24 (100.00%)	9231 (100.00%)
Basque	43754	1584 (95.77%)	8918 (99.34%)
Belarusian	1025983	56 (100.00%)	8769 (100.00%)
Bengali	163424	58 (100.00%)	3759 (100.00%)
Bulgarian	593920	95 (100.00%)	11290 (100.00%)
Czech	598680	180 (94.24%)	12056 (99.90%)
Danish	7274570	14 (100.00%)	11205 (100.00%)
Dutch	2031211	25 (100.00%)	11181 (100.00%)
English	10006487	5 (100.00%)	16245 (100.00%)
Estonian	641478	108 (100.00%)	10262 (100.00%)
Finnish	508684	197 (100.00%)	18231 (100.00%)
French	1029926	49 (100.00%)	15220 (100.00%)
German	2108502	37 (100.00%)	13174 (100.00%)
Greek	811576	177 (100.00%)	13668 (100.00%)
Hebrew	1095028	54 (100.00%)	8957 (100.00%)
Hindi	246605	211 (100.00%)	8916 (100.00%)
Hungarian	1062552	93 (100.00%)	16747 (100.00%)
Irish	2248336	89 (100.00%)	13169 (100.00%)
Italian	990860	51 (100.00%)	16016 (100.00%)
Kannada	133094	95 (100.00%)	3049 (100.00%)
Kurmanji	2836118	104 (98.11%)	16209 (99.99%)
Latin	447718	151 (100.00%)	16141 (100.00%)
Latvian	984308	80 (100.00%)	14127 (100.00%)
Persian	375639	136 (100.00%)	9323 (100.00%)
Polish	1023982	111 (100.00%)	14779 (100.00%)
Portuguese	668308	76 (100.00%)	12921 (100.00%)
Romanian	945689	59 (100.00%)	12380 (100.00%)
Russian	978081	89 (91.75%)	17227 (99.94%)
Sanskrit	822359	120 (100.00%)	8473 (100.00%)
Slovak	2026778	39 (100.00%)	7442 (100.00%)
Slovene	900301	99 (100.00%)	10189 (100.00%)
Sorani	246077	244 (97.99%)	10158 (99.95%)
Spanish	725601	70 (100.00%)	14445 (100.00%)
Swahili	207967	207 (100.00%)	6419 (100.00%)
Turkish	304609	288 (96.00%)	12650 (99.91%)
Urdu	245343	217 (100.00%)	5192 (100.00%)
Uzbek	7312	84 (100.00%)	936 (100.00%)
Welsh	799086	63 (100.00%)	8820 (100.00%)
Zulu	348500	228 (98.70%)	9613 (99.97%)

Table 4: Statistics of high resource languages from Sigmorphon 2019. Languages in bold are also present in Sigmorphon 2016. Note that only a training set is available for high resource languages. Languages in red have less than 50000 analogies.

Language	# Analogies	# Features with analogies (% of all features)	# Words with analogies (% of vocabulary)
Azeri	247	20 (47.62%)	144 (81.82%)
Bengali	183	28 (54.90%)	125 (80.65%)
Breton	196	29 (63.04%)	119 (86.86%)
Classical-Syriac	333	18 (58.06%)	145 (92.36%)
Cornish	166	26 (45.61%)	73 (70.19%)
Crimean-Tatar	809	8 (66.67%)	176 (96.17%)
Czech	148	23 (34.85%)	110 (57.89%)
Friulian	199	32 (78.05%)	164 (90.61%)
Greek	162	21 (33.33%)	107 (56.61%)
Ingrian	320	21 (84.00%)	137 (96.48%)
Irish	328	17 (45.95%)	140 (77.78%)
Kabardian	450	14 (82.35%)	168 (98.25%)
Karelian	202	30 (69.77%)	102 (90.27%)
Kashubian	447	14 (100.00%)	109 (100.00%)
Kazakh	446	14 (100.00%)	115 (100.00%)
Khakas	397	15 (93.75%)	139 (99.29%)
Ladin	222	26 (63.41%)	144 (84.71%)
Latin	137	23 (32.39%)	103 (52.02%)
Lithuanian	218	20 (37.74%)	128 (67.72%)
Livonian	211	23 (47.92%)	132 (80.49%)
Maltese	402	16 (94.12%)	165 (98.80%)
Middle-High-German	232	26 (68.42%)	91 (87.50%)
Middle-Low-German	239	26 (72.22%)	125 (91.91%)
Murrinhpatha	226	29 (82.86%)	108 (95.58%)
Neapolitan	187	35 (83.33%)	129 (94.85%)
North-Frisian	187	32 (69.57%)	108 (90.76%)
Occitan	190	32 (72.73%)	149 (87.65%)
Old-Church-Slavonic	340	20 (95.24%)	162 (98.78%)
Old-English	162	27 (47.37%)	138 (70.41%)
Old-Irish	156	23 (35.94%)	94 (67.63%)
Old-Saxon	194	20 (35.71%)	123 (66.85%)
Pashto	156	27 (45.00%)	113 (71.97%)
Portuguese	164	28 (49.12%)	139 (72.02%)
Quechua	115	11 (12.64%)	48 (25.26%)
Russian	193	25 (53.19%)	147 (77.37%)
Scottish-Gaelic	382	17 (89.47%)	105 (98.13%)
Sorani	121	14 (16.87%)	59 (35.12%)
Swahili	124	20 (25.64%)	71 (46.41%)
Tatar	863	6 (66.67%)	172 (96.63%)
Telugu	248	8 (61.54%)	66 (92.96%)
Turkmen	497	12 (100.00%)	144 (100.00%)
Votic	282	23 (88.46%)	139 (97.20%)
West-Frisian	341	20 (100.00%)	124 (100.00%)
Yiddish	357	19 (90.48%)	160 (98.16%)

Table 5: Statistics of low resource languages from Sigmorphon 2019, for the training data. Languages in bold are also present in Sigmorphon 2016.

	# Analogies	# Features with analogies	# Words with analogies
Language		(% of all features)	(% of vocabulary)
Azeri	256	22 (48.89%)	141 (81.98%)
Bengali	196	29 (64.44%)	138 (86.25%)
Breton	166	30 (54.55%)	113 (81.29%)
Classical-Syriac	370	16 (64.00%)	162 (94.19%)
Cornish	65	11 (29.73%)	31 (53.45%)
Crimean-Tatar	760	8 (66.67%)	175 (96.15%)
Czech	6982	146 (89.02%)	1687 (98.65%)
Friulian	195	34 (82.93%)	158 (94.05%)
Greek	8338	123 (79.35%)	1745 (96.78%)
Ingrian	100	16 (84.21%)	78 (96.30%)
Irish	22975	76 (88.37%)	1715 (98.96%)
Kabardian	522	13 (86.67%)	171 (98.28%)
Karelian	70	11 (31.43%)	39 (58.21%)
Kashubian	139	11 (78.57%)	64 (94.12%)
Kazakh	128	13 (92.86%)	67 (98.53%)
Khakas	119	14 (93.33%)	76 (98.70%)
Ladin	189	30 (65.22%)	151 (86.78%)
Latin	5129	145 (96.67%)	1914 (99.53%)
Lithuanian	8421	127 (93.38%)	1604 (99.44%)
Livonian	168	24 (41.38%)	114 (74.03%)
Maltese	411	15 (88.24%)	150 (98.04%)
Middle-High-German	93	15 (62.50%)	48 (84.21%)
Middle-Low-German	80	16 (57.14%)	64 (80.00%)
Murrinhpatha	85	14 (51.85%)	60 (81.08%)
Neapolitan	209	30 (73.17%)	123 (91.79%)
North-Frisian	200	29 (64.44%)	112 (88.19%)
Occitan	199	28 (65.12%)	151 (87.28%)
Old-Church-Slavonic	325	21 (100.00%)	157 (100.00%)
Old-English	9011	89 (95.70%)	1608 (99.57%)
Old-Irish	61	7 (17.07%)	29 (38.67%)
Old-Saxon	8665	125 (86.81%)	1402 (98.94%)
Pashto	149	27 (42.86%)	112 (69.14%)
Portuguese	7417	76 (100.00%)	1816 (100.00%)
Quechua	2230	260 (64.84%)	1332 (89.40%)
Russian	11380	68 (90.67%)	1879 (99.31%)
Scottish-Gaelic	125	12 (63.16%)	57 (87.69%)
Sorani	3362	212 (93.39%)	1175 (98.74%)
Swahili	127	19 (24.68%)	75 (51.72%)
Tatar	794	9 (75.00%)	173 (97.74%)
Telugu	170	8 (57.14%)	58 (90.62%)
Turkmen	166	10 (83.33%)	75 (94.94%)
Votic	278	24 (96.00%)	138 (99.28%)
West-Frisian	366	18 (94.74%)	137 (99.28%)
Yiddish	346	17 (73.91%)	158 (94.61%)

Table 6: Statistics of low resource languages from Sigmorphon 2019, for the development data. Languages in bold are also present in Sigmorphon 2016.

	# Analogies	# Features with analogies	# Words with analogies
Language		(% of all features)	(% of vocabulary)
Azeri	261	18 (41.86%)	137 (77.84%)
Bengali	190	32 (71.11%)	148 (88.10%)
Breton	197	28 (58.33%)	115 (84.56%)
Classical-Syriac	341	19 (76.00%)	161 (95.27%)
Cornish	79	12 (38.71%)	38 (67.86%)
Crimean-Tatar	869	6 (54.55%)	170 (94.44%)
Czech	6810	133 (82.61%)	1668 (97.20%)
Friulian	202	29 (69.05%)	146 (88.48%)
Greek	8380	121 (76.58%)	1735 (96.44%)
Ingrian	96	15 (68.18%)	69 (88.46%)
Irish	24330	71 (87.65%)	1698 (99.01%)
Kabardian	430	15 (78.95%)	178 (97.27%)
Karelian	78	11 (35.48%)	42 (68.85%)
Kashubian	137	12 (92.31%)	70 (98.59%)
Kazakh	129	14 (100.00%)	67 (100.00%)
Khakas	129	11 (68.75%)	72 (90.00%)
Ladin	196	29 (64.44%)	150 (86.71%)
Latin	5449	142 (94.67%)	1907 (99.17%)
Lithuanian	9070	123 (93.89%)	1606 (99.38%)
Livonian	212	17 (32.69%)	114 (69.51%)
Maltese	401	16 (94.12%)	160 (98.77%)
Middle-High-German	94	15 (62.50%)	52 (85.25%)
Middle-Low-German	79	15 (53.57%)	59 (77.63%)
Murrinhpatha	90	16 (66.67%)	59 (89.39%)
Neapolitan	200	33 (82.50%)	128 (94.81%)
North-Frisian	193	29 (64.44%)	110 (89.43%)
Occitan	214	27 (62.79%)	154 (88.00%)
Old-Church-Slavonic	323	20 (95.24%)	144 (99.31%)
Old-English	9333	91 (98.91%)	1593 (99.94%)
Old-Irish	58	8 (19.05%)	27 (38.57%)
Old-Saxon	8671	123 (87.23%)	1387 (98.86%)
Pashto	175	18 (28.57%)	95 (59.38%)
Portuguese	7573	76 (100.00%)	1826 (100.00%)
Quechua	2294	245 (61.71%)	1327 (88.17%)
Russian	10246	69 (90.79%)	1880 (99.37%)
Scottish-Gaelic	121	15 (83.33%)	69 (97.18%)
Sorani	3313	209 (89.70%)	1168 (98.07%)
Swahili	129	23 (31.08%)	83 (54.97%)
Tatar	781	8 (72.73%)	177 (97.79%)
Telugu	124	10 (50.00%)	51 (85.00%)
Turkmen	146	11 (100.00%)	85 (100.00%)
Votic	279	23 (88.46%)	137 (97.86%)
West-Frisian	364	17 (85.00%)	131 (97.76%)
Yiddish	359	18 (78.26%)	162 (94.74%)

Table 7: Statistics of low resource languages from Sigmorphon 2019, for the test data. Languages in bold are also present in Sigmorphon 2016.

4 Bilingual Analogies in Sigmorphon 2016 (Feature Coming Soon)

Multilingual analogies are not implemented yet.

5 Bilingual Analogies in Sigmorphon 2019 (Feature Coming Soon)

Multilingual analogies are not implemented yet.

We will consider two settings: (*i*) the analogies we can build from the high/low resource language pairs of Sigmorphon 2019, and (*ii*) the analogies we can build by making pairs of high resource languages.

References

[1] Ryan Cotterell et al. "The SIGMORPHON 2016 Shared Task—Morphological Reinflection". In: *Proceedings of the 2016 Meeting of SIGMORPHON*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016.

[2] Marzena Karpinska et al. "Subcharacter Information in Japanese embeddings: when is it worth it?" In: *Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*. Melbourne, Australia: ACL, 2018, pp. 28–37.

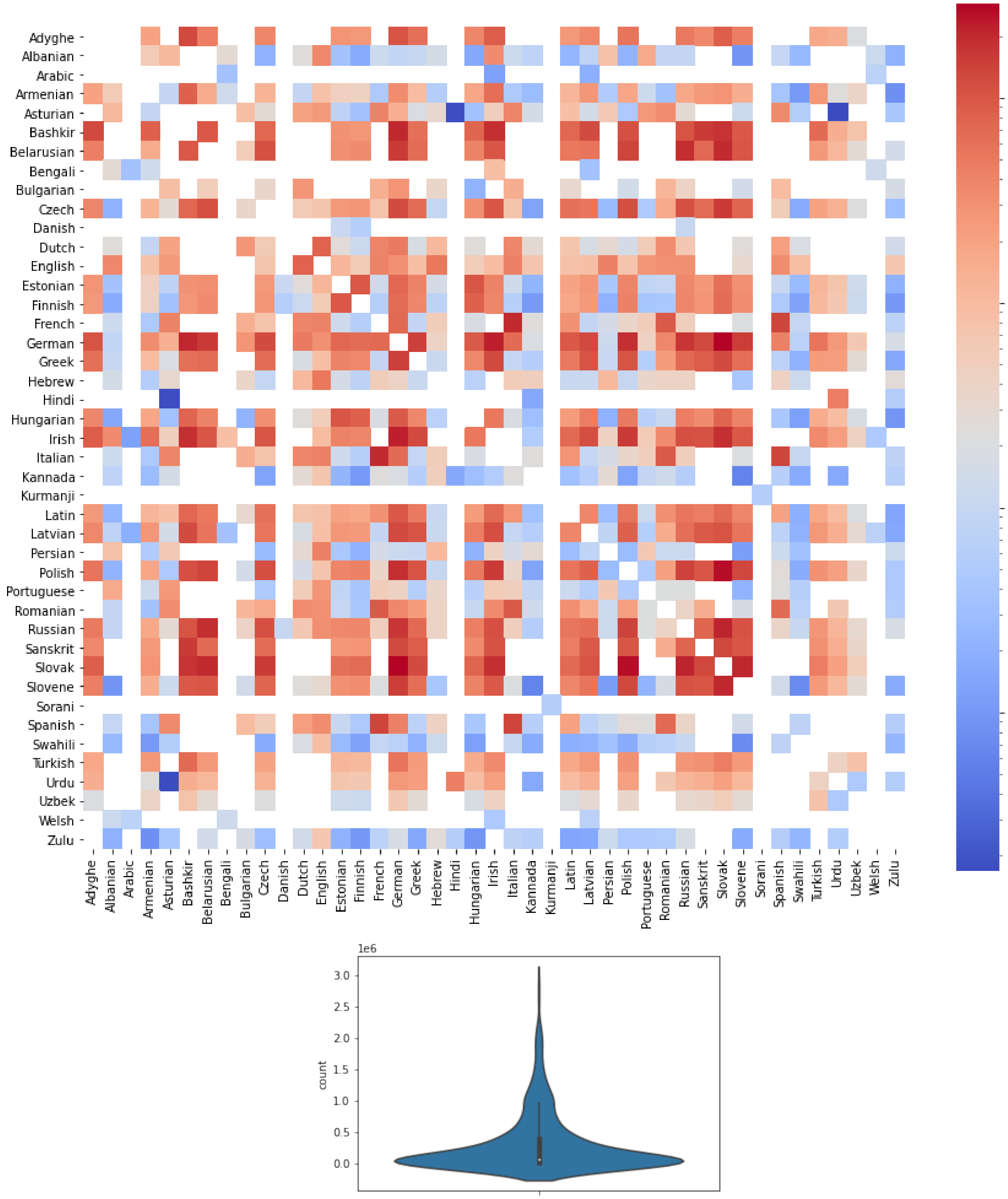

Figure 1: Number of analogies between high resource languages. Range is [168, 2866777].