Cultural Differences in the Understanding of History on Wikipedia

Robin Giecka, Peter Gloor, Hanna-Mari Kinnunenb, Yuanyuan Lic, Mohsen Moghaddamc, Maria Paasivaara, Franziska Pradel, Matthäus Zylka

Abstract This paper sheds light on cultural differences in the understanding of historical military events among Chinese, English, French, German and Swedish Wikipedia language editions. This is due to the fact that differences in understanding can lead to intercultural misinterpretation and conflicts. We identified the most important historical events, mined cross-cultural relations and investigated word usage in war related pages, performed a network analysis, as well as complexity- and sentiment analysis. We also analyzed the usage of war-related words and the quantity of war events mentioned in different languages. Our findings suggest that World War I and World War II are the most important historical events among English, French and German cultures. English Wikipedia has more violence- and war-related content as well in addition to a higher complexity compared to other editions.

1 Introduction and Related Work

In August 2015, Wikipedia has 280 (active) different language editions and therefore plays a very important role in communicating between different languages and cultures. Previous research validated Wikipedia as a great data provider for resolving different research questions (Medelyan et al. 2009; Schroeder/Taylor 2015; Xu/Li 2015: 275 ff.). However, most studies about differences between various language editions focused on the interaction and communication between editors and only few papers aimed to investigate the cultural similarities and dissimilarities among different language editions of Wikipedia.

The MIT Media Lab developed the project Pantheon that used Wikipedia data and information of Murray's "Human Accomplishment" (2003) to map cultural production (Yu et al. 2015). Cultural differences in Wikipedia have been

examined by analyzing editing behaviour in various Wikipedia editions (e.g., Nemoto and Gloor 2011) or by investigating the description of cultural practices in Wikipedia like food cultures (Laufer et al. 2015). Other studies focused on important historical persons by using network analysis (Aragon et al. 2012; Eom/Shepelyansky 2013; Gloor et al. 2015), quantitative and qualitative content analysis (e.g., Callahan/Herring 2011) or different kinds of discussion spaces (Hara et al. 2010). Furthermore, some studies focused on articles in different languages to explore controversial topics and showed the emergence of different preferences and interests in Wikipedia (Yasseri et al. 2014; Bicli/Bulian 2014).

Our paper aims to study the cultural similarities and differences between Chinese, English, German, French and Swedish Wikipedia by focusing on articles related to the most important and influential historical war events of these five languages in Wikipedia. We use network, content, sentiment and complexity analysis to investigate the cultural differences. Since Wikipedia articles are written in different languages, we assume that the editors of the articles are influenced by their own culture. Therefore, the different understandings of historical events are unavoidable. According to Laufer et al. (2015), we use Wikipedia language as the proxy for cultural communities. Defining transmitted information as culture is a common practice among scholars (Yu et al. 2015). It is important to investigate cultural similarities and dissimilarities because different understandings of historical events can lead to conflicts and confusion.

2 Methods

We used articles of historical war-related events in Chinese, English, French, German and Swedish Wikipedia, which are categorized as military historical events to investigate cultural differences and similarities between different languages. As Wikipedia does not contain a universal categorization for wars, we had to find another way for extracting the most important war events. Fortunately, many countries have their own article in Wikipedia that contains a list of wars in which the respective country has taken part in. In this research the historical war event is a war that one of the listed countries has participated.

In order to get the most important war-related events, we created a Java program that fetched a list of war-related events from Wikipedia, counted the indegree of each event (number of incoming links) and ordered events by their popularity. We considered the measured indegree as a key figure of importance, which is in line with Charles Murray's work (2003) who used number of mentioning as determination of importance. We used Wikipedia pages containing a list of wars involving the respective country for mining war events. Those pages existed both in English and the original language (Chinese, German, Swedish and French). The way of data extraction enabled us to gather information of 93 English, 28 Swedish, 253 German, 201 Chinese, (69 Finnish), 104 French war articles.

3 Results

Mining cross-cultural relations: We used the Jaccard similarity coefficient to measure the similarities among the 10, 20 and 50 most important historical events (accordingly to Laufer et al. 2015). Here, we have used this measure to compare the size of intersection of the same historical wars divided by the size of potentially intersection (sample union), for example J between Chinese and English culture:

$J=(Chinese \cap English)/(Chinese \cup English)$

Table 1 shows the similarity among top 20 most important events between the different languages of Wikipedia. For example, English and Chinese Wikipedia have 15 % of the 20 most important events in common (equivalent to three of the top 20 wars). From this result we can conclude that the French and German culture, the German and English culture as well as the French and English culture are most similar. Chinese and Swedish cultures have the least in common with the other cultures.

Table 1. Similarities among the 20 most important historical wars

	Chinese	English	German	French	Swedish
Chinese	1	0.15	0.15	0.1	0
English		1	0.35	0.5	0
German			1	0.45	0.1
French				1	0.1
Swedish					1

War like Words Analysis: In order to investigate the popularity of war-related words, we performed an analysis by conducting a search by Google to determine usage frequency of different war and violence related words. Words like “kill”, “war”, “battle”, “murder”, “victory”, “defeat” and “revolution” were translated and used in the target language. The number of hits was divided by the size of the relevant language edition (Wikipedia 2014) to obtain comparable results.

Results listed in table 2 show that English Wikipedia contains a significant amount of mentioned words such as “war” and “battle” compared to other languages. “Revolution” was named most in French Wikipedia. The ratio between victory and defeat shows how the different Wikipedias focus on victories in comparison to defeats. Based on this ratio, French Wikipedia clearly used the most positive and Chinese the most critical language. All five Wikipedia languages focus more on victories than defeats.

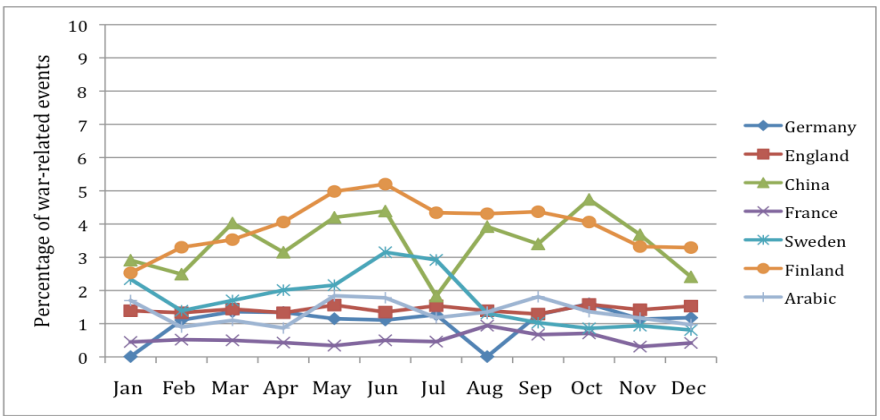
Table 2. Wikipedia war-like word analysis

Word	English	German	French	Swedish	Finnish	Chinese	Arabic
Kill	0.0572	0.0236	0.0210	0.000847	0.0540	0.0510	0.575
War	190	0.0722	0.196	0.0103	0.0326	0.0853	0.148
Battle	149	0.0312	0.0686	0.00661	0.0534	0.0339	0.0719
Murder	0.0548	0.0149	0.0195	0.00510	0.0195	0.0153	0.0575
Victory	0.0650	0.0417	0.0644	0.00303	0.0497	0.0370	0.0485
Defeat	0.0390	0.0215	0.0139	0.00109	0.0200	0.0300	0.0161
Revolution	0.0539	0.0257	0.0805	0.00259	0.0145	0.0492	0.0593
Victory/Defeat	1.67	1.94	4.63	2.78	2.49	1.23	3.01

Wikipedia Date Page Analysis: As a separate Wikipedia analysis method we created a Java program that used the Wikimedia API for searching and analyzing the amount of war events in date pages. All our target languages have a page that contains a list of events that have happened on that day during history. We wanted to analyze how many of the events that are listed in those date pages are war-related compared to the total amount of events.

Figure 1 contains the results of a date page analysis divided into monthly values. It seems that Finns and Chinese focus most on war-related events in Wikipedia date pages. French mentions least war events compared to other countries. Furthermore, there are less war events during wintertime than in summer or autumn. This emphasizes that most of the war events seem to have happened between June and October.

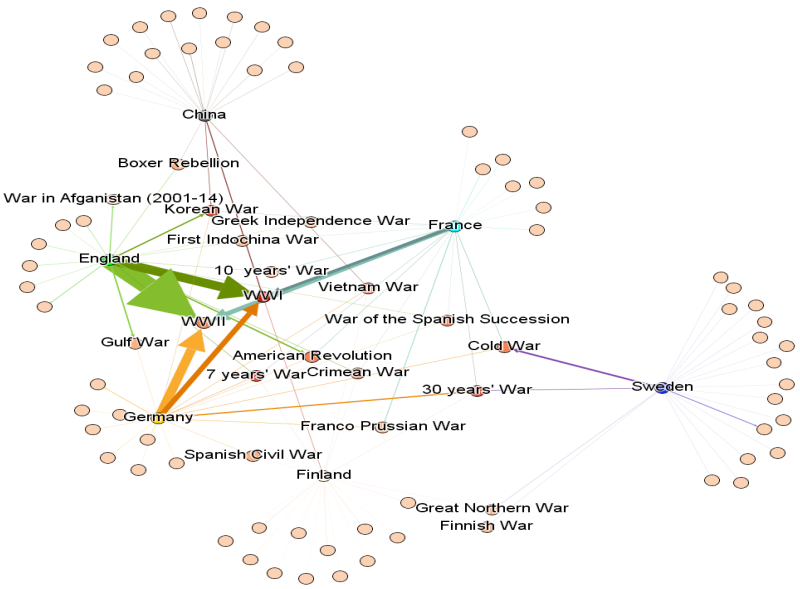
Fig. 1. Wikipedia date page analysis



Network of War Events: Based on previous results, we have been able to create a "Network of War Events" for the previously named nations plus Finland (see Figure 2). For creation and analysis of the network we used the open source tool Gephi (Bastian, Heymann, & Jacomy, 2009). We have developed a model of

different types of nodes and weighted edges in a directed network. As node types we defined two categories. The first category represents the considered nations (in this case China, England, Finland, France, Germany, Sweden), because it is a directed network each “nation-node” has an out-degree-value of 20 and an in-degree-value of 0. For a better illustration, nodes with the type “nation” are shown in different colours. While nodes of the second category (type “events”) are displayed in a gradient from light to dark red, depending on the indegree, the importance (weighted indegree) or number of incoming links is represented by an increasing weighting and thickness of the edges. A detailed viewing of the graph shows that World War I, with the highest In-Degree (5) and Centrality (1), is the most frequently mentioned event among our analyzed nations while World War II is the most important event by considering the number of incoming links. In particular, the English and German Wikipedia contribute a large amount of backlinks to World War II, but this is also related to the size of the Wikipedia Language versions.

Fig. 2 Network of war events

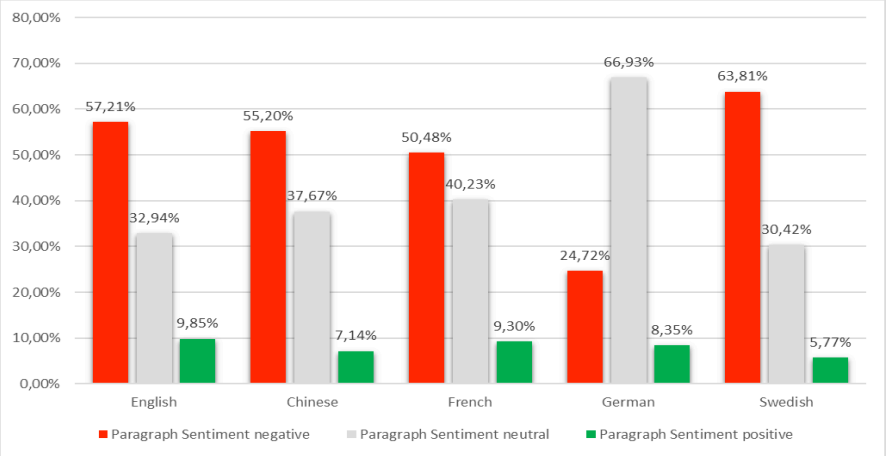


Sentiment Analysis: Sentiment Analysis focuses on the analysis of people’s sentiments, opinions, attitudes and emotions towards elements like themes, individuals and organizations (Serrano-Guerrero et al. 2015). We used “Semantria for Excel” to analyze the cultural differences among Chinese-, English-, German-, French- and Swedish Wikipedia. Semantria is a multilingual sentiment engine, which masters several languages such as English, German, Chinese or French and

weighted each paragraph of a document based on the document's components (themes, topics, entities) and their sentiment values (Semantria 2012).

Figure 3 shows the proportion of negative, neutral and positive paragraphs for each language. Considering the whole content of the viewed events for each language, the Swedish Wikipedia comprises the largest proportion of negative paragraphs with 63.81%, followed by English (57.21%), Chinese (55.20%) and French (50.48%). The German Wikipedia consists of neutral evaluations to a large extent (66.93%). A similarity between all five languages is the low percentage of positively evaluated paragraphs. Considering the Event's category, these results are as expected.

Fig. 3. Proportional view of the sentiment analysis for each language



Language Complexity Analysis: We tried to determine readability, the ease with which a reader can understand a written text (Dale/Chall 1949). Gunning Fog Index (DuBay 2004) has been used to analyze the top 20 war events' content for English, French, German and Swedish Wikipedia. Because of the completely different syllable structure of Chinese compared to European languages, Yang's Index specially developed for Chinese has been used for Chinese Wikipedia (Sung et al. 2013). Due to the different interpretation of output value for these two indexes, we translated the outputs into their individual equivalent reading level to compare war-related articles written in Chinese and other languages. The most complex article of all languages is the one about Cold War in the English Wikipedia with the highest Gunning Fog of 32.02, while the most "comfortable" Gunning Fox index 8. The average reading level needed for Chinese war-related articles is only of a 16 years old, which means every high school student should perceive these articles as relative readable. Meanwhile most English and Swedish high school students might perceive the language of their Wikipedia as more complex, as their average reading level requires an age over 25 years.

4 Discussion and Conclusion

In this paper, we analyzed how war-related articles are written in different language editions of Wikipedia, to determine the impact of Wikipedia among the speakers of each language. Several analyses showed differences among Wikipedia articles in selected languages. War-related word analysis has shown that English Wikipedia has the biggest amount of war-related words. This result suggests that English Wikipedia has more violence- and war-related content compared to other language editions. The Swedish and Chinese Wikipedias focused more on war-related events than other languages according to Data Page Analysis. High complexity like English and Swedish Wikipedia and therefore a low readability for English and Swedish speakers could cause a long-term effect and lead to a loss of reader groups. It could lead to a decreasing influence of these two language editions for all ages among English and Swedish speakers in the future. The most similar Wikipedia language versions based on Network Analysis are English, German and French, which have similar top events among their top 20s. Our findings suggest that World War I and World War II are the most important historical events among these cultures.

The presented work can be improved in several ways. The extraction of historical events can be improved by finding a more systematical way of data extraction for example by using key words in the title (e.g., “War”, “Revolution”, “Battle”). Furthermore, it would be interesting to analyze more language versions and cluster them by using the gathered information. According to Murray (2003), we used the number of incoming links (indegree) as measure for the importance of an event, but other measures could also be used and compared with each other (e.g., number of edits or views). Another supplement could be to add a variable, which weights the size of each Wikipedia language-version for transnational contemplation. The event network could be improved by using an own tool that visualizes the gathered data and automatically combines it with more information like time of the event or involved people. In addition, adding more events and thus more nodes could show a higher connectedness between different Wikipedia language versions. Due to the different standard of output value in the complexity analysis, a comparison of the readability between Chinese and the other four languages is not accurate enough in this work. A generally recognized index for both Asian and European languages should be developed for future work to gain more accurate results. Due to the missing sentiment scores for the Swedish language, we had to perform the sentiment index analysis without it. The development of a comprehensive tool would help to enhance the analysis without restrictions.

Despite these limitations our analyses show that Wikipedia reflects cultural differences. We hope they will spur additional research in other settings using alternative ways of data extraction and analyses of cultural dissimilarities. While Wikipedia is increasingly significant for our every-day life, our data show the importance of further investigation of cultural dissimilarities.

5 References

Aragon, P., Laniado, D., Kaltenbrunner, A., & Volkovich, Y. (2011) Biographical Social Networks on Wikipedia A cross-cultural study of links that made history. In WikiSym '12, 3–6

Bastian M., Heymann S., Jacomy M. (2009) Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media

Bilic, P., & Bulian, L. (2014) Lost in translation: Contexts, computing, disputing on Wikipedia. iConference 2014 Proceedings

Callahan, E. S., & Herring, S. C. (2011) Cultural bias in Wikipedia content on famous persons. Journal of the American society for information science and technology, 62(10), 1899-1915

Dale, E., & Chall, J. S. (1949) Techniques for Selecting and Writing Readable Materials. Elementary English, 26(5), 250-258

DuBay, W. H. (2004) The Principles of Readability. Online Submission

Eom, Y. H., & Shepelyansky, D. L. (2013). Highlighting entanglement of cultures via ranking of multilingual Wikipedia articles. PloS one, 8(10), e74554.

Gloor, P. A., Marcos, J., de Boer, P. M., Fuehres, H., Lo, W., & Nemoto, K. (2015). Cultural Anthropology through the Lens of Wikipedia: Historical Leader Networks, Gender Bias, and News-based Sentiment. arXiv preprint arXiv:1508.00055.

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

Laufer, P., Wagner, C., Flöck, F., & Strohmaier, M. (2015) Mining cross-cultural relations from Wikipedia-A study of 31 European food cultures. arXiv preprint arXiv:1411.4484

Medelyan, O., Milne, D., Legg, C., & Witten, I. H. (2009) Mining meaning from Wikipedia. International Journal of Human-Computer Studies, 67(9), 716-754

Murray, C. (2003) Human accomplishment: The pursuit of excellence in the arts and sciences, 800 BC to 1950. Harper Collins

Nemoto, K., & Gloor, P. A. (2011) Analyzing cultural differences in collaborative innovation networks by analyzing editing behavior in different-language Wikipedias. Procedia-Social and Behavioral Sciences, 26, 180-190

Schroeder, R., & Taylor, L. (2015) Big data and Wikipedia research: social science knowledge across disciplinary divides. Information, Communication & Society, 1-18

Semantria (2012) Frequently Asked Questions, API version 2.0

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015) Sentiment analysis: A review and comparative analysis of web services. Information Sciences, 311, 18-38

Sung, H.T., Chen,R.L., Lee,Y.X., Cha,R.S, Tseng,H.Q., Lin,W.J., Chang,D.X., Chang, G.E. (2013) Investigating Chinese Text Readability Linguistic Features, Modeling, and Validation. Chinese Journal of Psychology, 55(1), 75-106

Wikipedia. Suomenkielisen Wikipedian koko (2014) Url: https://fi.wikipedia.org/wiki/Wikipedia:Tilastot (Read 11/2015)

Yasseri, T., Spoerri, A., Graham, M., & Kertész, J. (2014) The most controversial topics in Wikipedia: A multilingual and geographical analysis

Xu, B., & Li, D. (2015 An empirical study of the motivations for content contribution and community participation in Wikipedia. Information & Management, 52(3), 275-286

Yu, A. Z., Ronen, S., Hu, K., Lu, T., & Hidalgo, C. A. (2015) Pantheon: A Dataset for the Study of Global Cultural Production. arXiv preprint arXiv:1502.07310