Manypedia: Comparing Language Points of View of Wikipedia Communities

Paolo Massa
Bruno Kessler Foundation
Via Sommarive, 18
Trento, Italy
massa@fbk.eu

Federico Scrinzi
Bruno Kessler Foundation
Via Sommarive, 18
Trento, Italy
fscrinzi@fbk.eu

ABSTRACT

The 4 million articles of the English Wikipedia have been written in a collaborative fashion by more than 16 million volunteer editors. On each article, the community of editors strive to reach a neutral point of view, representing all significant views fairly, proportionately, and without biases. However, beside the English one, there are more than 280 editions of Wikipedia in different languages and their relatively isolated communities of editors are not forced by the platform to discuss and negotiate their points of view. So the empirical question is: do communities on different language Wikipedias develop their own diverse Linguistic Points of View (LPOV)? To answer this question we created and released as open source Manypedia, a web tool whose aim is to facilitate cross-cultural analysis of Wikipedia language communities by providing an easy way to compare automatically translated versions of their different representations of the same topic.

Keywords

Wikipedia, cross-cultural comparison, Linguistic Point of View, languages, automatic translation, web tool, open source.

1. INTRODUCTION

Wikipedia is becoming one of the most accessed Web resources for information needs. 53% of American Internet users look for information on Wikipedia as of May 2010 and this number increased from 36% in February 2007 [21]. A survey found that 88% of 2,318 university students use Wikipedia during a course–related research process, even if an instructor advised against it [11]. According to Alexa, Wikipedia is the sixth most visited site of the entire web [1].

It is hence clear that a large share of people rely on Wikipedia for forming their representations of facts, of what is true and what is not. This point is even more interesting considering that every single word on which so many people rely could have been added by anyone. In fact Wikipedia's slogan is "the free encyclopedia anyone can edit" and indeed the almost 4 million articles of the English Wikipedia, since its inception in 2001, have received more than 526 million edits by more than 16 million registered users[1]. It is even possible to edit Wikipedia without performing the login with the personal username and hence to edit the encyclopedia anonymously. Despite this ultimate openness, the quality of Wikipedia articles is relatively high. A 2005 investigation by the scientific journal Nature found out that "Wikipedia comes close to Britannica in terms of the accuracy of its science entries". They also report how more than 70% of Nature authors consult Wikipedia on scientific topics [9].

Given the importance in shaping the wisdom and world view of so many people, we believe it is important to raise awareness on who are the people who edit Wikipedia. This concern is fully shared by the Wikipedia community itself: for instance the "Wikipedia" namespace devoted to policies and rules contains a page titled "Wikipedia:Systemic bias"[2] which states that "the Wikipedia project suffers from systemic bias that naturally grows from its contributors' demographic groups, manifesting an imbalanced coverage of a subject, thereby discriminating against the less represented demographic groups." which states "the Wikipedia project suffers from systemic bias that naturally grows from its contributors' demographic groups, manifesting an imbalanced coverage of a subject, thereby discriminating against the less represented demographic groups." The page clearly lists the main biases: "The average Wikipedian on the English Wikipedia is a male, technically inclined, formally educated, an English speaker (native or non-native), European–descent, aged 15–49, from a majority-Christian country, from a developed nation, from the Northern Hemisphere, and likely employed as a white-collar worker or enrolled as a student rather than employed as a labourer". There is even a project, described and coordinated at the page "Wikipedia:WikiProject Countering systemic bias", which lists what Wikipedians can do in order to counter this important issue.

In this paper we are not interested in the biases internal to a specific wiki such as the English Wikipedia but we focus on the existence (or absence) of different biases in different language communities of Wikipedia. In fact, while the largest and oldest

WikiSym'12, Aug 27–29, 2012, Linz, Austria.

[1] The data reported in this paper are taken from the Wikipedia site as they appeared on April 3, 2012.

[2] Since this article deals with Wikipedia pages, it cites many of them. In order not to clutter the paper with too many footnotes or citations, we simply report the title of the Wikipedia page between brackets such as "Page title". Wikipedia pages were accessed on March 10, 2012 and to get a version of any "Page title" at that date using the history feature of Wikipedia, the reader can visit the URL http://en.wikipedia.org/w/index.php?action=history&limit=1&offset=20120310000000&title=Page_title where the offset parameter indicated the retrieval date.

Wikipedia is in English language, there are currently more than 280 editions of Wikipedia in as many different languages, ranging from many with more than 700,000 articles such as the German, French, Polish, Italian, Japanese and Spanish ones, up to smaller ones in languages such as Wolof, Catalan, Piedmontese, Latin, Esperanto, Tibetan, Haitian and more.

So the motivating question for this contribution is "do people who self-elect for editing the page about Palestine in the English Wikipedia have and represent the same points of view of people who self-elect to edit the counterpart article on the Arabic Wikipedia or on the Hebrew Wikipedia?" We call this lens of investigation, Linguistic Point of View (LPOV).

This paper is structured as follows. First, in Section 2, we describe Wikipedia editing policies and in particular the neutral point of view (NPOV) policy and, in Section 3, we highlight the richness of the multi-cultural phenomenon represented by the communities of the different languages Wikipedias. Then Section 4 is devoted to presenting Manypedia, the Web tool we have created and deployed whose aim is to facilitate the comparison and analysis of different points of view represented in the equivalent articles about the same topic as they appear on two different language Wikipedias. We conclude by introducing examples of comparisons that show how the tool can be used for research and scientific investigation purposes and maintenance of Wikipedia as a healthy and balanced cross-cultural project.

2. POINTS OF VIEW AND NEUTRALITY ON WIKIPEDIA

A project which exhibits such a large openness and inclusiveness such as Wikipedia would hardly be possible without precise rules and guidelines. In fact, over the years, a complex and vast set of rules were developed by the Wikipedia community through the distributed contributions and negotiations of thousands of people, just as content articles did. Policies and rules are pages on the namespace "Wikipedia:".

Among the most important rules, there are the three core content policies reported on the pages "Wikipedia:Neutral point of view", "Wikipedia:Verifiability" and "Wikipedia:No original research".

The first one, neutrality, is the policy defining Wikipedia itself or, as Roy Rosenzweig in "Can History Be Open Source? Wikipedia and the Future of the Past" [23] puts it, the "founding myth". Its definition in a nutshell is "Editors must write articles from a neutral point of view, representing all significant views fairly, proportionately, and without bias." The neutral point of view (NPOV) policy "says nothing about objectivity" and "in particular, the policy does not say that there is such a thing as objectivity in a philosophical sense—a "view from nowhere" [18], such that articles written from that viewpoint are consequently objectively true." "Rather, to be neutral is to describe debates rather than engage in them. In other words, when discussing a subject, we should report what people have said about it rather than what is so." Once defined the goal, the page goes on pondering on the feasibility of such a task: "is it possible to characterize disputes fairly? This is an empirical issue, not a philosophical one: can we edit articles so that all the major participants will be able to look at the resulting text, and agree that their views are presented accurately and as completely as the context permits? It may not be possible to describe all disputes with perfect objectivity, but it is an aim that thousands of editors strive towards every day" (quotations from page "Wikipedia:Neutral point of view/FAQ").

The two other content core policies are less controversial. Verifiability refers to the fact that material written on Wikipedia must be attributed to a reliable, published source so that readers can check that the specific material has already been published; again, the goal is not truth but verifiability.

The "no original research" content policy states that "Wikipedia does not publish original thought: all material in Wikipedia must be attributable to a reliable, published source. Articles may not contain any new analysis or synthesis of published material that serves to advance a position not clearly advanced by the sources."

Of course, writing "without bias" is "difficult" since "all articles are edited by people" and "people are inherently biased" [23]. This is testified by the many "edit wars" on Wikipedia pages [14,27]. An edit war occurs when two or more users who disagree about the content of a page repeatedly override (revert) each other's contributions, rather than trying to resolve the disagreement by discussion.

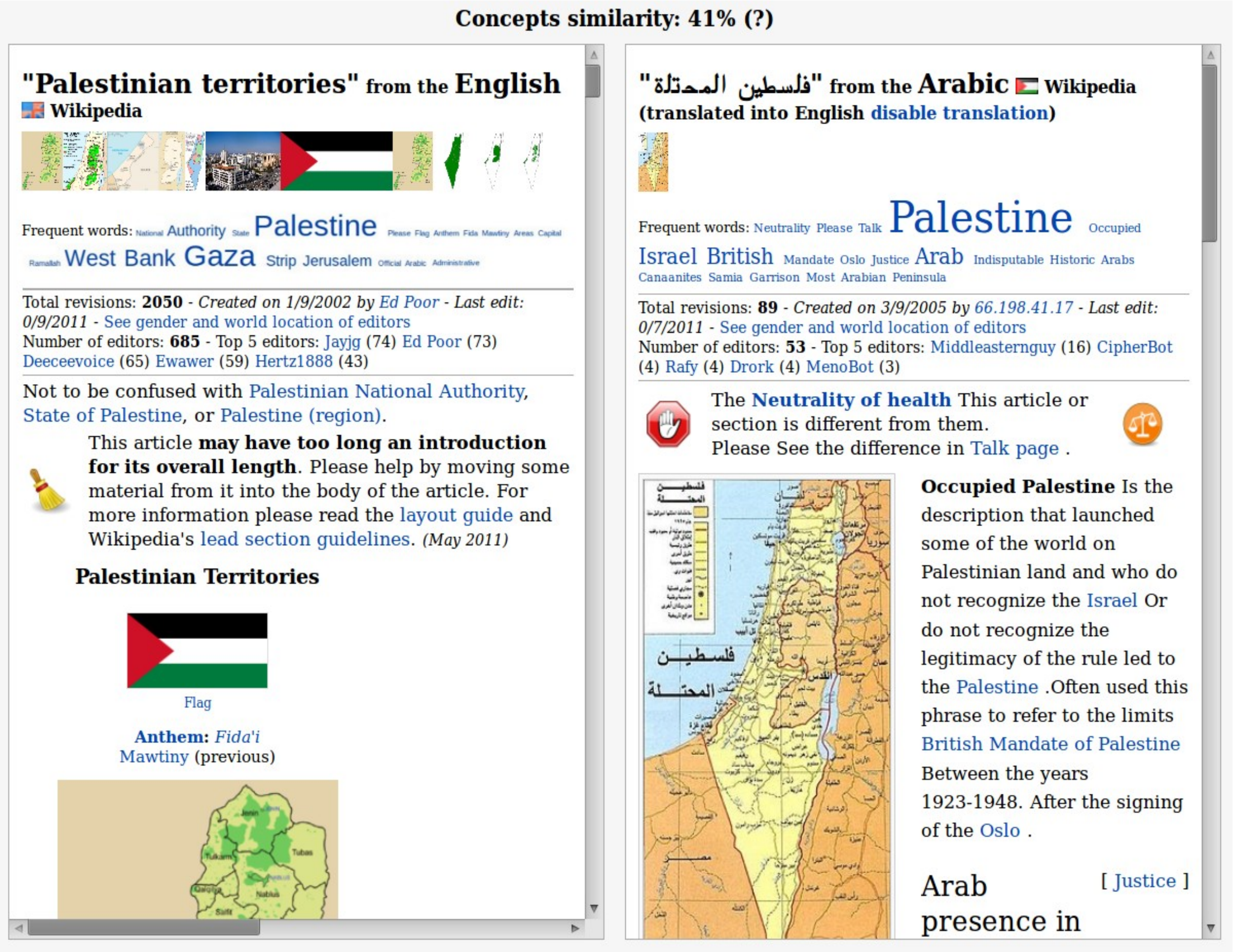
In fact, consensus is the primary way in which editorial decisions are made on Wikipedia with the goal of establishing and ensuring neutrality and verifiability. Usually consensus is reached as a "natural and inherent product of editing; generally someone makes a change or addition to a page, then everyone who reads it has an opportunity to leave the page as it is or change it". However, "when editors cannot reach agreement by editing, the process of finding a consensus is continued by discussion on the relevant talk pages" (page "Wikipedia:Consensus"). In fact, each article page, fox example "Palestinian territories" has an associated discussion page in the "Talk" namespace such as "Talk:Palestinian territories" where editors can discuss changes and improvements.

Rosenzweig argues that the most frequent debate topic on discussion pages is whether the article adheres to the NPOV and cites the "Armenian genocide" page as one example of the fact "those debates can go on at mind-numbing length, such as literally hundreds of pages" [23].

There are articles, as we will see in the following, on which it seems to be harder to reach consensus and civil discussions. Many of these articles are linked from a page titled "List of controversial issues" and often are flagged with a warning message, signaling that at least one Wikipedian believes this pages in not neutral and sometimes these pages are even blocked in editing. In fact, users with additional powers (administrators) can block a page of Wikipedia, in order to stop edit wars and cool down discussions, during periods in which consensus is not reached and discussions are particularly heated. It is interesting to note again that editing an article and talk pages can be performed even by anonymous users, identified only by their Internet address.

Beside the apparent theoretical difficulty of reaching consensus among hundreds of editors sometimes very vocal about a certain article topic, perhaps surprisingly the social process works quite well and the community is, up to now, able to self-control itself and edit wars are very limited considering the dimension of the active community.

After this short description of the main collective process happening around each page and each edit, we would like to go back to the neutral point of view concept. "Neutrality requires that each article or other page in the mainspace fairly represents all significant viewpoints that have been published by reliable sources, in proportion to the prominence of each viewpoint" ("Wikipedia:NPOV"). Each article should "accurately indicate the relative prominence of opposing views. Ensure that the reporting of different views on a subject adequately reflects the relative levels of support for those views, and that it does not give a false impression of parity, or give undue weight to a particular view.



Concepts similarity: 41% (?)

"Palestinian territories" from the English Wikipedia

Frequent words: National Authority State Palestine Please Flag Anthem Fida Mawtiny Areas Capital Ramallah West Bank Gaza Strip Jerusalem Official Arabic Administrative

Total revisions: **2050** - *Created on 1/9/2002 by* Ed Poor *- Last edit: 0/9/2011* - See gender and world location of editors
Number of editors: **685** - Top 5 editors: Jayjg (74) Ed Poor (73) Deeceevoice (65) Ewawer (59) Hertz1888 (43)

Not to be confused with Palestinian National Authority, State of Palestine, or Palestine (region).

This article **may have too long an introduction for its overall length**. Please help by moving some material from it into the body of the article. For more information please read the layout guide and Wikipedia's lead section guidelines. *(May 2011)*

Palestinian Territories

Flag

Anthem: *Fida'i* Mawtiny (previous)

"فلسطين المحتلة" from the Arabic Wikipedia (translated into English disable translation)

Frequent words: Neutrality Please Talk Palestine Occupied Israel British Mandate Oslo Justice Arab Indisputable Historic Arabs Canaanites Samia Garrison Most Arabian Peninsula

Total revisions: **89** - *Created on 3/9/2005 by* 66.198.41.17 *- Last edit: 0/7/2011* - See gender and world location of editors
Number of editors: **53** - Top 5 editors: Middleasternguy (16) CipherBot (4) Rafy (4) Drork (4) MenoBot (3)

The **Neutrality of health** This article or section is different from them.
Please See the difference in Talk page .

Occupied Palestine Is the description that launched some of the world on Palestinian land and who do not recognize the Israel Or do not recognize the legitimacy of the rule led to the Palestine .Often used this phrase to refer to the limits British Mandate of Palestine Between the years 1923-1948. After the signing of the Oslo .

Arab presence in

[Justice]

Figure 1: Manypedia comparison of "Palestinian territories" page on English and Arabic Wikipedia (http://www.manypedia.com)

For example, to state that "<According to Simon Wiesenthal, the Holocaust was a program of extermination of the Jewish people in Germany, but David Irving disputes this analysis> would be to give apparent parity between the supermajority view and a tiny minority view by assigning each to a single activist in the field".

In fact by giving different relevant prominence to specific points of view it is possible to write different histories. As George Orwell wrote in 1944, during the 2nd World War, "a Nazi and a non-Nazi version of the present war would have no resemblance to one another, and which of them finally gets into the history books will be decided not by evidential methods but on the battlefield" [20].

Interestingly in 2011 we are in a different situation: paraphrasing Orwell, we could argue that what eventually gets into Wikipedia is decided, not on the battlefield, but in a bottom-up collaborative fashion by millions of people, through discussions and negotiations of different points of view.

This bottom-up freedom can be considered undesirable by some governments and central institutions which might want to give more emphasis to "official" top-down points of view or even censor some points of view. This is hard to do on Wikipedia since every article is the result of the negotiations among the points of view of the people who self-elect for editing it. An interesting attempt to detect which organization is behind anonymous changes to Wikipedia pages which can be classified as propaganda is Wikiganda [6].

Along a similar line, there are reports about states censoring and making unreachable Wikipedia possibly in an attempt to control the spread of unwelcomed information. It is not easy to track when a web site is blocked on a country and if the block is for every page or just for some pages. An interesting example of this is China and an obviously partial report of the situation over the years is written at the Wikipedia page "Blocking of Wikipedia by the People's Republic of China". Sometimes the block is reported to be only partial on selected articles such as "Falun Gong" (a persecuted religious practice) and "Tiananmen Square protests of 1989". In China, possibly also because of these blocks, there are two wiki-based knowledge repositories that are larger than the Wikipedia in Chinese language, at least according to the numbers reported by them: Hudong.com has more than 4 million articles and Baidu Baike has almost 3 million articles while Chinese Wikipedia has around 350.000 articles. As we will see, on the Chinese Wikipedia there is no evident bias on information coming from the government point of view.

Similar motivations might have pushed the government of Cuba to launch its own online encyclopaedia, Ecured, "with the goal of presenting its view of the world and history" [3]. Interestingly the wiki clearly mentions that Ecured is built "from a decolonizer point of view". The entry on the United States, for example, describes it as the "empire of our time, which has historically taken by force territory and natural resources from other nations, to put at the service of its businesses and monopolies" and that "it consumes 25% of the energy produced on the planet and in spite of its wealth, more than a third of its population does not have assured medical attention" [3]. These quotations are a clear example of specific points of view and of the different prominence they can get in articles of online encyclopedias.

The large number of discussions abour what is the major point of view and which are the minor ones and how much relative prominence they should receive is one the most discussed topics on Wikipedia talk pages [23], especially because it is hard for someone holding a certain POV to be neutral and balanced.

Moreover, as the Cuba example above summarizes, relevant viewpoints can be different for different communities and surely the prominence of each viewpoint can be very different. Anarchopedia (http://anarchopedia.org) and Conservapedia (http://conservapedia.com) are two online encyclopedias which constitute themselves as alternatives to Wikipedia. Their "founding myth" [23] is in fact a specific point of view, respectively the "anarchistic point of view" and the "conservative viewpoint". While they gathered a community that is much smaller than Wikipedia, they exemplify a phenomenon: different histories are written depending on the point of view the community choose to adopt.

On the other hand, Wikipedia has neutrality as its "founding myth" and "it is an aim that thousands of editors strive towards every day" ("Wikipedia:Neutral point of view/FAQ").

Some people are skeptical that neutrality can be reached. For example, Larry Sanger, one the two founders of Wikipedia but who left it in 2002, argues that "over the long term, the quality of a given Wikipedia article will do a random walk around the highest level of quality permitted by the most persistent and aggressive people who follow an article" [24].

Beside being optimistic or pessimistic about the fate of Wikipedia in the long run, we believe the simple act of discussing is central to democracy, a healthy global society and peaceful coexistence of different points of view. Rosenzweig considers that "those who create Wikipedia's articles and debate their contents are involved in an astonishingly intense and widespread process of democratic self-education" and reports that the classicist James O'Donnell has argued that the benefit of Wikipedia may be greater for its active participants than for its readers: "A community that finds a way to talk in this way is creating education and online discourse at a higher level" [23].

These levels of discussions and interactions are indeed admirable but in this paper we ask if similar levels of intra-Wikipedia community negotiation and self-education happen also inter-Wikipedia communities. Is it the case that users who strive to reach consensus on the page "Palestine" in the Arabic Wikipedia discuss and try to balance their views with users who self-elect for editing the equivalent page "Palestine" in the Hebrew Wikipedia? As we will see, this process is not too encouraged by the current socio-technical platform powering Wikipedia. So our empirical question is "will relatively isolated language communities of Wikipedia develop their own divergent representations for topics? Their own Linguistic Point of View (LPOV)?" This question has important implications for the cross-cultural mutual understanding and peaceful coexistence of world communities.

To this end, in next section we report on the richness of the different language Wikipedia communities while in Section 4 we present Manypedia, the web mashup we have created as a tool for helping in answering the previous question.

3. LANGUAGE WIKIPEDIA COMMUNITIES AND THEIR POINTS OF VIEW

There are more than 280 language editions of Wikipedia. Some of the most active ones are reported in Table 1. The largest one is the English Wikipedia which, started in 2001, currently counts almost 4 million articles which have received more than 526 million edits by more than 16 million registered users, as of April 2012. The users who performed at least an action in the past month, considered the active part of the community, are 138,763.

The next largest Wikipedia communities assemble around languages widely spoken in the world especially in countries with a good level of Internet penetration, such as German, French, Polish, Italian, Spanish and Japanese. On the other hand, a language spoken by billions of people such as Chinese has a relatively smaller community but we have already reported how there are two other online encyclopedias in Chinese which gathered more users.

For our purposes it is especially interesting to look at the column of users and in particular of active users in Table 1, referring to the current "force load" of the community. In fact different language editions of Wikipedia started at different times and it is important to understand the current situation of the community of Wikipedians. For instance, the Wikipedia in Catalan language can count on a relatively small number of very dedicated users and was able to create the 15th Wikipedia as number of articles. Statistics in Table 1 show that a small community of really dedicated users can generate a large number of articles, especially when they care significantly about their language and, probably, their cultural heritage and world view.

Each Wikipedia has its own history and, partially, its own community. In fact each language edition of Wikipedia is an independent installation of the Mediawiki server software. A relatively new feature, "Unified login", of Wikipedia allows to use the same username on all Wikipedias, as long as this is not already used by someone else. But of course this feature is used mainly by Wikipedians who know at least two languages and that are confident in contributing in both of them.

Different language Wikipedias are connected mainly, if not only, by interwiki links. In fact it is possible to link the article about, for example, "Palestine" in the Hebrew Wikipedia with its equivalent in Arabic Wikipedia simply by inserting an interwiki link of the form [[language code:Title]], for example [[ar:فلسطين]]. The Wikipedia server interprets this interwiki syntax and offers links to the equivalent page in the other language Wikipedia, on the left hand side of each Wikipedia page under a "Language" menu. These interwiki links must be inserted manually (or with the help of semi-automated programs called bots) by users who, at least in theory, know both the source and target language.

The page "Wikipedia" on Wikipedia reports that "translated articles represent only a small portion of articles in most editions" and also "in part because automated translation of articles is disallowed." While this claim should be empirically validated, it is surely interesting that one policy warns against automated translations of articles and hence it is expected that each article, in each language edition of Wikipedia, is written by a human who knows, at least partially, the language. There are many examples

Language	Prefix	Articles	Edits	Users	Active Users
English	en	3,912,105	526,077,676	16,551,900	138,763
German	de	1,385,907	106,360,859	1,403,006	22,218
French	fr	1,232,680	82,236,250	1,255,780	16,471
Dutch	nl	1,035,566	30,563,443	447,929	4,724
Italian	it	906,636	52,569,842	734,348	8,061
Polish	pl	889,056	30,490,485	482,397	5,126
Spanish	es	880,350	58,577,562	2,171,257	14,931
Russian	ru	840,061	45,884,948	824,528	12,638
Japanese	ja	799,975	42,776,541	609,565	11,971
Portuguese	pt	718,501	30,249,021	982,460	5,347
Swedish	sv	452,121	17,015,139	267,277	3,149
Chinese	zh	434,117	20,427,481	1,176,713	6,569
Vietnamese	vi	395,102	6,565,432	275,115	1,195
Ukrainian	uk	377,111	9,371,511	129,879	2,075
Catalan	ca	369,852	9,431,935	110,711	1,427
Norwegian	no	333,538	10,858,373	229,329	2,169
Finnish	fi	293,424	12,009,060	200,609	1,947
Czech	cs	226,545	8,544,830	180,433	2,201
Hungarian	hu	213,940	11,946,177	195,196	1,996
Korean	ko	194,252	9,694,187	182,338	1,938
Indonesian	id	186,820	6,154,782	346,585	1,942
Turkish	tr	183,745	11,671,387	396,736	2,291
Persian	fa	178,493	8,451,247	275,661	2,237
Romanian	ro	176,413	6,560,778	219,432	1,205
Arabic	ar	172,948	9,877,157	484,123	3,681

Table 1: Statistics of some of the most edited Wikipedia by language (http://meta.wikimedia.org/wiki/List_of_Wikipedias accessed on April 3, 2012). Active users performed at least one action in the last 30 days.

of article topics that are present in many different language communities, for example the page about "Osama Bin Laden", according to the interwiki links present in the English Wikipedia article, is present in 116 language Wikipedias and the page about "George W. Bush" in 190 Wikipedias.

An excellent analysis of diversity of knowledge represented in 25 different Wikipedias is presented in [13] and a surprisingly small amount of concept overlap is found between languages of Wikipedia, as over 74 percent of concepts are described in only one language and only 0.12 percent of them are described in all the 25 investigated language Wikipedias. Moreover it has been found that each language Wikipedia exhibits a self-focus bias towards articles about regions where that language is largely spoken [12].

So the empirical question is: on articles that are present in different language Wikipedias and given also the fact that automatic translation of articles is discouraged, do different language communities develop very diverse versions of equivalent articles?

Actually the page "Wikipedia:Neutral point of view/FAQ" at the section "Anglo-American focus" states that "Wikipedia seems to have an Anglo-American focus. Is this contrary to the neutral point of view? Yes, it is, especially when dealing with articles that require an international perspective. The presence of articles written from a United States or European Anglophone perspective is simply a reflection of the fact that there are many U.S. and European Anglophone people working on the project. This is an ongoing problem that should be corrected by active collaboration between Anglo-Americans and people from other countries. But rather than introducing their own cultural bias, they should seek to improve articles by removing any examples of cultural bias that they encounter, or making readers aware of them." And then "this is not only a problem in the English Wikipedia. The French Language Wikipedia may reflect a French bias, the Japanese Wikipedia may reflect a Japanese bias, and so on." This is acknowledged also by Rosenzweig in [23] when he states “but the largest bias—at least in the English-language version—favors Western culture (and English-speaking nations), rather than geek or popular culture.”

We call this phenomenon “Linguistic Point of View” (LPOV). The presence of diverse points of view on different language editions of Wikipedia would disprove the “global consensus hypothesis” which posits that “two articles about the same concept in two different languages will describe that concept roughly identically” [13].

On the page “Wikipedia:Describing points of view”, it is clearly written that “English language Wikipedia articles should be written for an international audience”. Two questions can arise from this aim. The first one is if this is really what is happening in the English Wikipedia and in the other Wikipedias: are they written for an international audience or do they reflect a specific Linguistic Point of View? The second question is about the fact this aim is good for our world or not: are we going towards a globalized knowledge losing specificities and traditions or do we risk to go towards fragmentation of society in language specific communities?

The contribution of this paper focuses on the first question, in order to provide a tool which makes it easier to assess the current situation. Speculations and arguments about which path is better for the world can start as a natural consequence of an informed debate on the current situation.

Note that few studies started to emerge comparing different language Wikipedias. For example, authors of “Cultural Differences in Collaborative Authoring of Wikipedia” compared French, German, Japanese and Dutch Wikipedia [22] while “Cross-cultural analysis of the Wikipedia community” analyzed English, Hebrew, Japanese, and Malay [10]. Arabic, English, and Korean Wikipedias were compared by Stvilia et al. [25]. These analysis were performed with manual content analysis of some article pages from the different Wikipedias. There is also one published paper that focuses on one single Wikipedia, the Chinese one, and compares point of regional differences of its contributors based on four regions of origin (Mainland, Hong Kong / Macau, Taiwan, and Singapore / Malaysia) [15]. Authors claim that the main issue threatening the potential growth of Chinese Wikipedia are not the internal conflicts, nor the external competition by Baidu Baike but the evolution of the newly established “Avoid Region-Centric Policy”. On [19] instead social network analysis is used as a lens for comparing English, German, Japanese, Korean, and Finnish language Wikipedias finding a difference between egalitarian cultures such as the Finnish, and quite hierarchical ones such as the Japanese.

A specific analysis of the comparative cultural biases present in articles about famous persons in English and Polish Wikipedia is presented in [5]. Authors perform quantitative and qualitative content analysis revealing systematic differences related to the different cultures, histories, and values of Poland and United States.

Such studies, when involving manual analysis of the articles, required for the authors knowledge of all the involved languages in order to compare the knowledge products created by the different language communities.

Hecht and Gergle published research that is closer to ours in scope [13,12,2]. In [13] they report how there is a surprisingly small overlap in the concepts present in different language Wikipedias. Moreover, when the same concept exists in two different editions of Wikipedia, they find that the sub-concept diversity, defined as overlap in links to other Wikipedia pages, is lower than expected. In [12], each language edition of Wikipedia is characterized for its level of self-focus bias operationalized as number of links directed at articles located in the region of the world where that language is largely spoken. In both cases, their focus is at the level of characterizing the entire Wikipedia and the main considered element are number of links to other pages and not the text present in the page itself. In a recent paper [2], Bao et al. describe Omnipedia, a tool able to show, for a specific Wikipedia page, which other pages are discussed in only a single language edition of the concept, indicating topics specific to a certain culture and which ones are discussed more broadly in many different language editions of the page. Instead our work aims at providing a Web tool for pairwise comparison at the level of single pages so that anyone with a web browser can investigate the presence (or absence) of different Linguistic Points of View and possibly improve, correct and discuss them. By releasing it as open source, our aim is also to make it easier for other researchers to extend our initial effort.

In the next section we present Manypedia, the web tool that, exploiting automated machine translation, aims at lower the bar for cross-cultural studies and research of different language Wikipedia communities.

4. MANYPEDIA WEB TOOL

Through Manypedia it is possible to compare Linguistic Points of View of different communities of language editions of Wikipedia. Manypedia is accessible at http://www.manypedia.com. Precisely, Manypedia can be used to search for a page title in a specific Wikipedia, for example the English one (left side of Figure 1), and to compare it with the equivalent page from another Wikipedia, for example the Chinese one, with the possibility of translating it into English with one click (right side of Figure 1). In this way, even if automatic translation, powered by Google Translate online service, is not perfect, the requirement of knowing the two languages for cross-language studies is relieved. We believe that being able to "understand" the result of hundreds of edits by Wikipedians who edited a certain page in, for example, Chinese (without knowing Chinese) using a single pairwise web interface is a great opportunity for cross-cultural studies. Every link in Wikipedia articles gets transformed into a new comparison so that navigation can conveniently continue inside Manypedia. Currently 56 languages are supported in translation both as source and target language, ranging from English, Spanish, German to Yiddish, Tagalog, Catalan, Swahili and more.

In Figure 1, there is a screenshot of Manypedia comparing the page “Palestinian territories” from English Wikipedia (left) with the equivalent page, titled "فلسـطين المحتلـة", in the Arabic Wikipedia (translated into English).

On top of embedded Wikipedia pages (both left and right sides), Manypedia shows information which can help in forming a first idea about the differences of the knowledge products created by the two different language communities. First of all, Manypedia finds images included in the two Wikipedia articles and show them also on top of the page in order to get a first visual understanding of the points of view represented. A word cloud of the most frequent words is also presented in order to quickly spot the main textual differences of the two pages. Statistics about the pages are required at runtime via Ajax to out PHP scripts running on toolserver.org, where a copy of Wikipedia databased is made available. Statistics comprise number of total edits received, useful for comparing the attention received by the page from the two language communities, while taking into account that the English Wikipedia community is much larger than, for example, the Japanese Wikipedia community that is much larger than the Swahili one. The number of different editors who contributed to the page is shown as well. In general noting few edits by one or two editors could warn the Manypedia visitor about the possibility the article does not reflect an at-least-partially shared vision but only the points of view of the few involved editors. On the other hand, if the page has received a large number of edits by a large number of editors, it is more plausible to assume that the current page is the up-to-date neutral result of the negotiation of all the significant viewpoints about the issue shared by the specific language Wikipedia community. Creation date and creator are shown as well in order to provide evidence about the fact the page has existed since enough time to get enough attention and diverse points of view. The date of last edit allows to ponder how much the page is settled down or received recent attention by the community. Moreover, signs of vandalism or very biased points of view can be more easily found on pages edited very recently, for which the community didn't have enough time to react and fix the vandalism yet [14,27].

On top of the two pages Manypedia also shows the 5 Wikipedians who edited the page the most, with a link opening additional statistical data about them, along with the number of edits they contributed to the page. This information is useful in order to understand if there is one single user "owning" the page: Wikipedia policy clearly states that "you do not own articles". Moreover it is possible to get an idea of the relative influence exercised by the top editors of this page by comparing their edits and the total number of edits: again, a small percentage might indicate a more shared and neutral point of view. Even more interestingly, there might be cases in which the same Wikipedian is one of the most active editors in both the articles from the two different language Wikipedias. All the statistics shown on top of article pages go in the direction of improving transparency of Wikipedia pages by highlighting some important but not so visible aspects of the process involved in the creation and maintenance of the page by the community. This is similar to what the project Wikidashboard does with the goal of increasing social transparency [4].

An additional automatic instrument for comparing the two pages is the concept similarity percentage. This is computed at runtime based on the sub-concept diversity index introduced in [13]. The concept similarity is computed based on outlinks, or links in one Wikipedia article pointing to another article. The intuition is that “if two articles on the same concept in two languages define the concept in a nearly identical fashion, they should link to articles on nearly all the same concepts. If, on the other hand, there is great sub-concept diversity, these articles would link to very few articles about the same concepts” [13]. The measure is not meaningful for each comparison because many factors are involved in the differences in links to pages such as cultural differences but also differences in linking behaviours (a page might refer to a concept without linking to it while the other one links to it) [13]. Current work is ongoing with the aim of adding additional comparisons at the level of the meaning of each sentence.

Since Wikipedia articles are released under Creative Commons Attribution Share Alike License, anyone, including Manypedia, is allowed to copy, distribute, transmit and also remix the content as long as he or she attributes it to the authors and copyright holders: Manypedia does so by giving credit to the specific Wikipedia articles incorporated in each comparison specifying that the source is Wikipedia and linking to the specific article. As a consequence, the content of Manypedia is released under a Creative Commons Attribution-Share Alike License as well so that anyone, including researchers, can copy, redistribute and remix the content simply by citing Manypedia as a source.

The code powering Manypedia and the scripts running on toolserver.org extracting statistics at runtime for each page and user have been released as open source so that other researchers can build on them and are available at https://github.com/volpino/

5. DISCUSSION ON POSSIBLE USES OF MANYPEDIA AND FUTURE WORK

In this section we briefly highlight possible foreseen uses of Manypedia. We are not experts of cross-cultural studies and carefully conducted investigations in this realm about similarities and dissimilarities on how different communities represent the same concept go over the scope of this paper and are future work.

Manypedia interface provides (on top right, see Figure 1) a list of featured comparisons, as selected by hand by authors, as well as a list of the latest comparisons performed by Manypedia users and of the comparisons most popular in the last 20 days in order to continuously highlight what are the topics more cross-culturally investigated recently. These lists can possibly provide interesting starting points for cross-cultural investigations, considering also that each link present in Wikipedia pages is transformed into a comparison inside Manypedia as well. We plan to also offer an additional list of the comparisons recently performed by users whose concept similarity percentage is smaller than, say, 10% along with a large number of links present in both pages.

An interesting starting point for investigation is the page “List of controversial articles”. These pages, just as every Wikipedia page, can be analyzed using Manypedia. For example the URL http://www.manypedia.com/#|en|List_of_controversial_articles|zh is the comparison the the page “List of controversial articles” from English Wikipedia (en) and Chinese Wikipedia (zh), translated into English[3].

It is possible to observe that the page from English Wikipedia (which groups the many controversial articles into 15 main classes such as Politics/ economics, History, Religion, Science / Biology / Health, Sexuality, Sports, Entertainment, Environment, Law and Order, Linguistics, Philosophy, Psychiatry, Technology, Media/culture, People/ public figures/ infamous persons) is slightly centered around topics important for US and Western culture. On the other hand the Chinese Wikipedia page lists pages such as "Anti-Japanese War", "Nanjing Massacre", "Taiwan", "Human Rights in China", "Falun Gong", "Tiananmen Incident", "Mao Zedong", "List of sites blocked by China". Many of the links contained in both pages will possibly result in an interesting start for a journey on cross-cultural comparisons. The same argument is visible for most language communities, for example the “List of Controversial articles” in the Catalan Wikipedia refers predominantly to issues about the term “country” and “region” and the concept of Catalan country itself.

Another interesting example is the page “Human rights in the United States” whose Chinese counterpart starts with "Most Americans think the U.S. is a free country” and then “U.S. double standards on human rights is hypocritical”.

In general all topics related to recent history can be biased, especially if there are two or more fighting nations involved. We have already reported the article in which George Orwell,

[3] Wikipedia is an ongoing work and we are aware that each page can be changed in any moment. For this reason we saved the HTML page of the comparisons to which we referred in this paper at http://sonetlab.fbk.eu/data/manypedia_saved/

referring to the ongoing 2nd World War, argues that “a Nazi and a non-Nazi version of the present war would have no resemblance to one another, and which of them finally gets into the history books will be decided not by evidential methods but on the battlefield” and in which he reminds that “history is written by the winners” [20].

Surely the interesting part of Wikipedia is that it can be edited and “fixed” in real-time while this is much harder and slower with history books which are taught in schools for example. As Rosenzweig puts it, “like journalism, Wikipedia offers a first draft of history, but unlike journalism's draft, that history is subject to continuous revision. Wikipedia's ease of revision not only makes it more up-to-date than a traditional encyclopedia, it also gives it (like the Web itself) a self-healing quality since defects that are criticized can be quickly remedied and alternative perspectives can be instantly added”. In fact recent work on the formation of collective memories of recent events exploits this feature of Wikipedia for which recent events tend to get created few minutes or hours after it happens and the community strives to fairly represent it as it unfolds [8]. This is especially interesting in the case of traumatic events such as, for example, the recent North African revolutions [7]. With regard to history, Manypedia offers for example a tool for comparing the different representations of the “Vietnam war” between the English Wikipedia and the Vietnamese one, or to get an understanding of the reception of “Abu Ghraib torture and prisoner abuse” by different language communities in Wikipedia.

Ongoing struggles for disputed states might also be represented in diverse ways, especially by the language communities which are more closely involved in the issue. We have already reported about Catalonia in Catalan language but similar arguments can be made for Galicia in Galician Wikipedia, Taiwan and Tibet in Chinese Wikipedia. Northern Cyprus is an especially interesting comparison where the Greek Wikipedia reports it “is under Turkish occupation since 1974 in violation of international legal norms” while the Turkish Wikipedia states it “is an independent state”. The community of editors on English Wikipedia is possibly less involved and more neutral and claim Northern Cyprus “is a de facto independent state (...). Tensions between the Greek Cypriot and Turkish Cypriot populations culminated in 1974 with a coup d'état, an attempt to annex the island to Greece and a military invasion by Turkey in response. (…) Northern Cyprus has received diplomatic recognition only from Turkey.”

A paradigmatic example with this regard is the ongoing conflict between Israeli and Palestinians that can be analyzed in terms of Linguistic Point of View on pages such as “Palestine”, “Israel”, “Israeli–Palestinian conflict” and the dozens of other pages in the “Category:Israeli–Palestinian conflict” by comparing, for example, the Arabic and Hebrew Wikipedia representations. The page “Jerusalem” is a related example which is possibly even more controversial since it involves also issues related to religion. Religion is surely a topic on which it can be harder to remain neutral involving faith and basic believes: pages interesting with this regard are for example “Crusades”, “Islamofascism”, “Poligamy”.

Moreover some knowledge areas might be more or less treated in relative terms by different language communities and hence reveal an imbalanced coverage. For example the English Wikipedia has an impressive coverage of topics related to sexuality both with regard to extreme practices and sexual orientation and, thanks to Manypedia, it is possible to check if other Wikipedias such as the Arabic or Japanese ones exhibit different coverage in relative terms and by number of edits and editors involved.

The feature of grouping all images of a Wikipedia page on top of it can be particularly useful with this regard because it can be easier to just spot how many and more importantly which images are used to represent a specific concept. This can be done on generic pages such as “1970 year” or “Black people” and also on sex-related pages. Just as an intriguing example of this, we report that in 2010, Larry Sanger, cofounder of Wikipedia in 2001, reported the Wikimedia Foundation to the FBI for "knowingly distributing child pornography". The suspect material were 27 images in the "Pedophilia" and "Lolicon" categories on Wikimedia Commons [17]. This testifies that each language community is probably faced and must reach consensus between representation and self-censorship of sensitive topics.

The last examples we report here are the pages “Recent deaths” and “Portal:current events”. By comparing them across different language Wikipedias it is possible to quickly appreciate which are the people whose death is encyclopedia worthy for the editors of a specific languages and which events are important enough to be reported in the portal. Is it possible to write these pages with an international audience in mind? It is reasonable to ask to the different language communities of Wikipedia to do it? We will address these questions and the broad implications of Manypedia as a tool for investigating Linguistic Points of View in the next section.

In this section we just reported few examples of comparisons which can act as starting points for cross-cultural investigations made possible by Manypedia. Our future work involves developing automatic ways for highlighting differences at the sentence level and conducting case studies with cross-cultural researchers in order to empirically validate the utility of Manypedia.

6. CONCLUSIONS

In this paper we have presented Manypedia, a web mashup which allows to compare the same page on two different language Wikipedias. Manypedia exploits automatic machine translation and hence does not require knowledge of the second language for the comparison. Moreover, the summarization provided through images, most frequent words and statistics of the creation process of the Wikipedia page allows to complement the investigation about the differences (if any) in representation of the same concept by the two different language Wikipedia communities of thousands of editors.

As Wikipedia itself states there are systemic biases in its process which naturally grows from the characteristics of people who self-elect for writing its millions of articles. However, in this paper, we are not interested in biases intra-specific Wikipedia but on differences in inter-Wikipedia representations: are there Linguistic Points of View in the different language editions of Wikipedia?

As we have seen, Manypedia is a tool which allows to answer this question and makes easier to conduct cross-cultural studies on Wikipedia. Moreover Manypedia can be used to maintain balanced, coherent and convergent points of view across different language Wikipedias since the current Wikipedia socio-technical platform does not provide many opportunities for editors of different language Wikipedias to discuss and share points of view.

As a consequence of Manypedia allowing to assess the current situation in terms of the magnitude of Linguistic Points of View, it is hence possible to enter into more philosophical questions and speculate on the fact writing from an internationally neutral point of view is possible in every language Wikipedia and, more interestingly, if this is desirable for the future of our world.

Do we risk of going towards a globalized knowledge losing specificities and traditions of local cultures or do we risk to go towards fragmentation of world society in language specific communities? In the first case the model is the tyranny of majority in which minority views and diversity get not represented and only few major points of view survive [16]. On the other hand of the spectrum of possibilities, there are so-called echo chambers [7]: different communities (identified by the language they speak, or by the founding point of view they chose as in the examples of Ecured, Anarchopedia and Conservapedia) develop their own representations of facts and these representation become more and more biased and diverge in such a way that fragmentation of society and in-communicability among groups is reached as Sunstein warns against in his book Republic.com [26].

Which extreme shall the Wikipedia platform encourages, tyranny of the majority or echo chambers [16]? Or what is the best balance among them? Our aim with Manypedia is to help starting a global informed debate about these important issues.

7. ACKNOWLEDGMENTS

Manypedia could not have been possible without Wikipedia, its community of volunteers and the Creative Commons license which encourages remix of content. We also thanks Google for providing a real-time language translation API and the developers of open source software we used for developing Manypedia, in particular the Javascript library jQuery and its dozens of plugins.

8. REFERENCES

[1] Alexa ranking. http://www.alexa.com/siteinfo/wikipedia.org Retrieved on April 4, 2012.

[2] Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M. and Gergle, D. 2012. Omnipedia: Bridging the Wikipedia Language Gap. ACM Conference on Human Factors in Computing Systems (CHI 2012). New York: ACM Press.

[3] BBC. 2010. Cuba launches online encyclopaedia similar to Wikipedia. 14 December 2010.

[4] Bongwon S., Chi Ed H., Kittur A., and Pendleton B. A. 2008. Lifting the veil: improving accountability and social transparency in Wikipedia with wikidashboard. ACM Conference on Human Factors in Computing Systems (CHI 2008). New York: ACM Press.

[5] Callahan, E., and Herring, S. C. 2011. Cultural bias in Wikipedia articles about famous persons. Journal of the American Society for Information Science and Technology.

[6] Chandy R.. 2008. Wikiganda: Identifying Propaganda Through Text Analysis. Caltech Undergraduate Research Journal. Winter 2008-2009

[7] Ferron, M., and Massa, P. 2011. Collective memory building in Wikipedia: the case of North African revolutions. Wikisym 2011.

[8] Ferron, M., and Massa, P. 2011. Studying Collective Memories in Wikipedia. 3rd Digital Memories Conference. Prague, March 2011.

[9] Giles J. 2005. Internet encyclopaedias go head to head. Nature; 438: 900-901.

[10] Hara, N., Shachaf, P., & Hew, K.F. 2010. Cross-cultural analysis of the Wikipedia community. Journal of the American Society for Information Science and Technology, 61(10), 2097–2108.

[11] Head A. J. and Eisenberg M. B. 2010. How today's college students use Wikipedia for course–related research. First Monday, Volume 15, Number 3. 1 March 2010.

[12] Hecht, B., & Gergle, D. 2009. Measuring Self-Focus Bias in Community-Maintained Knowledge Repositories. Proceedings of Communities and Technologies 2009 (C&T 2009), pp. 11-20. New York: ACM Press.

[13] Hecht, B., & Gergle, D. 2010. The Tower of Babel Meets Web 2.0: User-Generated Content and its Applications in a Multilingual Context. ACM Conference on Human Factors in Computing Systems (CHI 2010). New York: ACM Press.

[14] Kittur A., Suh B., Pendleton B. A., Chi E. H. 2007. He says, she says: Conflict and coordination in Wikipedia. ACM Conference on Human Factors in Computing Systems (CHI 2007). New York: ACM Press.

[15] Liao, H. 2009. Conflictual Consensus in the Chinese Version of Wikipedia. IEEE Technology and Society Magazine.

[16] Massa P. and Avesani P. 2007. Trust Metrics on Controversial Users: Balancing Between Tyranny of the Majority. International Journal on Semantic Web and Information Systems (IJSWIS), 3(1), 39-64.

[17] Metz C. 2010. Wikifounder reports Wikiparent to FBI over 'child porn'. The Register. Retrieved on April 4, 2012, from http://www.theregister.co.uk/2010/04/09/sanger_reports_wikimedia_to_the_fbi/

[18] Nagel T. 1986. The View from Nowhere. Oxford University Press.

[19] Nemoto, K. Gloor, P. 2010. Analyzing Cultural Differences in Collaborative Innovation Networks by Analyzing Editing Behavior in Different-Language Wikipedias. Proceedings COINs 2010, Collaborative Innovations Networks Conference, Savannah GA, Oct 7-9, 2010

[20] Orwell G. 1943. As I Please. Tribune. GB, London.

[21] Pew Research Center. 2011. Wikipedia, past and present. A snapshot of current Wikipedia users.. Jan 13, 2011. Retrieved on August 4, 2011, from http://www.pewinternet.org/Reports/2011/Wikipedia.aspx

[22] Pfeil, U., Zaphiris, P. and Ang, C. S. 2006. Cultural Differences in Collaborative Authoring of Wikipedia. Journal of Computer-Mediated Communication, 12, 88–113.

[23] Rosenzweig R. Can History Be Open Source? Wikipedia and the Future of the Past. 2006. The Journal of American History 93(1): 117-146.

[24] Sanger L. M. 2009. The Fate of Expertise after Wikipedia. Episteme. Volume 6, Page 52-73

[25] Stvilia, B., Al-Faraj, A., & Yi, Y. 2009. Issues of cross-contextual information quality evaluation—The case of Arabic, English, and Korean Wikipedias. Library & Information Science Research, 31(4), 232-239.

[26] Sunstein C. R. 2001. Republic.com. Princeton University Press, Princeton, NJ, USA.

[27] Viegas F. B., Wattenberg M. and Dave K. 2004. Studying cooperation and conflict between authors with history flow visualizations. ACM Conference on Human Factors in Computing Systems (CHI 2004). New York: ACM Press.