The Workshops of the Tenth International AAAI Conference on Web and Social Media
Wiki: Technical Report WS-16-17

Topical Interest and Degree of Involvement of Bilingual Editors in Wikipedia

Sooyoung Kim
School of Computing
KAIST
Daejeon, Korea
sooyoungkim@kaist.ac.kr

Alice Oh
School of Computing
KAIST
Daejeon, Korea
alice.oh@kaist.edu

Abstract

Language reveals a lot of information about its speakers. Speakers of one language usually share common cultural habits or regional characteristics, and their similarities become more obvious within the context where there are multiple languages in use. We focus on studying bilingual users of Wikipedia, one of the largest multilingual user-generated content platforms. In Wikipedia, we can observe the patterns in the English edition, where users of multiple languages come together to express their thoughts and interests in the common language of English. To understand the specific topics edited by bilingual users, we analyze them in terms of revision counts, topics, and country names. We find that bilingual users are generally interested in more local topics, and their language is highly related with their topics. Also, we observe that the topical diversity decreases with the proportion of English edits, and more concentrates on topics related with countries and cultures.

Introduction

Collective intelligence has brought a huge change in knowledge sharing. Rather than spreading words of individual professionals, people can distribute their knowledge online and get reciprocal benefits. However, languages have become a barrier within this flood of information (Halavais 2000; Herring 2007). Not only is there a bias in search result, but also the lack of multilingual support in webpages hinder knowledge sharing. Wikipedia is one of the platforms that support many languages, in a form of language edition. With this, researches have been conducted to understand the characteristics of multilingual speakers.

(Hale 2014) has shown that multilingual users have unique characteristics in terms of high devotion to article creation, and (Kim et al. 2015) claims that among multilingual users, primary and non-primary users of English, Spanish, and German have similar interests, yet there are a few differences by topic and language. This phenomenon is not limited to Wikipedia. Researchers have found that users who can speak more than one language plays an important role in reducing homophily (informational in-breeding)(Kim et al. 2014; Eleta and Golbeck 2012) and information inequality(Etling et al. 2009; Otterbacher). Their online activities build bridges between two separate networks.

Bilingual users have unique position in online communities, but we need to examine them in more detailed way. One's language tells a lot of thing about the speaker(Gumperz 1982; Phinney et al. 2001; Ochs 1993). Ethnic group, culture or education that affects identity formulation is also significant when people learn first or second language. Multilingual produce a different impact on language communities, and examining the specific area that multilingual users focuses on is an important task. In editing Wikipedia in multiple languages, we expected that they would not act like a machine translator; there would be a consistent pattern that shows characteristics of a language. Thus, we performed individual analysis on six language group, Russian, Italian, Persian, Portuguese, Chinese, and Korean.

We conjecture that (1) the bilingual editors consistently edit articles that contain local or cultural information related to their language, and (2) this pattern is stronger with the editors who are not familiar with English.

Materials and Methods

Dataset

Wikipedia provides text and metadata of all revisions of all pages as XML file. We retain the list of 3,575,175 unique users who edited English Wikipedia more than once in 2015. After sampling 38,309 users from this list, we retrieve 500 recent edits of 110 languages of Wikipedia using the Wikipedia API. The edits include 117,559 articles edited in English and 1,855,923 articles are edited in 110 other languages.

Multilingual Users

We define a user knows a language if the user edited one or more Wikipedia articles in that language. Though a user can merely change the markup or layout, we assume that the contributions in a specific language context are possible if and only if the user knows that language. We can find that the number of languages a user edit decrease exponentially. 33,786 users edited only in English, followed by 3,369 users of two languages, and 583 users of three languages. In this paper, we only focuses on the users who edited only one or

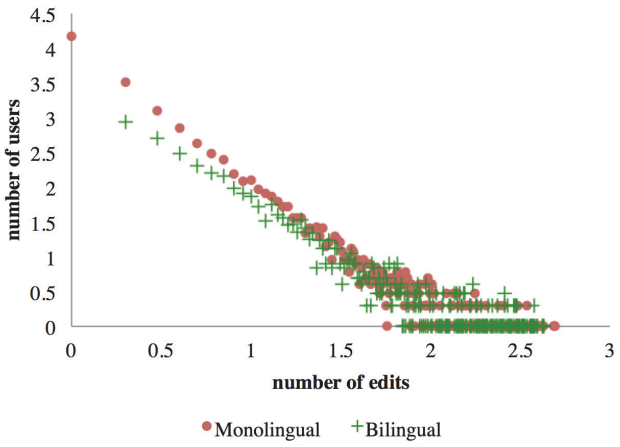

Figure 1: Log-log plot of the number of unique articles one user edited. Two user groups shows a linear relationship between the number of user and the number of articles, but multilingual users tend to edit more than monolingual users.

two languages, since most of multilingual users are bilingual in terms of their editing behavior and the primary language is ambiguous with trilingual and quad-lingual users.

Topic Extraction Using Latent Dirichlet Allocation

In order to observe the editing behaviors of Wikipedia users, we focus on the characteristics of articles a user edited. Since most users do not reveal information about themselves in their user page, the content of articles they actually made revisions tells much about a user. From the texts of articles edited by our user sample, we extract topics of Wikipedia articles using Latent Dirichlet Allocation. We set the number of topics to K = 50, and employ online variational inference algorithm to estimate parameters with maximum iteration count 200.

Appendix shows the result from LDA. We use this result in order to examine the interest of Wikipedia editors. Users are represented as the articles they edited, so each topic vector is averaged to exhibit topical interest of a user.

Topic Groups

Each topic represents the contents that users are interested in. However, not all topics are valid. There are topics that contain syntactic or meaningless words. Also, it is more convenient to cluster the topics into large groups in order to analyze the behavior of each user group.

Observing most probable word cloud in each topic, we manually classify the topics into two groups: country-related topic group, and country-independent general topic group. Country-related topic group consists of 36 topics that contains more than one region name, and general country-independent topics.

Among the country-related topics, there are three topics related to the United States (T1, T11, and T25 in Appendix). Contrary to other country-related topics, such as Greece, Russia or Pakistan, these topics can be frequently edited by all users. Thus, we also divide the U.S. topics and topics related to countries other that the U.S..

	Total	Monolingual	Bilingual
CI	0.813	**0.815**	0.809
CR	0.146	0.145	**0.151**
US	0.0452	**0.0483**	0.0338
non-US	0.102	0.097	**0.117**

Table 1: The difference in proportion of bilingual and monolingual users. (CI: Country-independent topics, CR: Country-related topics, US: topics related to the United States, non-US: topics related to countries other than the US.)

Results

Overall Activities

We compare the number of articles edited by monolingual and multilingual users. The average of monolingual users is 4.20; whereas the average of bilingual users is 22.91. Figure 1 shows the histogram of users by their number of edited articles. The log-log plot of histogram shows that multilingual users have overall tendency of editing more, which aligns with H. Scott (Hale 2014).

We also observed difference in topical interest of monolingual and multilingual users. We measure this by comparing monolingual and multilingual users with the total topic distribution. As shown in Table 1, multilingual users show high involvement in country country-related topics and low involvement in general, country-independent topic group. Also, among country-related topics, U.S. topics are popular among monolingual users, while multilingual users edits more on other countries, such as Greece, Israel, or Japan.

Per-Language Analysis

Although we detect bilingual users in 42 languages, some languages show distinct patterns in large scale. As shown in Appendix, a set of country names appears in one topic, and they are either linguistically or culturally similar. We sample 6 different languages from those topics, and measure the relationship between the language and user behaviors. We first measure the involvement in each topic group. In Figure 2, the first two column shows that the degree of involvement in general topic group and country-related topic group is different. Persian, Portuguese, and Chinese users actively edit general topics, and less actively edit country-related topics. However, all 6 language groups show low involvement in U.S related topics, and high involvement in countries other than the US. Appendix shows the distribution of 50 topics. The low involvement in country-related topic is due to the high-involvement in a few topics, such as T16 about music and T21 about weather. According to the language they edit, users show a skewness in certain topics, which is not related to countries but more detailed topics. We also measured the user interest in each country. Topics from LDA usually show several country names, which do not provide a detailed interest in one country. We counted the number of country names

q

	CI	CR	US	non-US
Bilingual	-0.004	0.009	-0.011	0.020
Russian	-0.014	0.014	-0.004	0.018
Italian	-0.001	0.012	-0.015	0.026
Persian	0.019	-0.018	-0.040	0.022
Portugese	0.034	0.000	-0.014	0.014
Chinese	0.012	-0.015	-0.024	0.009
Korean	-0.036	0.035	-0.018	0.053

(a)

	CI	CR	US	non-US
Non-Primary	-0.055	0.067	-0.002	0.068
Primary	-0.012	0.013	-0.006	0.019

(b)

Figure 2: The difference in proportion of each language groups. The proportion of each group is subtracted by proportion of total population. The degree of involvement in each topic group differ across language pairs.

in each article, and normalized by the total number. Table shows top 7 country names mentioned in articled edited by each language group. The result shows that the language of a user and their interest in countries are correlated. For example, Korean bilinguals are most interested in Korea, and Persian bilinguals show high interest in Iran. Though bilingual users of English and Italian do not frequently edit about Italy, we can observe that their interest is quite related with their geographical location.

Language Inclination

Although there are a few bilingual speakers who are equally fluent in two languages(Grosjean 1989), most people would be inclined to one language. The same fact applies to Wikipedia editors. We found that the number of articles a user edited in two languages seldom weighs the same. Figure 3 shows the histogram of normalized English article numbers by bilingual users. The distribution is roughly bimodal, which peaks at 0.05 and 0.98, indicating that most users exhibit inclination toward one language.

Based on this, we define the users in two categories: English primary user, who edits English more that 70% of their time; and English non-primary user, who edits less than 30%. We compare the average topic distribution and the topical diversity between them.

Topic analysis on this pair shows a similar behavior to the monolingual and bilingual pair. Figure 2(b) shows that the distribution of primary and non-primary user in four topic groups. Non-primary users are show more clear interest toward country-related topics, and exceptionally indifferent in the U.S and their culture. This indicates that users who are not familiar with English are also uninterested in a English-speaking country.

Topic analysis shows that those who infrequently use

Italian		**Korean**	
Country Name	*Proportion*	*Country Name*	*Proportion*
canada	0.048	korea	0.081
spain	0.04	japan	0.053
austria	0.037	kenya	0.039
ireland	0.037	australia	0.03
australia	0.035	china	0.029
denmark	0.03	afghanistan	0.02
china	0.026	canada	0.015
Persian		**Portugese**	
Country Name	*Proportion*	*Country Name*	*Proportion*
iran	0.042	brazil	0.084
egypt	0.019	canada	0.065
nigeria	0.007	japan	0.049
canada	0.003	australia	0.043
israel	0.003	china	0.036
russia	0.002	spain	0.031
switzerland	0.002	india	0.026
Russian		**China**	
Country Name	*Proportion*	*Country Name*	*Proportion*
russia	0.09	hong kong	0.112
jersey	0.057	china	0.088
china	0.05	mexico	0.061
india	0.05	japan	0.049
jordan	0.046	poland	0.045
canada	0.045	india	0.037
japan	0.045	australia	0.033

Table 2: The average mentions of country names based on user edits. Proportion means normalized number of appearance of one country name over all countries.

English tend to edit more local topics related to countries except for the US. Also, according to their language, Wikipedia users exhibit different topical interest. To observe the topical diversity of users, we examine the variance of topic vector of each article a user edited. Figure 4 shows the variance of topics with respect to the proportion of English articles. One dot represents one user. The overall trend is going up as the proportion of English gets higher. This indicates that the users who are not familiar with English focus on editing specific articles, and it is usually related with countries and local information.

Conclusion and Future works

Our preliminary findings suggest that bilingual editors of Wikipedia show high involvement in Wikipedia. We find repeatedly that the contents of bilingual users are related with local topics, and their language plays a big role in choosing their topics. Our results have important implications regarding the extent to which bilingual users transfer information about which topics. While there is no doubt that the user interest is diverse and sometimes very specific, we find that their language affects the majority of bilingual users. We further find that bilingual users show frequent and consis-

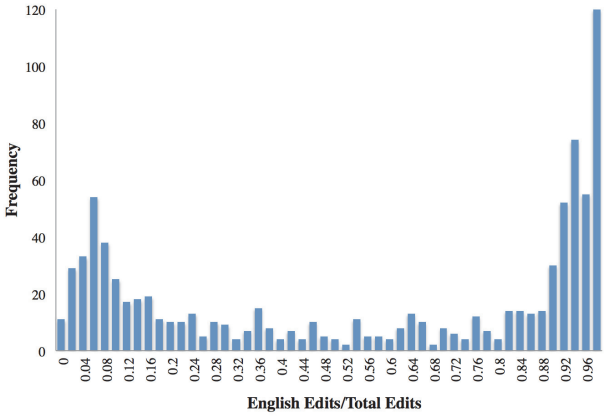

Figure 3: The number of users according to their relative contribution of English over total contributions. Bilingual users edit extreme large or small amount of English articles. (CI: Country-independent topics, CR: Country-related topics, US: topics related to the United States, non-US: topics related to countries other than the US.)

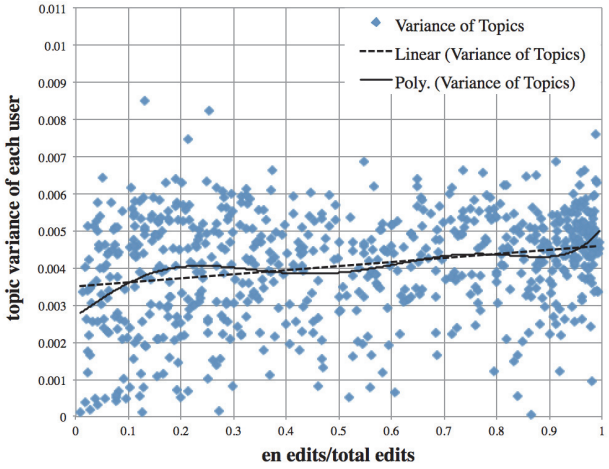

Figure 4: Variance of topic proportions across articles of a user. Plot represent the relationship between user involvement in English and the topic variance.

tent mentions about certain countries across Wikipedia editions. Also, we observe that users who are not familiar with English are more topically focused.

These findings suggest that many of the users editing two language editions of Wikipedia may have a high level of motivation of spreading information about themselves. This aligns with the claim that a good proportion of multilingual Wikipedia users may be power users who contribute to Wikipedia heavily(Kittur and Kraut 2008).

Meanwhile, the interest of users could be explored by analyzing the topics using other measures. Also, topic vectors of articles, averaged to represent the interest of users may be investigated independently. Not only the proposed ideas but also various research questions could be probed and we hope to work on the questions in the future works.

Acknowledgement

This work was supported by Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No. 10041313, UX-oriented Mobile SW Platform)

References

Eleta, I., and Golbeck, J. 2012. Bridging languages in social networks: How multilingual users of twitter connect language communities? *Proceedings of the American Society for Information Science and Technology* 49(1).

Etling, B.; Kelly, J.; Faris, R.; and Palfrey, J. 2009. Mapping the arabic blogosphere: Politics. *Culture, and Dissent/Berkman Center Research Publication* .

Grosjean, F. 1989. Neurolinguists, beware! the bilingual is not two monolinguals in one person. *Brain and language* 36(1):3–15.

Gumperz, J. J. 1982. *Language and social identity*, volume 2. Cambridge University Press.

Halavais, A. 2000. National borders on the world wide web. *New Media & Society* 2(1):7–28.

Hale, S. A. 2014. Multilinguals and wikipedia editing. In *Proceedings of the 2014 ACM conference on Web science*, 99–108. ACM.

Herring, S. C. 2007. A faceted classification scheme for computer-mediated discourse. *Language@ internet* 4(1):1–37.

Kim, S.; Weber, I.; Wei, L.; and Oh, A. 2014. Sociolinguistic analysis of twitter in multilingual societies. In *Proceedings of the 25th ACM conference on Hypertext and social media*, 243–248. ACM.

Kim, S.; Park, S.; Hale, S. A.; Kim, S.; Byun, J.; and Oh, A. 2015. Understanding editing behaviors in multilingual wikipedia. *arXiv preprint arXiv:1508.07266*.

Kittur, A., and Kraut, R. E. 2008. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, 37–46. ACM.

Ochs, E. 1993. Constructing social identity: A language socialization perspective. *Research on language and Social Interaction* 26(3):287–306.

Otterbacher, J. Chapter 3 our news, their events: A comparison of archived current events on english and greek wikipedias.

Phinney, J. S.; Romero, I.; Nava, M.; and Huang, D. 2001. The role of language, parents, and peers in ethnic identity among adolescents in immigrant families. *Journal of youth and Adolescence* 30(2):135–153.

Appendix

Bilingual	Russian	Italian	Persian	Portuguese	Chinese	Korean	Topics from LDA
0.0005	-0.0030	-0.0003	0.0003	-0.0030	0.0026	0.0143	0: ship + fire + aircraft + navi + korean + korea + japanes + crew + two + fleet + tank + naval
-0.0017	-0.0037	-0.0005	-0.0098	-0.0053	-0.0050	-0.0091	**1: england + british + london + ireland + english + scotland + irish + uk + 's + scottish + wale**
-0.0057	0.0036	-0.0025	-0.0128	-0.0036	-0.0014	-0.0153	2: citi + area + popul + town + locat + park + includ + 's + river + district + hous + also + north
0.0029	0.0071	0.0168	-0.0116	0.0061	0.0111	0.0064	3: number + speci + one + two + use + x + function + group + n + point + also + form + gener
0.0000	0.0064	0.0004	0.0100	0.0019	-0.0032	-0.0016	4: kill + attack + battl + death + power + war + forc + one + fight + destroy + escap + attempt + die
-0.0095	-0.0410	-0.0233	-0.0481	0.0109	-0.0268	0.0027	5: 's + show + seri + episod + season + appear + televis + also + star + film + charact + play + first
0.0037	0.0171	0.0074	0.0319	0.0028	-0.0017	-0.0067	6: empir + greek + 's + ottoman + centuri + rule + europ + egypt + war + muslim + turkish + greec
-0.0003	-0.0031	0.0080	-0.0029	-0.0019	-0.0046	-0.0059	7: church + christian + cathol + jesu + jewish + jew + religi + saint + pope + holi + christ + bishop + st.
0.0035	0.0069	0.0019	0.0728	-0.0049	-0.0026	0.0155	8: one + use + exampl + term + theori + differ + 's + person + social + view + mean + way + state
-0.0019	-0.0026	-0.0123	-0.0094	-0.0054	0.0115	-0.0138	9: compani + oper + servic + line + new + station + 's + train + million + open + plan + rout + build
-0.0006	-0.0081	0.0158	0.0082	-0.0018	-0.0044	0.0086	10: 's + work + book + publish + first + year + time + wrote + new + life + novel + later + write
-0.0008	-0.0007	-0.0041	-0.0108	0.0010	-0.0057	-0.0007	**11: american + white + black + south + new + african + coloni + state + unit + america + nativ**
0.0001	0.0025	-0.0050	-0.0078	-0.0001	0.0006	-0.0011	12: may + women + diseas + patient + use + treatment + medic + 's + caus + includ + also + sexual + infect
-0.0009	-0.0019	0.0037	-0.0035	-0.0054	0.0036	-0.0040	13: race + edit + block + car + pleas + one + vandal + wikipedia + made + revert + driver + appear
0.0031	-0.0028	-0.0015	-0.0043	-0.0007	-0.0026	0.0126	14: san + mexico + mexican + el + america + diego + santa + puerto + argentina + whale + spanish
-0.0034	-0.0168	-0.0172	0.0933	-0.0145	-0.0176	-0.0134	15: film + 's + award + movi + releas + best + million + role + star + director + actor + direct + critic
0.0050	-0.0046	0.0585	-0.0455	0.0978	-0.0159	0.0008	16: album + song + releas + 's + record + singl + music + chart + video + number + first + also + tour
0.0017	-0.0037	-0.0001	0.0267	-0.0074	-0.0051	0.0108	17: pakistan + israel + arab + isra + muslim + iran + iraq + islam + khan + palestinian + governaqi
0.0011	-0.0015	-0.0012	0.0170	-0.0041	0.0201	0.0005	18: effect + use + studi + system + human + may + result + research + increas + energi + measur + time
-0.0028	0.0022	-0.0094	-0.0025	-0.0089	-0.0027	-0.0004	19: match + titl + championship + world + defeat + wrestl + fight + first + 's + round + team + champion
-0.0012	0.0023	0.0178	-0.0010	0.0022	-0.0034	-0.0156	20: music + band + record + 's + play + perform + rock + song + guitar + includ + sound + instrument
0.0017	-0.0057	-0.0052	-0.0089	-0.0077	0.0581	-0.0055	21: island + water + storm + sea + wind + tropic + day + ocean + area + hurrican + year + temperatur
-0.0042	-0.0168	0.0000	-0.0059	-0.0210	0.0124	0.0040	22: school + univers + student + colleg + educ + program + institut + scienc + year + research + high + studi
0.0020	-0.0059	-0.0076	-0.0206	-0.0058	0.0038	-0.0105	23: use + engin + design + model + power + system + also + vehicl + car + 's + speed + first + electr + air
0.0047	-0.0020	0.0086	-0.0045	0.0318	-0.0027	-0.0005	24: de + french + la + franc + spanish + spain + pari + portugues + brazil + le + del + brazilian + barcelona
-0.0089	0.0005	-0.0103	-0.0198	-0.0098	-0.0133	-0.0079	**25: state + counti + york + new + 's + citi + american + age + year + unit + texa + mile + florida**
0.0015	0.0015	-0.0082	-0.0126	-0.0098	0.0311	-0.0059	26: india + chines + china + indian + also + asia + peopl + singapor + kong + hong + region + asian + taiwan
0.0002	0.0067	0.0019	0.0165	-0.0061	-0.0104	0.0087	27: languag + use + word + name + tradit + god + also + one + centuri + refer + form + mean + english
0.0004	-0.0040	-0.0010	-0.0052	0.0070	0.0048	0.0025	28: king + 's + son + princ + queen + royal + famili + duke + father + daughter + name + marri + ii + henri
-0.0054	0.0003	-0.0098	0.0093	-0.0029	-0.0107	0.0160	29: 's + state + court + report + law + said + polic + presid + offic + case + public + would + investig + right
-0.0001	0.0041	-0.0008	-0.0055	0.0021	0.0012	0.0089	30: charact + weapon + 's + use + seri + dragon + anim + robot + gun + appear + power + also + origin
-0.0054	-0.0199	-0.0076	-0.0246	-0.0076	-0.0070	0.0024	31: 's + get + one + time + make + would + take + n't + go + back + say + find + tri + want + tell + later
0.0003	0.0061	-0.0019	0.0029	-0.0058	0.0040	0.0058	32: forc + war + armi + militari + command + german + unit + air + oper + troop + divis + gener
0.0010	-0.0019	0.0020	0.0180	-0.0018	-0.0010	-0.0022	33: islam + ali + africa + muhammad + ibn + muslim + african + gaddafi + abu + arab + sudan + libya + kenya
-0.0036	-0.0077	0.0050	0.0330	-0.0304	-0.0320	-0.0238	34: utc + talk + articl + page + wikipedia + delet + edit + n't + pleas + 's + use + may + thank + ad
0.0033	0.0295	-0.0043	-0.0038	-0.0076	-0.0053	-0.0118	35: state + govern + 's + parti + nation + elect + polit + countri + presid + member + support + union + war
0.0019	-0.0020	0.0200	-0.0017	0.0094	0.0010	-0.0012	36: centuri + roman + italian + period + itali + rome + stone + citi + 's + wall + date + king + bc + one
0.0042	0.0022	-0.0049	-0.0041	0.0002	0.0128	0.0121	37: use + code + standard + type + system + comput + one + exampl + number + languag + program + differ
0.0001	0.0039	-0.0061	-0.0118	-0.0092	-0.0003	0.0038	38: market + countri + compani + bank + product + rate + use + state + increas + million + unit + tax + industri
-0.0005	0.0016	-0.0036	-0.0073	-0.0066	0.0071	-0.0057	39: use + water + acid + ga + form + cell + temperatur + chemic + heat + metal + energi + materi + carbon
0.0016	0.0023	0.0093	-0.0071	0.0044	-0.0064	0.0062	40: station + radio + broadcast + channel + network + 's + program + news + televis + v + air
0.0002	-0.0033	-0.0039	0.0056	0.0032	-0.0001	0.0016	41: contest + show + jone + judg + elimin + week + challeng + van + vote + episod + announc + winner
-0.0020	0.0130	0.0018	0.0015	-0.0044	-0.0011	-0.0031	42: food + use + plant + also + red + anim + product + speci + color + produc + dog + often + made + tree
0.0092	0.0203	-0.0031	0.0422	0.0137	0.0239	0.0061	43: use + system + user + softwar + data + support + develop + network + includ + provid + comput + file
-0.0006	0.0001	-0.0014	0.0059	0.0000	0.0005	-0.0036	44: earth + 's + test + space + star + moon + planet + sun + orbit + cricket + first + solar + mission + mar
0.0030	0.0094	0.0011	-0.0458	-0.0065	-0.0124	-0.0455	45: season + team + play + leagu + club + first + game + 's + player + win + footbal + score + cup + goal
-0.0022	-0.0067	-0.0038	0.0034	-0.0029	0.0002	-0.0045	46: 's + comic + australia + australian + batman + new + superman + sydney + issu + green + stori + clark
0.0025	0.0042	-0.0074	-0.0057	0.0260	0.0028	0.0317	47: game + player + 's + releas + video + also + develop + new + version + featur + seri + includ + play + one
0.0022	-0.0039	-0.0040	-0.0203	-0.0004	-0.0040	0.0434	48: world + 's + canada + sport + team + first + event + won + competit + nation + canadian + olymp + intern
0.0001	0.0215	-0.0054	-0.0086	-0.0045	-0.0039	-0.0056	49: airport + airlin + russian + flight + intern + india + russia + air + aircraft + ukrainian + ukrain + tamil

Top words in 50 topics(the last column), and the average topic proportions of each language group. Topic words in yellow means topics are related with countries except for the US, and the words in orange means topics are related with the US. The topic proportions are represented as a relative values to the total topic distribution (average proportion per group-total proportion). It is color-coded, and red is for high-values, yet blue is for low, negative numbers.