Cultural Identities in Wikipedias

Marc Miquel-Ribé
Universitat Pompeu Fabra
Roc Boronat, 138, 08018, Barcelona,
Catalonia, Spain
marcmiquel@gmail.com

David Laniado
Eurecat
Av. Diagonal, 177, 080018, Barcelona,
Catalonia, Spain
david.laniado@eurecat.org

ABSTRACT

In this paper we study identity-based motivation in Wikipedia as a drive for editors to act congruently with their cultural identity values by contributing with content related to them. To assess its influence, we developed a computational method to identify articles related to the cultural identities associated to a language and applied it to 40 Wikipedia language editions. The results show that about a quarter of each Wikipedia language edition is dedicated to represent the corresponding cultural identities. The topical coverage of these articles reflects that geography, biographies, and culture are the most common themes, although each language shows its idiosyncrasy and other topics are also present. The majority of these articles remain exclusive to each language, which is consistent with the idea that a Cultural Identity is defined in relation to others; as entangled and separated. An analysis of how this content is shared among language editions reveals special links between cultures. The approach and findings presented in this study can help to foster participation and inter-cultural enrichment of Wikipedias. The datasets produced in this study are made available for further research.

CCS Concepts

• **Human-centered computing→Wikis** • **Human-centered computing→Empirical studies in collaborative and social computing** • *Human-centered computing→Collaborative content creation* • *Information systems→Data analytics* • *Social and professional topics→Cultural characteristics* • Applied computing~Psychology

Keywords

Cultural Identity, Wikipedia, Cross-cultural studies, Online Communities, Analytics & Data Mining

SMSociety '16, July 11 - 13, 2016, London, United Kingdom

ACM 978-1-4503-3938-4/16/07…$15.00
DOI: http://dx.doi.org/10.1145/2930971.2930996

1. INTRODUCTION

Wikipedia is self-defined as "a free-access, free-content Internet encyclopedia"[1]. When Jimmy Wales and Larry Sanger started Wikipedia in 2001, they were already developing a free encyclopedia called Nupedia with this same purpose. It was the implementation of the wiki technology that completely changed their approach by allowing collaborative modifications directly from the browser. This grew into the current site we know. The result is a dual object: a social network that also serves the purpose of creating a knowledge repository. However, , Wikipedia does not encourage editors to build their identities based on personal traits, biography and social affinities[2], which is different from other online communities. Instead, Wikipedians are valued according to their activity, their writing skills, the languages they speak or acknowledgements they have received from other peers, such as barnstars[3] and praising comments.

To be a Wikipedian requires being involved in the community, learning how to edit articles, and a motivation that sustains their involvement. According to previous research, Wikipedians are motivated by reasons like the project ideology, the fun of writing, community values, and various other motivations [19, 28]. In this study we start from the observation that Wikipedians may also be motivated by their identities, and that apart from userpages, such identities may emerge in users' content choices. In this sense, there is evidence that a great part of content does not strictly follow a balanced coverage of the encyclopedic topics, and coverage analyses reflect an overrepresentation of culture and arts, and biographies; in particular celebrities, pop artists and media [16]. Likewise, interests in what content to create varies according to geography [13]. Therefore, we believe a greater understanding of the interplay between Wikipedia editors' identities and motivation could explain the cultural contextualization and composition of different Wikipedia language editions as well as provide new insights on editors' behavior.

In this study we want to explore how an identity-based motivation drives the editors of each Wikipedia language edition to contribute content related to their cultural identities. To this aim, we propose a computational approach to obtain articles related to editors' cultural identities, and run it on 40 Wikipedia language editions selected to validate the results across a diverse set of content. We measure the resulting proportion of articles as an indicator of the influence of editors' motivation related to cultural identities. Then, in order to enrich the understanding of

[1] https://en.wikipedia.org/wiki/Wikipedia

[2] https://en.wikipedia.org/wiki/Wikipedia:NOT

[3] https://en.wikipedia.org/wiki/Wikipedia:STAR

Cultural Identity related content, we analyze the topics they comprise and their availability across languages. Lastly, we provide recommendations regarding the gained insights to enhance growth and diversity in the overall Wikipedia project. The datasets produced in the study are made available to encourage further research [4].

2. BACKGROUND

2.1 Identity-based motivation in Wikipedia

The fact that Wikipedia allows anyone to edit articles without being registered has been one of the slogans to invite new editors. However, after a few hours of contributing, editors often realize Wikipedia is a community with goals, rules, and different tasks to perform. Then, new editors often establish a userpage so that other peers can recognize a user's skills and degree of expertise in the topics in which they can collaborate, and even track their contributions [2]. The importance of creating and developing an identity within the community is not a matter of narcissism, but of gaining credibility and trust. The value of a Wikipedian lies in their previous edits and areas of interest, as well as the competences demonstrated in different kinds of tasks. Unlike most online communities, disclosing personal identity aspects like hobbies, professional experience or social affiliations is not required or encouraged on Wikipedia. Nevertheless, aspects such as gender, religion or education can be inferred from content [23], suggesting that other identities, besides the Wikipedian one, can play a role on the site. In this sense, Oyserman's model of identity-based motivation [20] can provide background to explore and reflect on Wikipedia as a context where editors' identities matter.

Firstly, the main tenet of the model is that "people are motivated to act in identity-congruent ways" [21]. Therefore, a Wikipedian conciliates their activity goals in the encyclopedia with those derived from other identities. In fact, Oyserman follows that "identity is a dynamic function of the pragmatic options for action a particular situation" […] "and these options are imbued with identity-based meaning" [21]. Then, interactions in Wikipedia could be motivated by being part of the Wikipedia community, the encyclopedia characteristics and its place in society, but also by the meaning from the particular content they interact with. In other words, the possibility of contributing with certain content aligned to personal beliefs, values and interests allows editors to fulfill several aims associated to each identity. And since "identities can be subtly cued without conscious awareness" [20], an editor might choose to perform certain tasks oriented by a Wikipedian identity (e.g., correcting typography errors, or introducing specific data) and complement it by contributing content related to others identities.

Secondly, an identity-based motivation "may not necessarily be serving individuals' goal attainment." This also remains true in the scenario of Wikipedia, where the collective effort of constructing an encyclopedia revolves around the idea of "gathering the sum of all human knowledge."[5] We may consider that the vagueness of this goal can have considerable content implications at different levels; acting as an open call for a wide range of content, which may align with all kinds of identities, whether they are political, religious or related to other characteristics. For instance, if an identity involves the goal of expansion and proselytism, contributions may result in content that is not in line with the objectives of the encyclopedia. In order to prevent undesired content, Wikipedia has suitable norms and guidelines. At an article creation level, a 'Notability guideline'[6] avoids new unnecessary or inappropriate articles by requiring a specified minimum of verifiable sources. For content inside an article, the policy of 'Neutral Point of View' requires that any text must "represent fairly all the significant views published by reliable sources on a topic."[7] Even though these norms establish some limitations in order to correct the content, their appliance always depends on other editors' intervention, and in case of dispute solutions, are taken on a consensus basis. Therefore, editors' identities may also play a role in discriminating against new articles and points of view. Then, the overrepresentation of certain topics [16], the imbalance of articles in different language editions [27], or the different points of view on the same topic depending on the language edition [9, 17, 24], may be explained by shared identities. Consequently, the more common an identity is within the editing community, the easier it is for content related to it to remain in the encyclopedia, because editors may not be willing to delete such content taking an action incongruent to their identity.

Thirdly, an identity-based motivation sets people into "readiness to act and make sense of the world in terms of norms, values and behaviors relevant to the identity." As an online encyclopedia, Wikipedia requires immediacy in order to respond to the information needs in our society. One of the cases in which it is accessed most is when readers need to understand specific concepts to follow breaking news [14]. In fact, the aim of covering any kind of topic has positioned the site among the first results in search engines[8], which in turn triggered the attraction of new editors who helped to create more content in a type of feedback loop [25]. Hence, when Wikipedians contribute to the encyclopedia they may feel in the crossroad of fulfilling the expected readers' informational needs, in addition to the content they feel most congruent with.

Because of the dynamics of identity and characteristics of Wikipedia, Oyserman's model suggests us that any identity can influence both content creation and editor interaction. As an example, Neff et al. [18] studied the impact of community identification on political interaction in Wikipedia and observed that editors who self-presented with political affiliations in their user-page had also intensely identified as Wikipedians. Further, results also showed that editors who disclosed their political affinities tend to edit more content related to the political party they support, which suggests that the conciliation between political identity and being a Wikipedian is not only possible, but affects and permeates all the aspects of interaction. In this sense, when we apply the model to Wikipedia, its main advantage is that it sheds light on both the cultural and social nature of identity; providing a deeper understanding of identity-based processes and their outcomes in the encyclopedia.

[4] Available at: http://www.wikiidentities.org

[5] https://slashdot.org/story/04/07/28/1351230/wikipedia-founder-jimmy-wales-responds

[6] https://en.wikipedia.org/wiki/Wikipedia:Notability

[7] http://en.wikipedia.org/wiki/Wikipedia:NPOV

[8] https://econsultancy.com/blog/9009-why-wikipedia-is-top-on-google-the-seo-truth-no-one-wants-to-hear

2.2 Cultural Identity and Wikipedia

Cultural Identity refers to the sense of belonging to a group and is defined "in terms of cultural or subcultural categories (including ethnicity, nationality, language, religion, and gender among others)"[9]. Therefore, Cultural Identity is a broad and useful concept to analyze content created in Wikipedia as a result of an identity-based motivation. Mainly, to understand Cultural Identity it is necessary to delve into how it is constituted and created in an historical context.

Cultural theorist Stuart Hall [6] defines Cultural Identity as "the common historical experiences and the shared cultural codes." He adds that "culture is about shared meanings," and it can be the language, territory places, artistic creations, traditions, among others. He stresses the importance of the idea that meanings are originated around a place. This is a very prevalent idea in social sciences. Anthropologists Hofstede and Hofstede [12] affirmed that "culture is a collective phenomenon because it is shared with people who live or lived within the same social environment."

According to Hall, one of the most important aspects from Cultural Identity is its dynamic nature. It is a matter of becoming as well as of being. Its creation is not fixed, and it is in constant relationship with history, culture and power in territories. Likewise, individuals' cultural identities can undergo changes because of their integration into different places, mixing with communities, where different cultures are practiced. People's cultural identities are the sum of experiences that occur in precise places with other people. Therefore, Hall affirms that cultural identities are represented, and that happens when their "shared meanings or shared conceptual maps" use language system as a vehicle [7]. In fact, they can coexist in language: for instance, British and North American cultural identities may share meanings despite being in different territories. Some languages may also coexist in the same territory, giving place to different cultural identities with shared meanings about their surrounding environments. This makes the creation and representation of a cultural identity a variable geometry. Only in some cases in which territory sovereignty coincides with the territory of cultural practice, cultural identity shared meanings are coincident with those from a national identity (one case of this would be Icelandic cultural identity). This reaffirms the idea that cultural identities are tied to territory in their origin, their constant dynamic evolution and their representation.

In Wikipedia, the editors' geographical factor has been used to explain how diversity appears in each language edition. More generally, the diversity process has been referred to as a Cultural Contextualization and it happens in any user-generated content repository[10]. Wikipedia editors tend to contribute with information related to near by locations [11]. One of the consequences is the categorization of Wikipedia language editions as 'self-focus biased', which means that editors' attention is highly biased towards their own territories. Hecht and Gergle [8] detected this phenomenon by analyzing the prominence of the articles associated to the territories local to each language edition (analyzed in number of hyperlinks and PageRank scores coming from all the Wikipedia language edition articles). Other consequences of cultural contextualization are that each Wikipedia language edition has a very different set of unique content, and that instead, when content is shared to other languages this responds to geographical proximity factors [27].

In this study we propose Wikipedia's cultural contextualization process can be more fully explained by an identity-based motivation type that moves editors to act congruently with their cultural identities and represent them. In the following section we propose our research questions.

2.3 Research Questions

We assume that in the process of understanding their territory, editors will contribute to Wikipedia with those shared meanings they have learnt from their Cultural Identity, which include geographical places and also involve education, traditions, among many other subjects. The relative amount of such articles in each Wikipedia language edition will reflect the influence of the motivation. Even though previous research [5, 15] found patterns of multilingual editing activity in each language edition, they implied lower levels of activity and linguistic quality. We expect to find a considerable portion of each Wikipedia dedicated to cultural identities. Therefore, we ask:

> **RQ1-Extension:** What is the extent of editors' Cultural Identities representations in each Wikipedia language edition? (Section 4.1)

Since elements from cultural identities are shared between the editors of a language edition, we want to know what topics are required to understand their immediate context and make sense of the world. Looking at the topical coverage may allow us to inspect which shared meanings are more essential for the cultural identities in each language edition. We expect each language based cultural identities to require diverse topics to represent their context according to their location and their historical background. Therefore, we ask:

> **RQ2-Topics:** What is the topical coverage of editors' Cultural Identities representations in each Wikipedia language edition? (Section 4.2)

Cultural Identities are also framed in terms of difference and otherness. There exists a relativism between identities, implying that in cultures there is sometimes a certain lack of equivalence, and in order to translate meaning, it is necessary to move from one mindset to another [6]. In Wikipedia, different language editions show a considerable amount of unique content [27], which is partially explained by the fact that some languages split large topics into more than one article [9]. We expect content related to cultural identities to be mainly exclusive and part of this unique content found in every language edition. Therefore, we ask:

> **RQ3-Cross-language:** What is the availability of content representing editors' Cultural Identities across different Wikipedia language editions? (Section 4.3)

3. DATASET CONSTRUCTION: CULTURAL IDENTITY RELATED ARTICLES (CIRA)

In this section we describe our method for identifying a comprehensive set of Cultural Identity Related Articles (from

[9] http://www.oxfordreference.com/view/10.1093/oi/authority.20110803095652855

now on CIRA) in each Wikipedia language edition, and we assess the results obtained for 40 language editions.

The selection of language editions includes the biggest 30 in number of articles (as of July 2015), in addition to 10 more language editions to complete the picture with distinct sociolinguistic factors to include the five continents, different linguistic roots, different speaking community sizes, and also different editing community sizes. The 10 added language editions are Basque, Estonian, Greek, Macedonian, Hebrew, Swahili, Afrikaans, Icelandic, Nepali and Guarani.

3.1 Mapping Cultural Identities to Wikipedia language editions

In order to map Cultural Identities to each Wikipedia language edition content it was required to set a database with the territories where a language is spoken. Therefore, we chose ISO code 639 used by Wikimedia Foundation to classify Wikipedia language editions (e.g., 'es' for the Spanish language Wikipedia: es.wikipedia.org) and ISO codes 3166 and 3166-2 to identify each country and its subdivisions at regional level. These codes are widely used on the Internet in geolocation services.

This way we paired each of the selected language editions with its native words to specify the territories where it is officially spoken ('de iure' or 'de facto'), its inhabitants' demonym and language name (e.g., eswiki españa mexico … español castellano). This word list has been generated crossing ISO databases, and for cases such as a language spoken in a region that does not appear in the database, or a second name for a language, it has been manually revised and extended using information from the specific articles in the correspondent Wikipedia language editions.

3.2 Article Selection and Filtering

Each language's CIRA is expected to be a set of articles encompassing a wide variety of topics to represent the shared meanings related to the corresponding territories and cultures. For the purpose of gathering such articles for each language, we developed and in July 2015, applied several strategies.

First we gathered the articles considered to be more reliably identifiable: articles (i) including in their title keywords related to the language or the corresponding territories as defined in the previous section (e.g., "England National football team", "English law", etc.) or (ii) clearly located within such territories. Articles satisfying the first criterion were directly retrieved from the databases of each Wikipedia language edition, which are updated in real time and whose access was provided by Wikimedia Foundation[10]. The second criterion required examining article location tags such as the coordinates and the ISO code, and performing some validation. We noticed that coordinate implementation is unequal in different language editions and may contain errors. Therefore, articles with only a pair of coordinates were verified using a *reverse geocoder* util in Python, which provided a ISO code to check in our database. Later, we added the articles that were not tagged with coordinates neither with territory ISO code, but could be matched to the corresponding articles in other language editions, in which they were properly geolocated (e.g., an article about a city in Nepal which was not geolocated in the Nepali Wikipedia, but it was in the English Wikipedia).

[10] http://wikitech.wikimedia.org

These two criteria allow us to reliably include articles in a CIRA selection, but still leave out many other articles that should be included. The third criterion is a strategy to retrieve articles linked to particular keywords [22]. Wikipedia articles are classified according to categories, which are named according to the topics developed in the articles. Since these categories are organized in a hierarchical tree, starting from a few categories at a general level allows crawling down the classification structure and gathering all the articles about a particular topic. Similarly to article retrieval according to the first criterion, we used the keywords identified through the ISO codes, and retrieved all the categories including them in the title; for example: "Performing Arts in England" or "Disputes in English Grammar." These categories contain articles, and other categories containing in turn more specific articles (see Figure 1), until at a certain level the process of crawling and gathering articles finishes. This will depend on the way each editing community constructed the category structure, but it generally happens around the tenth level. The main advantage of this method is that it allows articles related to some top-level keywords to be obtained. However, the distance to the top matters: while category "Films directed by Charlie Chaplin," is part of "Performing Arts in England" category, its content will be far more specific. The downside of the method is that sometimes the categorization includes circular references or incorrect links (e.g., a more general category appears under a more specific one), which may produce interferences in the final gathering (e.g., "World War II" category placed under "Wars involving the United States" category would bring to include articles about the German army as related to the English Wikipedia related cultural identities). Possibly because of this interference issue, when [22] used this method in 2011 with the keywords territories, demonyms and language names, they only took into account the first four levels. Their results were the average proportion of 24.9% articles per language for a total of 20 language editions. In our case, we only put a limit of five levels of iteration to the English language edition, letting the rest of languages complete the iterations until the down category graph goes extinct.

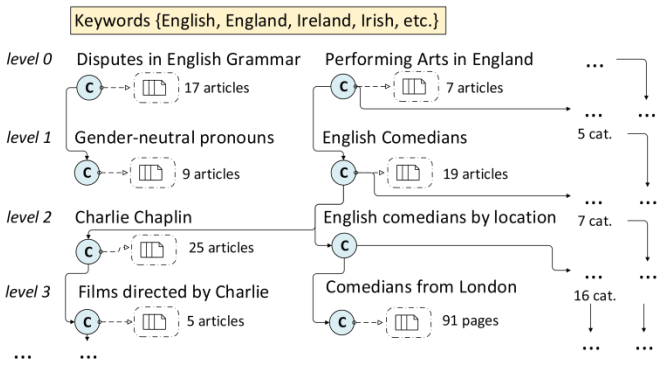

Figure 1: Crawling down the category graph with keywords.

Since most of the articles obtained using this third criterion and method can be considered CIRA, we tackled the interference issue with a filter. To be effective it had to discriminate whether and article relates to the editors' cultural identities in their text or the links contained in them that direct the reader to other Wikipedia articles. Fortunately, the geolocated articles and those including the keywords in their title could serve as an initial ground-truth. As such, when articles from the bulk category crawling selection had a 15% of their text links pointing out to ground-truth they could be added to this group for a further

iteration. While the algorithm usually did not add more articles after the third iteration, in large Wikipedia language editions like the English language we had to limit it to the fifth iteration because more articles considered for the new ground-truth had an attracting effect with interference from the bulk. Using this 15% threshold we obtained a definitive CIRA slightly smaller than the bulk selection, but avoided most of the interference.

Table 1 reports the total number of articles and the percentage of articles classified as CIRA at the end of the process for each of the 40 considered language editions. Furthermore, the table shows the percentage of articles that were identified through Criterion 1 (i.e., through keywords in the title) and Criterion 2 (geolocated articles), and the percentage of articles identified through the category titles, before applying the iterative filters. We omit the percentage of articles selected with this third criterion after applying the filter; as for most language editions it is very close, or almost equal to the final percentage of articles included in the CIRA set.

In the Venn diagram shown in Figure 2 we can see the average proportion of CIRA in the 40 language editions, and the proportion of these articles that were identified via geolocation tags and keywords in the title. As it can be observed, about 1 over 5 articles in the CIRA set was identified via geo-coordinates, while only about one over 20 was identified via keywords in the title. The intersection between the two criteria is rather small. Data for the articles identified via the category hierarchy are not shown, as they represent almost the totality of CIRA (29.5% on average).

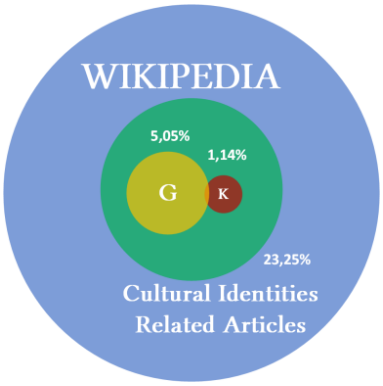

Figure 2: Average proportion of CIRA, and of CIRA detected through geolocation and keywords.

3.3 Manual assessment

To check the precision of the method and filter against interference we retrieved for each language edition 100 random articles classified as CIRA, and 100 random articles from the remaining ones for manual assessment. We used an automatic translator to translate the text of each article, and we manually classified them according to their content as belonging to CIRA or not. False positives were for instance articles totally unrelated about specific topics from nearby countries, or due to anecdotal relationships such as a football player who played a competition in one of the countries associated to a language. In few other cases, articles were considered to be part of CIRA despite not being exclusively focused on a country speaking the corresponding language, if they were relevant to a country's history or society, and this was reflected by the article content. For example, the article about the disputed French region of Lorraine was important to explain the history of Germany, especially during the first decades of 20th century, when it used to be part of the German Empire, and in consequence it is categorized in German Wikipedia as "Historical Territory (Germany)." In other Wikipedia language editions neither its text nor categories provide significant references about this historical period. Then, instead of debating between original or imported concepts, the CIRA selection should be seen as a continuum from those more central to a culture - in Hall's words, "shared historical codes" - to those more peripheral but still maintaining an important semantic value to explain a society's imaginary. Deleting periphery is possible by reducing the 15% threshold or adjusting the number of iterations lower than 5. Additionally, it would be interesting to try deleting interference if they belong to CIRA from other languages. Interference is a limitation we may address in future versions of CIRA to improve accuracy. The results of our manual assessment are shown in Table 1, which reports, for each language edition, the percentage of false positives (FP) and false negatives (FN), together with the corresponding F1 score.

Overall, we found that across the 40 languages there were on average 3.3% of false positives, and 3.4% of false negative. The average value of F1 is 0.48. The selections with more interference are Korean and Chinese (12% and 10% FP respectively). This is mainly due to the fact that the category hierarchy of these Wikipedias does not strictly follow a general-to-specific principle, and many articles are short and under developed and contain very few links, which makes the 15% threshold ineffective in filtering out anecdotal links. Some improvements might be achieved by setting a different value of the threshold for different languages, but on the other hand we believe that always using the same value for the parameter makes the results more coherent and comparable across languages, with acceptable accuracy levels.

4. RESULTS

4.1 Extent of the representation of Cultural Identities

As it can be observed in Table 1 and Figure 2, almost a quarter of each Wikipedia language edition (mean 23.2%, median 24.2, standard deviation 11.1%) belongs to Cultural Identity Related Articles **(RQ1-Extension)**. These results confirm the existence of an identity-based motivation, which emerges when editors edit and represent their cultural identities in the content of their Wikipedia language edition. The growth of Wikipedia language editions depends on factors such as number of speakers, language status, Internet access for the average speaker and their attitude towards their language [26]. Therefore, it is difficult to compare the proportion of CIRA across languages. The English Wikipedia is the biggest in number of articles, and its CIRA set is among the largest in proportion, with the 46.8% of the articles in the encyclopedia. Only the Japanese Wikipedia has a larger proportion of CIRA (49.2%). For all the other languages, the proportion of CIRA is below 40%. Low proportions of CIRA observed for some languages are due to the presence of automatically translated content. For example, the Vietnamese, Cebuano and Waray-Waray Wikipedia language editions are among the top ten in number of articles and only have strikingly

Table 1: For each of the 40 Wikipedia editions, columns show: total number of articles (WP art), percentage of CIRA (CIRA %), percentage of articles identified through geolocated tags in the corresponding territories (Geo %), percentage of articles identified through keywords in their titles (KW %), total percentage of articles identified through the category hierarchy, before iterative filtering (CC %), percentage of false positives (FP %), percentage of false negatives (FN %), resulting f1-score (F1), percentage of Featured articles among CIRA (FA %), average of Interlanguage links per article (ILL WP), average Interlanguage links in CIRA (ILL CIRA), percentage of CIRA having no ILLs (No ILL %)

ISO code	Language	WP Art.	CIRA %	Geo %	KW %	CC %	FP %	FN %	F1	CIRA FA %	ILL WP	ILL CIRA	No ILL %
af	**Afrikaans**	35966	19.20	5.95	0.91	19.53	1	1	0.50	13.75	40.12	4.45	56.16
ar	**Arabic**	375282	26.92	3.21	2.44	35.88	3	12	0.46	42.89	12.89	3.55	56.33
eu	**Basque**	208630	10.05	1.65	0.42	16.25	2	0	0.49	36.30	21.52	3.63	56.50
ca	**Catalan**	467486	16.17	7.93	0.83	18.58	0	0	0.50	17.91	14.98	1.56	56.53
ceb	**Cebuano**	1211531	0.03	0.00	0.06	0.06	2	0	0.49	0.00	4.81	8.85	56.75
zh	**Chinese**	851670	32.87	6.25	1.17	67.92	10	6	0.46	12.43	10.00	2.58	59.43
cz	**Czech**	326187	25.97	9.04	1.15	29.31	5	2	0.48	20.13	15.01	2.48	56.38
da	**Danish**	205764	31.70	6.11	1.00	39.56	6	5	0.47	30.77	17.93	4.15	59.24
nl	**Dutch**	1828148	7.77	1.64	0.33	9.29	1	2	0.49	19.53	6.81	1.46	56.69
en	**English**	4917741	46.84	9.84	2.75	58.62	4	12	0.46	75.07	3.46	2.36	59.87
et	**Estonian**	136362	31.06	6.06	1.73	33.51	2	5	0.48	50.00	20.16	1.83	58.30
fi	**Finnish**	375347	21.95	2.31	1.03	23.69	1	3	0.49	18.34	14.40	1.28	56.44
fr	**French**	1642276	29.00	6.88	1.70	31.25	9	5	0.46	32.83	7.64	4.83	57.26
de	**German**	1834147	36.77	8.76	1.85	37.89	9	6	0.46	45.53	6.04	2.92	59.77
el	**Greek**	108090	33.55	6.44	0.60	35.97	3	3	0.49	33.84	23.20	4.74	59.15
gn	**Guarani**	3031	23.59	13.96	3.27	24.05	0	5	0.49	-	82.07	24.18	56.11
he	**Hebrew**	174667	31.72	2.06	1.61	34.53	4	4	0.48	40.87	20.02	4.79	59.32
hu	**Hungarian**	326146	18.50	1.91	1.45	21.67	2	1	0.49	16.24	16.04	2.92	56.39
is	**Icelandic**	39554	30.70	2.19	1.49	32.18	1	2	0.49	20.00	33.74	2.39	58.32
id	**Indonesian**	363529	27.02	1.01	0.58	32.76	3	2	0.49	-	11.97	1.66	56.24
it	**Italian**	1210801	19.24	3.62	0.65	20.50	1	2	0.49	36.76	9.31	3.48	56.18
ja	**Japanese**	973955	49.24	3.42	1.01	56.36	0	9	0.48	38.82	7.05	1.15	76.57
ko	**Korean**	320742	32.60	2.37	0.83	99.88	12	7	0.45	23.17	14.14	7.76	59.45
mk	**Macedonian**	82743	15.88	2.46	1.33	20.47	5	1	0.48	12.88	25.32	3.34	56.50
ms	**Malay**	275031	19.47	1.40	0.75	22.08	1	1	0.50	32.43	15.52	1.81	56.19
ne	**Nepali**	29114	29.69	11.77	2.16	40.23	1	13	0.46	-	22.02	3.30	58.29
no	**Norwegian**	415015	26.82	5.51	0.77	29.55	2	1	0.49	24.42	12.96	1.82	56.30
fa	**Persian**	460523	11.03	10.33	0.71	30.86	2	13	0.46	6.83	12.40	2.26	56.51
pl	**Polish**	1122218	23.15	9.42	1.08	23.91	1	1	0.50	25.86	9.35	1.29	56.24
pt	**Portuguese**	880529	19.05	1.99	1.01	24.24	4	0	0.49	21.58	11.23	2.43	56.39
ro	**Romanian**	329925	20.74	7.24	1.11	24.11	3	2	0.49	19.02	16.89	3.45	56.19
ru	**Russian**	1237127	31.23	10.98	1.14	33.68	1	1	0.50	29.10	8.25	2.20	58.23
sr	**Serbian**	321912	12.05	3.22	0.14	13.04	2	2	0.49	22.75	16.04	4.72	56.33
es	**Spanish**	1147742	27.65	4.96	1.98	30.33	5	1	0.48	30.60	9.32	3.37	56.57
sw	**Swahili**	29168	18.30	3.58	0.99	21.26	2	2	0.49	31.84	39.97	3.67	56.38
sv	**Swedish**	1970808	11.42	4.34	0.42	12.31	9	2	0.47	13.64	5.98	1.45	56.85
tr	**Turkish**	249061	33.90	4.39	2.06	44.79	6	0	0.48	0.00	16.21	3.38	59.26
uk	**Ukranian**	581735	24.84	6.78	1.01	26.56	3	2	0.49	32.20	12.88	2.41	56.12
vi	**Viatnamese**	1137180	2.47	0.88	0.23	4.55	2	0	0.49	8.31	7.36	1.45	56.75
war	**Waray**	1259278	0.04	0.00	0.02	0.05	2	0	0.49	-	6.32	10.89	56.74

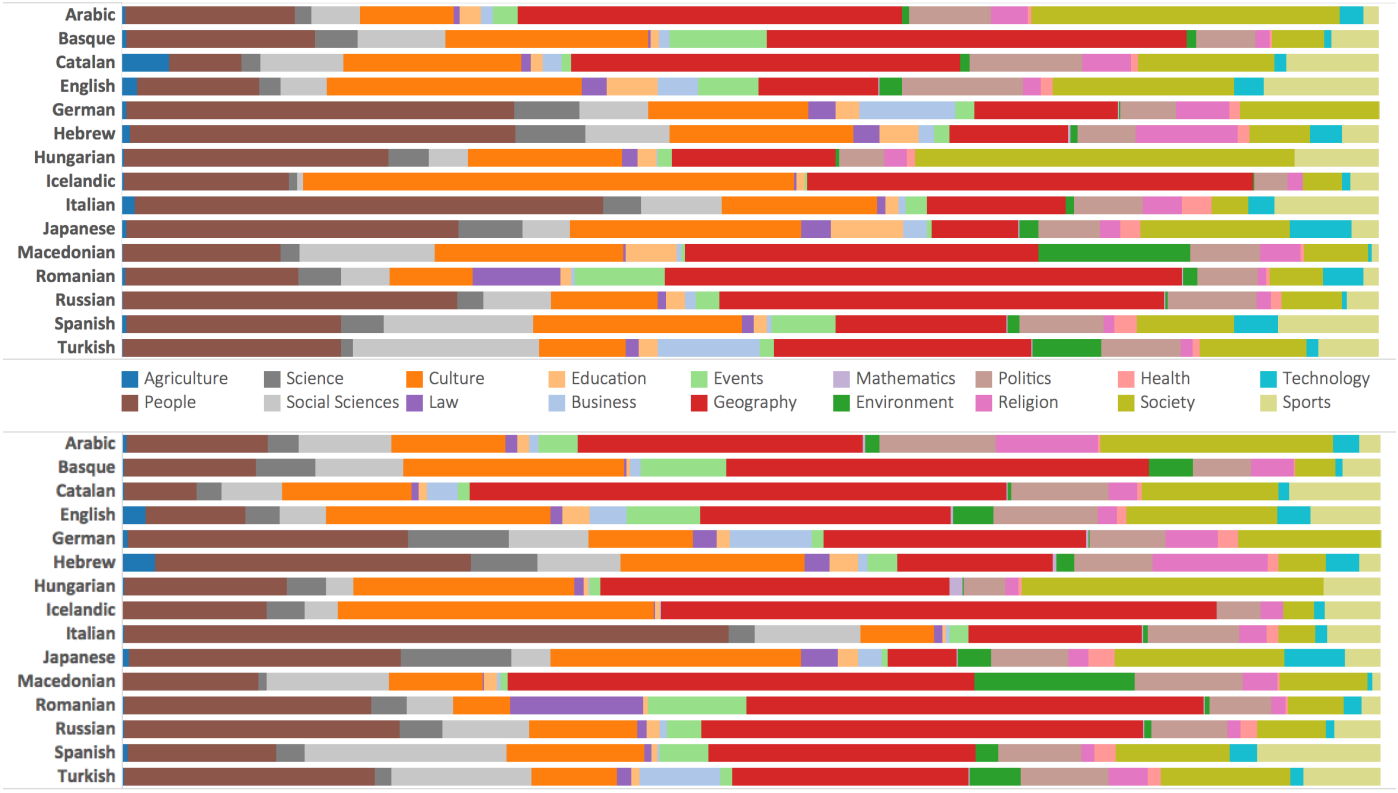

Figure 3: Topical coverage distribution in Cultural Identity Related Articles. Top: by number of articles (a), bottom: by number of Interlanguage Links (b).

low proportions of CIRA; this is because these editions have been mostly grown by an automatic program (bot) which massively created and translated articles from other language editions[11].

These cases are especially interesting because they indicate that CIRA may exist as long as there are editors involved in the community. To further investigate this relationship, we computed the Pearson correlation between CIRA percentage and number of editors. We found a correlation of 0.405 (p=0.013), which implies that the more editors contributing in a language edition, the more articles related to the corresponding cultural identities. This is consistent with the idea that identity-based motivation and cultural identity tend to affect all editors regardless of their activity level, who reaffirm it by contributing to cultural identity elements they share.

To inspect the quality of content related to the cultural identity of each Wikipedia, we looked at 'featured articles', a special category for those articles that according to editors deserve a mention of quality according to their characteristics[12]. We calculated the proportion of CIRA among featured articles for the 35 languages in our dataset in which this category exists, and we found an average of 27.8% (median 27.5%, standard deviation 13.7%). This proportion is higher than the proportion of CIRA articles, which indicates that high quality articles are more likely to be related to editors' cultural identities.

4.2 Topical Coverage of CIRA

We analyzed the topical coverage of CIRA to see the different shared meanings necessary to understand the editors' territories, and local contexts from each language edition. In order to do so, we used the method employed by [16], which consists of assigning each article's categories to one or more top level categories representing general topics, choosing the closest in the category hierarchy. Then, it is possible to obtain a distribution of topics for a group of articles. We expanded the top level categories according to [4] to a total of 18 main categories to cover all the very different encyclopedic themes, and only analyzed the 15 language editions having an equivalent category for each of them. The result of this is shown in Figure 3 (a) for the 15 Wikipedias. On average, we find Geography as the biggest category in CIRA (22%), followed by People (19.4%), Culture (14.7%), Society (9.8%), Social Sciences (6.2%), and others **(RQ2-Topics)**. When we compare the results for the English Wikipedia with the ones reported by [16], we see that these five categories represented a 82% of the encyclopedia vs. the 43% they represent in CIRA, and the order and proportions in the entire English language edition were quite different, with Culture (20.2%), People (9.6%), Geography and places (9.5%), Society and Social sciences (3.6%). Although this change can be due partly to the time between the two studies, a strong difference appears between CIRA and the entire encyclopedia, being the first more distributed into different topics. In fact, the Geography and People categories (whose sum makes 41.4%) are

[11] http://www.wsj.com/articles/for-this-author-10-000-wikipedia-articles-is-a-good-days-work-1405305001

[12] http://en.wikipedia.org/wiki/Wikipedia:Featured_articles

the dominant in every language edition's CIRA. This was expected because of the Cultural Identity selection criteria.

The cross-cultural comparison of the different CIRA topical coverage shown in Figure 3 allows us to see which topics have more representation in each language edition. We note that some patterns appear to confirm common knowledge about cultures. For instance, the Japanese cultural identity appears as the one with most articles categorized as technological, while the Hebrew the one with more religion, and the Icelandic has a strong prominence of culture and geography. Across all these data, it is readily apparent that the CIRA from each language edition include specialized topics as if they were local encyclopedias placed inside Wikipedia, and a wider analysis with more categories could provide valuable insights.

4.3 Difference and Otherness: CIRA Cross-language Availability

We examined the cross-language availability of CIRA from the 40 selected language, expecting to see uniqueness, since Cultural Identities are defined as shared meanings in a group but also in terms of difference from one another. In Wikipedia, an article is available in other language editions when it has Interlanguage links (ILL), which can be placed by an editor of any of the two languages, or by an automatic program (bot). In a way, the bigger encyclopedias act as leaders and the other editions can copy, translate, and adapt content [27]. An analysis of ILL shows, first and foremost, the degree of uniqueness of content related to cultural identity. Secondly, the analysis shows the relationship between different language editions in integrating one another's specific content, as well as the process of creating content in the overall Wikipedia as a multilingual project.

As seen in Table 1, the average number of ILLs per article is variable across languages - both in CIRA and WP. However, the average for CIRA is 4.5 times lower than for the entire language editions **(RQ-3-Cross-language).** Even though the average number of ILLs in CIRA is lower in all cases, the ratio is also variable. In fact, minor language editions like Icelandic, Afrikaans, Estonian and Swahili have between 9 and 13 times less ILLs in CIRA than in the total of their language editions. On the contrary, languages like English, French, Korean, German and Italian show a much smaller difference with CIRA having about the half of the ILLS than the whole encyclopedia average. These latter cases are coincident with some of the biggest Wikipedia language editions, which in a way confirms that both language status and Wikipedia size and development matter also for CIRA. Interestingly, one pattern that remains with great stability across languages is the percentage of cultural identity related articles totally unique to one language (zero ILLs), with a majority of 57.7% (median 56.5%, standard deviation 3.3%).

In order to see how the average value of ILLs in CIRA changed across time, we compared it with that obtained in previous research for a similar dataset from 2011 [22]. We observe that in four years the number of ILLs for the entire Wikipedia language editions has doubled, while for CIRA ILLs have remained in similar low proportions. This indicates that while the rest of the encyclopedic content is increasingly shared and "globalized" between languages over years, CIRA tend to remain mainly of local interest.

To further investigate the ILLs in CIRA, we crossed these results with those from topical coverage, so to understand whether certain topics from one language's cultural identities appear more relevant to other cultures. The results are shown in Figure 3(b), which represents the number of ILLs per article, by topics, showing a classification of topics according to cross-language availability. Similarly to the previous topical coverage results with articles, the most representative category is Geography, which exhibits a generally higher proportion of ILLs than number of articles (26.1% vs 22.0%), while the second one, People, has a slightly lower percentage (17.4% vs 19.4%). This suggests that when editors from a language edition import content from another language 's cultural identity, they will likely consider these topics as the most notable things to learn first. Although the two graphics in Figure 3(a) and 3(b) are generally similar, some remarkable differences can be noticed for some categories, such as Religion in the Arabic CIRA, that contains few articles, but has a much higher proportion of ILLs, indicating that these articles are often shared with other language editions. We observe a similar effect for example for Sports in the Spanish Wikipedia.

All in all, differences observed in CIRA cross-language availability have shown us that there are two very distinct types of articles, whose proportions will depend on each language edition: (i) those unique to a language edition and whose meaning will probably be shared by few editors, and (ii) those with many ILLs, and therefore shared by many languages, as a result of becoming an important symbol for that Cultural Identity.

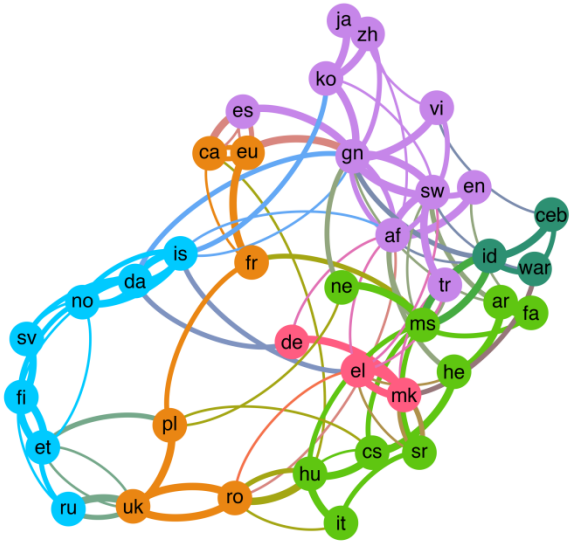

Figure 4: Network graph with CIRA.

Taking a closer look at CIRA's Interlanguage links, it is possible to obtain a better understanding of the proximity between cultural identities or their expansion. In Figure 4 we depict a network of languages to show which have a higher proportion of articles associated to other languages' cultural identities. More exactly, for each Wikipedia we computed the proportion of articles corresponding to other languages' CIRA. Then we selected, for each CIRA, the three languages in which it is represented in higher proportion and we drew the corresponding edges. Following a standard convention in graph representation, edges are curved and drawn in clockwise direction. Colors are assigned according to the clusters identified by an automatic clustering algorithm (the Louvain method), to highlight groups of language editions that are closer to each other.

We can see that Nordic languages form a cluster together with Russian, while Iberic languages are tightly close to each other, as well as Asian languages, and Middle East languages. These results confirm the importance of geographic proximity according to Tobler's Law, which states that things near tend to be similar [25], and results obtained comparing the availability of biographies in different languages [1, 3]. However, some less expected relationships also emerge, such as the relevance of Italian CIRA in the Hungarian Wikipedia.

5. CONCLUSIONS AND FUTURE LINES OF RESEARCH

5.1 Findings and Recommendations

What motivates Wikipedians has been largely studied in order to explain their dedication and contributions into the free encyclopedia [19, 28]. However, to our knowledge no previous research considered looking at editors' identity as a motivation to act in Wikipedia, where social identity may also play a central role. In this sense, Oyserman's model of an identity-based motivation has been useful to illuminate the process of contributing with content as an identity-congruent act. Articles imbue meanings related to any editor identity, including the Cultural Identity codes associated to the territories in which they live in.

5.1.1 Main Findings

Our findings confirm identity-based motivation as the driver of cultural contextualization of Wikipedia language editions, in which the extent of Cultural Identity Related Articles proves its influence. Below we summarize our main conclusions and some recommendations derived from this study.

Our first research question **(RQ1-Extension)** concerned the relevance of local cultural identities in each corresponding Wikipedia language edition. According to our method, a range of 7% to 49% of the articles in the 40 analyzed languages is on topics related to editors' cultural identities. CIRA have been spontaneously produced, with no policy or guideline recommending it, as the cumulative effect of editors' choices in content. Even though the method has been run on very different language editions, the relative size of CIRA correlates with the total number of editors and not with the current active editors. This is in agreement with the concept of Cultural Identity, which ties all editors sharing those meanings, independently of their level of involvement in the Wikipedia community.

Our second research question **(RQ2-Topical)** referred to the topical coverage of the content representing cultural identities, in order to understand which meanings conform them and how they can explain editors' context. We found that Geography and People categories occupy a dominant position, however other categories also play a role in expressing the diversity within the group of CIRA. Cultural Identity has been conformed in relation to a territory and power. Editors need to understand their very environment and reflect all these meanings in Wikipedia. Therefore, the result of their contributions is similar to a local specialized version of an encyclopedia.

The third and last question **(RQ3-Cross-language)** was about cross-language availability of CIRA, in order to see if the opposition marked by the definition between Cultural Identities is also effective in the selection of articles. According to our analysis based on ILLs in the 40 languages, CIRA articles are 4.5 times less shared than the average. Furthermore, an average proportion of 57.7% CIRA articles do not exist in any other language, which proves Cultural Identity as a source of articles with local audience. This value is very stable (standard deviation 3.3%) in opposition to the variability of ILL found in entire language editions by previous research [27]. Regarding topical coverage, articles on geography are the most shared across languages.

5.1.2 Recommendations for Intercultural enrichment

Both in the Wikipedia and in the research community the geographical imbalance of content has been considered an issue, explained by several demographic and territory factors. Our results showing the significant extent and uniqueness of CIRA are in line with previous studies. This study aims to provide an explanation of how editors create this content, digging into the need to act congruently with their cultural contexts, involving not just geography but many other themes.

Often the Wikipedia English language edition has been considered as a possible neutral language, for several reasons: being the first in creation; its leadership in number of articles; and importantly its status of lingua franca as a global reference with editors from all countries. In this regard, the English Wikipedia is the second language in which multilingual editors contribute [5], and in general it is the one containing more Cultural Identity Related Articles (CIRA) from other languages in absolute terms. However, far from having a reduced proportion of CIRA as one could expect from a markedly multicultural encyclopedia, it has a 46.8% of articles related to its cultural identities, second only after the Japanese.

As the relative importance of CIRA does not decrease with increasing size of an encyclopedia, but relates to one of the main editors' unconscious motivations, we believe the imbalance will persist as long as the Wikipedia project continues under the same content notability guidelines. Consequently, we suggest not to consider these imbalances in content as a bias, but rather to embrace cultural diversity by promoting and facilitating editors from each language edition to spread their cultural identities across languages. This is especially important when considering the influence identity-based motivation has demonstrated to have. Therefore, we suggest that the translator and the article recommendation tool[13] developed by the Wikimedia Foundation could include CIRA or subparts of it (e.g. articles including cultural identity related keywords in their title) as preferential content to translate and export across languages. Collaboration across languages can be useful to bring each of them closer to the goal of achieving the sum of human knowledge at an encyclopedic level, while at an article level the contrast of culturally different points of view can help to reach a more neutral point of view.

5.2 CIRA Datasets and Future Research Lines

We provided a methodology to obtain Cultural Identity Related Articles that takes a territory and its people as a reference to obtain a set of articles and filters it against interference. Manual assessment resulted into a 3.3% of false positives and 3.4% of false negatives. To improve accuracy, thresholds could be adjusted, although the more a Wikipedia language edition grows and geolocates its articles, the more reliable the ground-truth

[13] http://recommend.wmflabs.org/

will become. Other strategies to diminish interference would be to use articles solidly included as CIRA for another language as a negative ground-truth. Machine learning approaches could also be used to improve accuracy. We want to remark that the method we proposed in this study could either be applied to other kinds of editor identities across languages, such as religion, professional careers, hobbies, gender, etc. This would require finding proper keywords and setting additional filtering to ensure low interference.

As an important contribution of this paper, we make available both the code we used to process the Wikipedia language editions as well as the processed datasets. We believe this can motivate and encourage new research on cultural identities. The two approaches we used to verify and understand cultural identity, a topical coverage and a cross-language analysis, can be developed into more depth to bring new insights on particular cultural identities within a language (e.g., British with respect to the English language edition) or even across different ones in the same territory (e.g., assessing differences and similarities between English CIRA and Gaelic CIRA about Ireland).

An interesting aspect to be further evaluated is the overlap between CIRA and other groups of articles such as the most read ones (in terms of page views) or those which cover breaking news and current events. In general, many avenues of research especially in the field of Humanities can use CIRA as a source to study particular subjects from a cultural identity perspective. In the same way, one interesting aspect that we left unattended in this research is the representation of multiculturalism in Wikipedia, or the study of which specific meanings originary from some cultural identities end up reaching world attention.

6. REFERENCES

[1] Aragón, P., Laniado, D., Kaltenbrunner, A. and Volkovich, Y. 2012. Biographical social networks on Wikipedia: a cross-cultural study of links that made history. *Proc.WikiSym*.

[2] Bryant, S.L., Forte, A. and Bruckman, A. 2005. Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. *Proc. SIGGROUP*.

[3] Eom, Y.-H., Aragón, P., Laniado, D., Kaltenbrunner, A., Vigna, S. and Shepelyansky, D.L. 2015. Interactions of cultures and top people of Wikipedia from ranking of 24 language editions. *PloS one*. 10, 3 (2015), e0114825–.

[4] Farina, J., Tasso, R. and Laniado, D. 2011. Automatically assigning Wikipedia articles to macrocategories. *Proc. Hypertext*.

[5] Hale, S.A. 2014. Multilinguals and Wikipedia editing. *Proc. Web Science*.

[6] Hall, S. 1990. *Cultural identity and diaspora*. Editorial Jonathan Rutherford.

[7] Hall, S. 1997. *Representation: Cultural representations and signifying practices*.

[8] Hecht, B. and Gergle, D. 2009. Measuring self-focus bias in community-maintained knowledge repositories. *Proc. C&T*.

[9] Hecht, B. and Gergle, D. 2010. The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context. *Proc. CHI*.

[10] Hecht, B.J. 2013. *The Mining and Application of Diverse Cultural Perspectives in User-Generated Content*, PhD Thesis, Northwestern University.

[11] Hecht, B.J. and Gergle, D. 2010. On the localness of user-generated content. *Proc. CSCW*.

[12] Hofstede, G., Hofstede, G.J. and Minkov, M. 2010. *Cultures and Organizations: Software of the Mind*, McGraw Hill.

[13] Karimi, F., Bohlin, L., Samoilenko, A., Rosvall, M. and Lancichinetti, A. 2015. Quantifying national information interests using the activity of Wikipedia editors. *arXiv preprint arXiv*: 1503.05522.

[14] Keegan, B., Gergle, D. and Contractor, N. 2013. Hot Off the Wiki: Structures and Dynamics of Wikipedia's Coverage of Breaking News Events. *American Behavioral Scientist*. 57, 5, 595–622.

[15] 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*.

[16] Kittur, A., Chi, E.H. and Suh, B. 2009. What's in Wikipedia?: mapping topics and conflict using socially annotated category structure. *Proc. CHI*.

[17] Massa, P. and Scrinzi, F. 2011. Exploring linguistic points of view of Wikipedia. *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*.

[18] Neff, J.J., Laniado, D., Kappler, K.E., Volkovich, Y., Aragón, P. and Kaltenbrunner, A. 2013. Jointly They Edit: Examining the Impact of Community Identification on Political Interaction in Wikipedia. *PloS one*. 8.4: e60584.

[19] Nov, O. 2007. What motivates Wikipedians? *Communications of the ACM*. 50, 11, 60–64.

[20] Oyserman, D. 2009. Identity-based motivation and consumer behavior. *Journal of Consumer Psychology*. 19, 3, 250–260.

[21] Oyserman, D. and Destin, M. 2010. Identity-based motivation: Implications for intervention. *The Counseling Psychologist*. 38, 7, 1001–1043.

[22] Ribé, M.M. 2011. Cultural configuration of Wikipedia: measuring Autoreferentiality in different languages. *Proc. RANLP*.

[23] Rizoiu, M.-A., Xie, L., Caetano, T. and Cebrian, M. 2015. Evolution of Privacy Loss in Wikipedia. *arXiv preprint arXiv*:1512.03523.

[24] Rogers, R. and Sendijarevic, E. 2012. Neutral or National Point of View? A Comparison of Srebrenica articles across Wikipedia's language versions. *Proc. Wikipedia Academy*.

[25] Suh, B., Convertino, G., Chi, E.H. and Pirolli, P. 2009. The singularity is not near: slowing growth of Wikipedia. *Proc. WikiSym.*

[26] Van Dijk, Z. 2009. Wikipedia and lesser-resourced languages. *Language problems & Language planning*. 3, 33.

[27] Warncke-Wang, M., Uduwage, A., Dong, Z. and Riedl, J. 2012. In search of the ur-Wikipedia: universality, similarity, and translation in the Wikipedia inter-language link network. *Proc. WikiSym*.

[28] Xu, B. and Li, D. 2015. An empirical study of the motivations for content contribution and community participation in Wikipedia. *Information & Management*. 52, 3, 275–28