(In press, 2011). *Journal of the American Society for Information Science and Technology*.

Cultural Bias in Wikipedia Content on Famous Persons

Ewa Callahan
Quinnipiac University

Susan C. Herring
Indiana University, Bloomington

Abstract

Wikipedia advocates a strict 'neutral point of view' (NPOV) policy. However, although originally a U.S-based, English-language phenomenon, the online, user-created encyclopedia now has versions in many languages. This study examines the extent to which content and perspectives vary across cultures by comparing articles about famous persons in the Polish and English editions of Wikipedia. The results of quantitative and qualitative content analyses reveal systematic differences related to the different cultures, histories, and values of Poland and the United States; at the same time, a U.S./English-language bias is evident throughout. In conclusion, the implications of these findings for the quality and objectivity of Wikipedia as a global repository of knowledge are discussed, and recommendations are advanced for Wikipedia end-users and content developers.

Introduction

The user-generated, global online encyclopedia Wikipedia espouses a neutral point of view (NPOV) policy, according to which every entry should "represent[…] fairly, proportionately, and as far as possible without bias, all significant views" on the topic covered (Wikipedia, 2011a). In this respect, as Lih (2004, n.p.) notes, Wikipedia "has implicitly adopted the same types of operational policies facing modern news operations—sticking to the facts, attributing sources and maintaining balance." The quality of Wikipedia content has been examined in various studies, including through comparison with print encyclopedias, which it has been claimed to resemble in scientific accuracy (Giles, 2005) and formal language and tone (Emigh & Herring, 2005). However, less research has addressed the extent to which Wikipedia coverage is fair and balanced.

This issue comes to the fore when one considers that Wikipedia, although originally a U.S-based, English-language phenomenon, now has versions—or "editions," as they are called on Wikipedia—in many languages, with content and perspectives that can be expected to vary across cultures. With regard to coverage of persons in different language versions, Kolbitsch and Maurer (2006) claim that Wikipedia "emphasises 'local heroes'" and thus "distorts reality and creates an imbalance" (p. 196). However, their evidence is anecdotal; empirical research is needed to investigate the question of whether—and if so, to what extent—the cultural biases of a country are reflected in the content of Wikipedia entries written in the language of that country.

While answering this question fully would require a large-scale multilingual and multinational sample, the present study contributes to this goal by presenting a comparative analysis of two language editions situated in distinctive cultural contexts: English and Polish. More precisely, we ask: Are Wikipedia entries on famous persons different in English and Polish, and if so, how? How neutral and balanced is the coverage of entries in each language with regard to the disclosure and/or omission of controversial information and overall tone? The results of structural and thematic content analysis of 60 entries reveal quantitative and qualitative differences in entries in the different language versions related to the different histories, cultures, and values of Poland and the United States. Only limited evidence emerges in support of Kolbitsch and Maurer's (2006) claim that Wikipedia entries favor "local heroes;" rather, a US/English language advantage is evident throughout. Overall, the findings suggest that monolingual Polish and English readers would get different amounts and kinds of information about famous people through Wikipedia, and that both versions incorporate cultural biases to some extent. In concluding, implications of these findings for the quality and objectivity of Wikipedia as a global repository of knowledge are discussed, and recommendations are advanced for Wikipedia end-users and content developers.

Literature Review

Along with the growing popularity of Wikipedia, there has been a growth in studies that have addressed issues relating to the site's coverage, quality, editorial processes, and neutrality. A number of early studies concentrated on the possible risks arising from the democratic nature of the wiki platform, emphasizing the importance of quality control. Kolbitsch and Maurer (2006) dubbed Wikipedia "the people's encyclopedia," opining that "the main argument against the Wikipedia project is that with an open editing process the correctness of the information provided cannot be guaranteed" (p. 195). Denning et al. (2005) identified six possible risks: lack of accuracy, unknown motives, uncertain expertise, volatility, unconfirmed or lacking sources, and selective coverage biased by the specific interests of the contributors. The imbalance in coverage resulting from the specific interests and knowledge of a self-selected group of contributors means that Wikipedia is more oriented toward current events than historical knowledge.

This imbalance in coverage was empirically confirmed by Halavais and Lackaff (2008), who examined 3,000 random articles and concluded that Wikipedia coverage is good in some sciences and popular culture, but is more limited in the humanities, social sciences, medicine, and law. In addition, other topics are often found in Wikipedia that traditionally are not part of printed encyclopedias, such as colloquial expressions (terms like *fuck* or *oggy oggy oggy*) and unusual terminology (e.g., *folk metal*) (Lih, 2004). Kolbitsch and Maurer (2006) compare Wikipedia to a set of specialized encyclopedias, noting further that Wikipedia articles are often much longer and contain more details than corresponding articles in printed reference sources. This makes Wikipedia a new type of encyclopedia that is not comparable to traditional encyclopedias and dictionaries. Lih (2004) also asserts that Wikipedia is qualitatively different, in that it provides more current and more frequently-updated information. Despite these

differences, the language of Wikipedia articles has been found to be similar to entries in print encyclopedias in its structure and level of formality (Emigh & Herring, 2005).

The issue of the accuracy of Wikipedia content attracted widespread attention when the journal *Nature* published an article by Giles (2005) that claimed that Wikipedia was nearly as accurate as *Encyclopedia Britannica* for scientific articles. Giles's research found that among 42 entries tested, the average entry in Wikipedia contained about four errors versus about three errors in Britannica; moreover, in both sources only eight errors (four in each) were significant. In a further empirical study, Chesney (2006) found that 13% of the articles in Wikipedia examined by experts in their field contained mistakes.

Sanger (2004), one of the co-founders of Wikipedia, believes that more than its actual (in)accuracy, public perception of the site's lack of credibility is a serious problem, and that this problem is created by the low number of experts among its contributors. According to Sanger, a lack of respect for expertise, anti-elitism, and tolerance for vandalism are the major reasons why experts are reluctant to contribute. Other possible reasons include lack of proper recognition for their contributions and *edit wars* with non-experts (Lipsch, 2009). At the same time, the difficulty of assessing the authorship of Wikipedia articles raises questions about the composition of the active editorial community. Automated analysis of all Wikipedia activity from 2001 to 2006 shows a trend over time according to which a core of active ("elite") users contributes a higher proportion of edits as compared to casual users, although at a certain point this pattern levels off (Kittur et al., 2007). According to Ortega et al. (2008), fewer than 10% of the authors are responsible for 90% of the total number of contributions—a finding that holds across multiple linguistic editions of Wikipedia.

Alongside issues of accuracy and credibility is concern about potential bias introduced by the opinions of editors. To minimize possible bias, adherence to the site's NPOV (Neutral Point of View) policy is advised for all contributors. A Wikipedia article currently describes in detail how the NPOV policy, which was first formulated in December 2001, should be understood by authors and readers (Wikipedia, 2011a). Encyclopedic accounts should be presented as fact and not opinions. However, since different viewpoints cannot be avoided, it is also important to present an *editorially neutral* point of view, which means that all majority and significant minority views should be presented in a disinterested tone, without suggesting that one of them is more correct. In this respect, Wikipedia has implicitly adopted the norms of modern news organizations (Lih, 2004). Wikipedia's guidelines note further that even if an article is written with an emphasis on fact rather than opinion, the selection of the facts, their organization, and their presentation can result in a biased article. Thus articles should represent all significant views *fairly* and *proportionately,* where proportionality of views means that opinions should be included according to popularity. Specifically, the guidelines state that "articles should not give minority views as much of or as detailed a description as more widely held views. Generally, the views of tiny minorities should not be included at all" (Wikipedia, 2011a, n.p.).

Reagle (2005), in his essay *Is the Wikipedia Neutral?,* contrasts the term NPOV with the concept of *unbiased* content. Since Wikipedia is the product of thousands of contributors with conflicting viewpoints, Reagle acknowledges that some bias is unavoidable. However, he suggests that with the increasing number of international contributors and linguistic versions, criticism that Wikipedia is America-centric should lessen as the encyclopedia strives to be international and addresses problems like spellings and place names.

Other scholars view the increasing number of linguistic versions as challenging whether a single NPOV is possible in a multilingual environment. As noted by Kolbitsch and Maurer (2006), "even if an article is written in compliance with the "neutral point of view[,]" the varying cultural, social, national and lingual backgrounds can have an enormous influence. Hence, content in Wikipedia can only be as professional and balanced as it authors and their demography are" (p. 196). Similarly, Hecht and Gergel (2010) challenge the notion of a *global consensus hypothesis* that assumes encyclopedic knowledge is consistent across cultures and languages. The results of their study of 25 different Wikipedia versions suggest that common encyclopedic knowledge accounts only for one-tenth of one percent of content. In their study, 74% of articles were described in one language only, while 95% appeared in fewer than seven languages. Moreover, even if the same article existed in two languages, the two versions of the article frequently were not linked together. In another comparative study, Oh et al. (2008) found that only about 7% of the articles in the English language Wikipedia were linked to Japanese articles, while 38% of the Japanese articles were linked to the English articles—a reflection, perhaps, of the different sizes of the two Wikipedia editions.

These results problematize the NPOV policy, considered from an international perspective. It seems that articles on the same topic may present different views in different linguistic editions, although the articles in a specific linguistic version should adhere to NPOV, consistent with the majority opinions of that cultural/linguistic region. The description of the NPOV policy also varies somewhat across linguistic versions of Wikipedia. The NPOV article in the Polish language edition, for example, describes in some detail how to avoid *Polonocentrizm*,[1] acknowledging that it is difficult for authors to separate themselves from a world view characteristic of Poles and Western civilization in general. The article suggests ways to eliminate bias; for example, contributors should not use pronouns like "our," but instead should substitute the adjective "Polish." Likewise, contributors are advised to separate Polish-specific issues from general or international content (example: *Ministry of Finance* and *Ministry of Finance of Polish Government*). However, the article acknowledges that the best way to minimize the Polish perspective would be to entice foreigners, naturalized Polish citizens, and Polish emigrants to contribute to the Polish edition of Wikipedia (Wikipedia, 2011b).

The NPOV relates to the content of Wikipedia articles; another concept, Notability, provides guidance as to which topics are worthy of inclusion. Notability, in the English Wikipedia, is determined by a set of general guidelines, specific guidelines for different categories of articles, Wikipedia participant discussions, and voting. The

general guidelines state that a topic must have "received significant coverage in reliable sources that are independent of the subject," specifying further that sources should be secondary, multiple, objective, and subject to "editorial integrity" to allow verification of the notability of the subjects (Wikipedia, 2011c). The Notability of persons as article subjects (i.e., in biographies) follows the same criteria, with further guidelines for specific categories of people (academics, athletes, creative professionals, criminals, victims, etc.). The coverage of living subjects raises special issues, e.g., of privacy, and the guidelines explicitly caution against sensationalism and gossip (Wikipedia, 2011d).

The notability guidelines in the English and Polish Wikipedias are very similar, based on the same criteria and voting process. Indeed, the Polish version cites its English counterpart as one of its sources. Both editions of Wikipedia allow automatic inclusion of biographies of people included in leading print encyclopedias. The main difference is in the specification of criteria for categories important in Polish culture, such as soccer players and ski jumpers (two popular Polish sports) and officials in the Catholic hierarchy (the leading religion in Poland) (Wikipedia, 2011e,f).

The Polish Wikipedia is just one example of a Wikipedia edition in another language. As of January 15, 2011, 10 years after the launch of the English-language Wikipedia, there were 278 different linguistic editions of the online encyclopedia (Wikipedia, 2011g). The initial proliferation was caused, in part, by the bulk creation of articles by software robots (or "bots") "which can add a large set of articles using databases of information" (Lih, 2004, n.p.) that speakers of a language can then modify by deleting or adding content. Studies show that other linguistic versions of Wikipedia do not develop at the same pace, and economically developed countries have a higher rate of participation in Wikipedia than do underdeveloped countries (Rask, 2007). Nonetheless, the different linguistic editions all seem to follow the same overall pattern of development identified by Kittur et al. (2007); that is, in the beginning, many people submit, while after a few years, it is mostly a core group that contributes (Ortega et al., 2008).

Several recent studies have attempted comparisons of Wikipedia editions, raising questions about possible cultural differences and the origins thereof. Pfeil et al. (2006) examined the editing patterns in the French, German, Dutch, and Japanese Wikipedias for the article 'game' and concluded that the probability of different tasks being performed by editors of different ethnic background is correlated with their country's scores on Hofstede's (1991) dimensions of culture. Hofstede's dimensions also served as the theoretical background for a content analysis of talk pages in English, Japanese, Hebrew, and Malay by Hara, Shachaf, and Hew (2010); the authors found variations across language versions that correlated especially with Hofstede's power distance index. They report that "Eastern Wikipedias had significantly more courteous messages on each type of talk page than did the Western Wikipedias[… whereas c]onflict and disagreement behaviors were more frequently observed in the West" (p. 2103). Hofstede's dimensions of power distance and individualism were also found to correlate with the presence of experts in the Portuguese, Dutch, and French Wikipedias in a study by Lipsch (2009): Countries with high power distance reported fewer unique registered

experts than countries with low power distance, while the trend was reversed for individualism.

Stvilia, Al-Faraj, and Yi (2009) adopted a different approach, examining perceptions of quality as regards 'featured articles' in the Arabic, English, and Korean editions of Wikipedia. They found that the quality guidelines differed in Korean from the other two, but that the actual content of the Korean articles was more similar to the English articles than the Arabic articles were, most likely because the English and the Korean sites shared more of the same editors and the same article topics.

These differences raise the important question of whether Wikipedia content should be translated into all possible languages, allowing all users access to the same information regardless of their mother tongue, or whether linguistic versions should represent the body of knowledge typical for the cultural region of the language. The positions taken on this issue are divided. While some researchers propose using machine translation to duplicate and increase content across different linguistic versions (Adar et al., 2009), others point out that the different linguistic versions of Wikipedia serve different audiences. For example, Jones (2009) describes the benefits of the Welsh language Wikipedia, which is especially strong on local Welsh- and Wales- related topics, but rather limited in coverage of "outside Wales" topics. Despite the low number of editors and articles, the Welsh Wikipedia has attracted attention, and thus, Jones argues, it is good for the Welsh language. Moreover, the Arabic and Korean participants in Stvilia et al.'s (2009) study, through their rankings of featured articles, indicated a desire to promote articles on local topics and priorities. As one participant commented, "We need encyclopedia articles that interest Arab readers.... I wish you had made this effort to write about a subject that benefits your people" (p. 237). This preference is supported by the empirical findings of Hecht and Gergle (2009), who analyzed links between and within countries/regions in content in 15 language Wikipedia editions as of fall 2008 and found that the home region tended to be the geographic focus of each edition—that is, that the editions exhibited what they called a high degree of "self-focus."

Topic choices and linking patterns are two indicators that the content of different linguistic versions of Wikipedia reflects the interest of the linguistic/cultural group of contributors; level of detail is another indicator. Kolbitsch and Maurer (2006) provide the example of an article on the American chess player Paul Morphy. The English version had 5,466 words, a photo of the subject, and was supported by citations and references to external resources, while the German version consisted of only 290 words and did not provide any additional information. "This example shows, on the one hand, that Paul Morphy is an important person for Americans. On the other hand, it distorts reality and creates an imbalance in that it emphasizes "local heroes"" (p. 196). In contrast, a study by Adafe and Rijke (2006) found that articles relating to famous people in their English-Dutch comparison contained a high number of similar phrases in both languages, as opposed to articles on generic topics like classicism or tennis, which showed less overlap of text. Kolbitsch and Maurer's observations are anecdotal, however, and Adafe and Rijke's analysis was automated and did not examine Wikipedia entries qualitatively. Moreover, their contradictory findings leave open the question of

the extent to which different language versions favor own-culture famous persons. The present study addresses this question through comparative in-depth content and qualitative analysis of articles about famous people in the English and Polish versions of Wikipedia.

Research Questions and Hypotheses

This study aims to address two research questions:

RQ1: Are there differences between English and Polish versions of articles about famous persons?

RQ2: Do Wikipedia language versions favor "local heroes" in the amount and nature of their coverage?

The literature reviewed in the previous section, especially observations reported by Kolbitsch and Maurer (2006), Jones (2009), and Stvilia et al. (2009), suggests the following hypotheses:

H1: Systematic biases will be found in the English and Polish versions of articles about famous persons.

H2: Articles about Americans in English and about Poles in Polish (own-culture famous people; "local heroes") will have more content and more favorable coverage than will articles about Americans in Polish and about Poles in English (other-culture famous people).

Methodology

Language Selection

For our investigation, we chose two languages that differ in their social contexts. English is a global language spoken as a mother tongue by approximately 328 million people, mostly in the US, but also in Canada, the UK, Australia, and elsewhere, and as a second language by up to 1.4 billion people (Lewis, 2009). Wikipedia originated on January 15, 2001 in the US in English, which makes it an important basis for comparative analysis.

In contrast, Polish is a smaller language associated with one country. It is spoken by approximately 40 million people, mostly in Poland, and by Polish expatriates (Lewis, 2009). The Polish Wikipedia started in September 26, 2001 as the ninth language edition. It is currently the fourth largest edition of Wikipedia in terms of number of published articles (after English, German, and French), indicating that it is actively used, especially taking into consideration the relatively small number of Polish speakers.

This observation is borne out by Table 1, which shows the activity levels of the English and Polish Wikipedias as of October 2008, when the data were collected for this study. There were 27 times as many contributors to the English Wikipedia as to the

Polish Wikipedia, and they contributed five times as many articles, three times as many edits per article, and twice as many words per article. However, each individual contributor to the Polish Wikipedia contributed on average to 5.6 times as many articles, roughly three times as many edits, and twice as many words as the average contributor to the English Wikipedia.

Table 1. English Wikipedia and Polish Wikipedia activity as of October 2008 (source: http://stats.wikimedia.org/EN/TablesWikipediansContributors.htm)

	Articles	Words per article	Edits per article	Contributors	Articles per contributor	Words per contributor	Edits per contributor
ENGLISH	2,600,000	519.62	62.40	428,681	6.07	3151.53	378.46
POLISH	543,000	263.35	19.90	15,973	33.99	8952.61	676.50
Ratio EN:PL	5:1	2:1	3:1	27:1	0.2:1	0.4:1	0.6:1

Since Polish is a language spoken officially in only one country, choosing subjects of Polish nationality was straightforward. In the case of English, however, the subjects could potentially be chosen from any country where English was the first language. We decided that using the criterion of country rather than language in subject selection would lead to more coherent interpretations. Thus our English-language sample is composed of famous Americans, as representing the largest English-speaking country in the world and also the country in which Wikipedia originated.

This enabled a further basis for comparison relevant to our research questions and hypotheses: political ideology. Poland until recently belonged to the Soviet bloc, while the US has a democratic political system anchored in a capitalist economy. We reasoned that these differences might plausibly result in different ideological perspectives, or biases, in the entries, and thus that their two language versions would comprise useful cases to test our first hypothesis.

Data Sample: Famous People

We chose to analyze entries about famous people in order to address the conflicting claims in the literature about the treatment in Wikipedia of same-culture as opposed to other-culture persons. Moreover, biographical entries are a common genre of article in print encyclopedias, and we could be assured of finding entries whose content was prima facie comparable across the two languages. Entries on famous persons from Poland and the US written in Polish and English were compared, representing four nationality-language pairings: Americans-English, Americans-Polish, Poles-English, and Poles-Polish. We analyzed 60 entries (15 in each nationality-language pairing—that is, 30 individuals from each country and 30 entries in each language) totaling 254,826 words.

The people were selected from five domains commonly associated with fame: sports, politics, music, movies, and an intellectual domain including academe and

religion. For each language, three persons were selected for each domain. An attempt was made to include people who had more than minimal Wikipedia entries in both languages and were known at least by name to both researchers. We also tried to match historical period for Americans and Poles in each domain and to ensure representation of both genders in the sample, although it was not possible to match both gender and historical period for each person, given the other selection criteria. All persons selected met the Notability criteria for inclusion in Wikipedia, as discussed in the literature review. Table 2 lists the subjects selected for the study, their area(s) of fame, and the years in which they were famous.

Table 2. List of research subjects

Domain	**Americans**	**Poles**
Sports	Muhammad Ali (wrestling; 1960s-1970s)	Irena Szewińska (track and field; 1960s-1970s
	Michelle Kwan (skating; 1990s-2000s)	Andrzej Gołota (wrestling; 1980s-2000s)
	Lance Armstrong (cycling; 1980s-2000s)	Adam Małysz (ski jumping; 1990s-2000s)
Politics	George Washington ("father of country"; 1750s-1790s)	Tadeusz Kościuszko (general, national hero; 1770s-1810s)
	Theodore Roosevelt (US president; 1890s-1910s)	Józef Piłsudski (Head of State; 1900s-1930s)
	Condoleezza Rice (US Secretary of State; 2000s)	Lech Kaczyński (Polish president; 1970s-2000s)
Music	George Gershwin (composer; 1920s-1930s)	Ignacy Jan Paderewski (composer; 1880s-1930s)
	Frank Sinatra (singer and actor; 1930s-1990s)	Violetta Villas (singer; 1960s–2000s)
	Britney Spears (singer; 1980s-2000s)	Edyta Górniak (singer; 1990s-2000s)
Movies	Lillian Gish (silent film actress; 1910s-1950s)	Pola Negri (silent film actress; 1910s-1960s)
	James Dean (actor; 1950s)	Zbigniew Cybulski ("the Polish James Dean"; 1950s-1960s)
	Mel Gibson (actor and director; 1980s-2000s)	Krzysztof Kieślowski (director; 1970s-1990s)
Academe/ Religion	Linus Pauling (scientist; 1930s-1970s)	Marie Skłodowska Curie (physicist; 1890s-1930s)
	Sylvia Plath (poet; 1950s-1960s)	Wisława Szymborska (poet; 1950s-1990s)
	Martin Luther King, Jr. (religious leader; 1950s-1960s)	Pope John Paul II (religious leader; 1950-2000s)

Analytical Methods

The study was carried out in three stages using two different methodologies. The first two stages made use of structural and thematic content analysis. After initial analysis of a subsample of the data, a number of categories were established and refined using an iterative grounded theory approach (Glaser & Strauss, 1967). The structural categories

included: entry length (in number of words), presence and frequency of outlines, lists, references, external links, sidebar content, and photos. The thematic categories included favorableness of coverage, as well as mentions of personal information (family, romantic relationships, etc.), education, nationality, political ideology, and controversy. One-quarter of the data was coded using two coders for each language to assess inter-coder reliability. After at least 80% agreement was reached for each set of categories, the first author coded the remaining Polish data, and the second author coded the remaining English data.

The third stage of the study consisted of qualitative analysis of the entries with a focus on controversy. This involved close comparison of the entries in the two languages as regards inclusion or omission of controversial information. This stage of the research was conducted by the first author.

The results are presented below in three sections corresponding to the stages of analysis. For the content analyses, the unit of analysis was the article (N=15 for each nationality-language category). The quantitative results are presented as descriptive statistics, since the numbers of tokens in each of the four categories, when broken down by the individual code values, were too small to permit statistical analysis.

Results

Structural Content Analysis

The first structural result is that the biographical entries analyzed are much (nearly 11 times) longer than average Wikipedia articles in both languages. This is not entirely surprising, given that we excluded entries with minimal content from our sample, and given that a certain number of Wikipedia articles are undeveloped 'stubs.' Consistent with article length averages for the two language versions overall (see Table 1), the English entries are roughly twice as long on average as the Polish language entries, regardless of the country of origin of the subjects. Moreover, the entries about Americans are overall almost 10 times longer in English than in Polish, whereas the entries about Poles are more equal in length in both languages. While *all* entries about Americans are longer in English than in Polish, the entries for 10 Poles are actually longer in English than they are in Polish, albeit in two cases only slightly so. In three cases (Pola Negri, Marie Curie, Krzysztof Kieślowski), the English language entries are 50% longer. The breakdown of the proportion of the four nationality-language categories calculated in terms of words per entry is given in Table 3.

Differences were also found in the outlines of the entries. Presence of an outline and average number of levels in the outline correspond roughly to differences in entry length. However, the numbers of main categories and subcategories are higher for the linguistic versions that correspond with the nationality of the person (same-culture persons) than for the mismatched versions (other-culture persons). These results are summarized in Table 3.[2]

Table 3. Entry length and outlines

	Length of main entry (avg. # words)	% of total words	outline	main categories in outline (avg.)	subcategories in outline (avg.)	levels in outline (avg.)
AM-EN	8248.67	48.55%	15.00	12.27	11.80	2.07
P-EN	3451.93	20.32%	13.00	7.23	4.85	1.77
AM-PL	957.80	5.64%	7.00	8.00	2.67	1.44
P-PL	4330.00	25.49%	13.00	10.38	14.54	1.77

Most of the entries had a least one photograph, usually a portrait of the subject, with the exception of Andrzej Gołota (no photo in either language) and Sylvia Plath (no photo in the Polish version). The highest number of photos (N=40) was found in the Polish version of the entry on Józef Piłsudski (Head of State and the leader of the Second Polish Republic). Entries on Polish subjects in Polish had the most photos, followed by entries on Americans in English; the English entry on Theodore Roosevelt also included two videos. The main picture was often the same in both languages (it differed only in four cases: Mohammad Ali, Michelle Kwan, Martin Luther King, Jr., and Józef Piłsudski), although its realization varied somewhat in size, cropping, and evidence of photoshopping (the image of John Paul II appears to have had its background photoshopped out in the Polish version). The structural results for photographs are summarized in Table 4.

Table 4. Photographs

	Total # photos	Average # photos per entry	Avg. number of all photos that are identical in both versions		Percentage of main photos identical in both versions	
AM-EN	124	8.27	Americans	1.40	Americans	78.57%
P-EN	114	7.60				
AM-PL	38	2.53	Poles	3.29	Poles	92.86%
P-PL	149	9.93				

As Table 4 shows, there is a tendency for same-culture entries to have higher numbers of photos in absolute terms and per entry than other-culture entries. However, entries about Poles in English have almost as many photos as the two highest categories, so the same-culture vs. other-culture pattern is not perfectly symmetrical.

We also counted the frequencies of numbered and unnumbered notes and references, external links organized at the bottom of the entry, and lists of activities, accomplishments, and awards. The English entries about famous Americans had the most of all of these, as shown in Table 4, followed by the English entries about famous Poles, although there is significant variability within both sets (SD=62; SD=67): For example, the English entries for Condoleezza Rice and Martin Luther King, Jr. had 194 numbered references each, while Michelle Kwan's entry had none.

In contrast, the Polish language entries were generally sparse in references, especially in entries about Americans, where in nine cases no references were included. The average is higher for English entries about Poles, but this is because the entry for Józef Piłsudski had 185 numbered references, whereas six entries contained no references at all. Some Polish entries that did not have numbered references had quite a few unnumbered references (e.g., Michelle Kwan had 39 of the latter), suggesting that some Polish authors, perhaps unfamiliar with Wikipedia citation norms, treated the two sections interchangeably. However, the combined frequency of numbered and unnumbered references is still lower for entries in Polish than for entries in English.

The only features for which this pattern differed are lists of accomplishments and "see also" Wikipedia-internal links. For both features, own-culture persons had somewhat more than other-culture persons (see Table 5).

Table 5. Links and references (averages per entry)

	Reference/Notes				External links	"See also" (internal) links	Lists of accomplishments
	Numbered		Unnumbered				
	Count	SD	Count	SD			
AM-EN	79	62	17.80	67	14.13	3.33	3.40
P-EN	22	49	11.27	69	8.00	1.73	2.33
AM-PL	1	1	0.73	2	2.47	0.60	1.07
P-PL	13	18	9.00	27	5.47	2.87	4.13

Finally, we analyzed the amount and types of information included in the right sidebar of each entry. In most cases, this consisted of a photo of the subject and a summary of the most important information about the person, usually focusing on demographics, family status, and career highlights. Again, the Americans-English sidebars have the most of each type of information, as well as more words (see Table 6). Overall, the distribution of content across the four categories is roughly proportional to article length. The only feature that patterns in a way reminiscent of own- versus other-culture is list of accomplishments; these are found more in own-culture entries, although they are not very common overall. Additionally, categories of 'other information' were favored by English-language sidebars; they occur hardly at all in Polish-language entries.

Table 6. Information in the sidebar (averages per entry)

	Number of words	Categories of personal information	Categories of professional information	Categories of other information	Lists of accomplishments
AM-EN	101.93	5.73	8.07	1.07	1.27
P-EN	50.13	2.47	4.80	1.20	0.53
AM-PL	34.67	2.20	1.20	0.13	0.20
P-PL	65.93	2.73	3.67	0.07	1.07

Thematic Content Analysis

The second part of the content analysis focused on thematic types of information covered in the entries for each nationality-language category. The thematic variables included the tone of the overall content coverage; type of information included (personal, career, other); mentions of education; mentions of nationality; mentions of political ideology (two variables: communism/socialism and democracy); mentions of controversy; and types of controversies mentioned (personal vs. career related). Personal information was further broken down into information about spouses and romances, family members, and non-family other. The results for each variable by nationality-language category are presented in the following sections.

Tone of coverage

Entries about Americans were more positive in tone overall in both languages, especially in Polish, where all-positive or mostly-positive content was found in 93% of entries. (These were also the shortest entries.) In contrast, while the entries for famous Poles were also positive—all of the subjects were people famous for their accomplishments, after all—more of them were categorized as 'balanced,' due mostly to their tendency to juxtapose career accomplishments and life hardships (e.g., loss of a parent, childhood poverty) experienced by the subject. This tendency was typical of the Poles-Polish entries, and it was mirrored in many of their English language counterparts, which sometimes included parts translated from the Polish versions. These results are summarized in Figure 1.

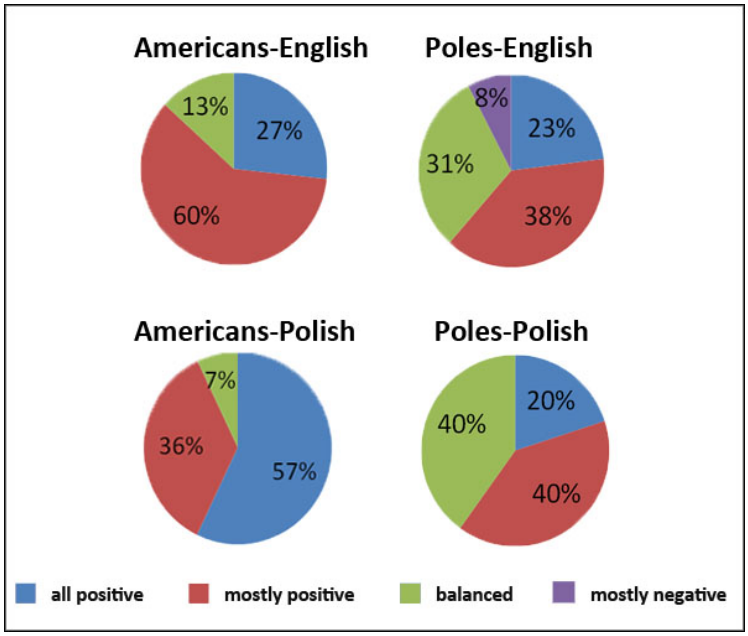

Figure 1. Tone of overall content coverage

Type of information covered

Each entry typically included more than one type of information, especially for Americans in English (see Figure 2). Only 12 entries' contents were limited to one type of information, professional work, most of them (N=7) in English entries about famous Poles.

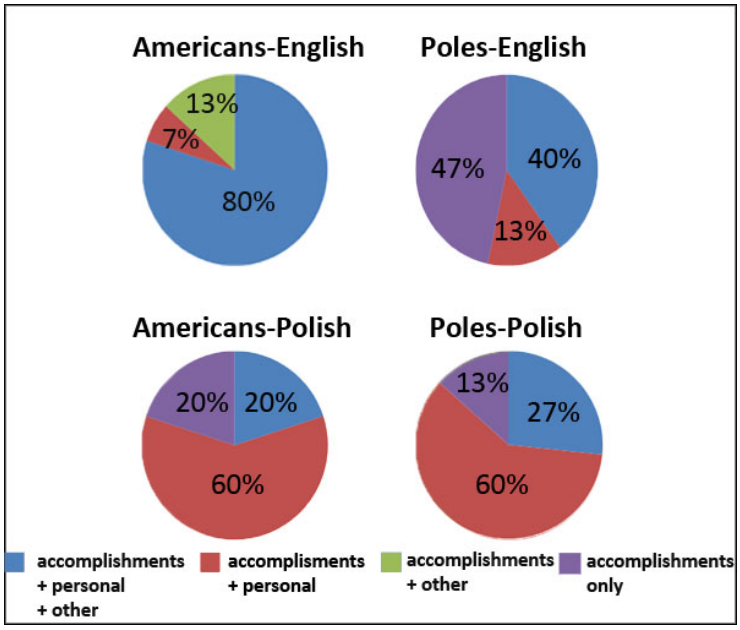

Figure 2. Types of content coverage

Personal content is considered in further detail below.

Personal information

While traditional print encyclopedia entries for famous people concentrate mostly on career highlights, the Wikipedia entries often provide personal information as well, including information about marriage and romance, family members, and other, e.g., health-related, information. The proportions of all types of personal information combined are shown in the pie charts in Figure 3. Entries about Americans in English include the most personal information; they are more likely to include 'a lot' and far more likely to include a 'moderate' amount of personal information than any of the other three nationality-language categories.

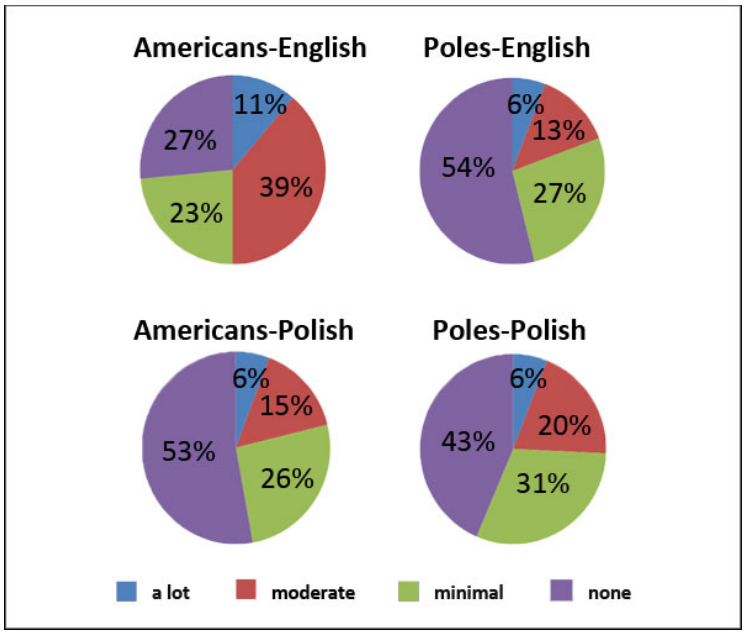

Figure 3. Amount of personal information (all types combined)

Results for the most common types of personal information included in the entries are presented below.

Spouses and romance

Information about spouses and/or romance was more frequently provided than any other type of personal information, especially for famous Americans. In three of the entries for Americans in English there was 'a lot' of information, and eight entries were coded as having a 'moderate' amount. Only two entries for Americans in English (Michelle Kwan, an Olympic skater, and a US Secretary of State, Condoleeza Rice) did not include such information. A considerable amount of romantic relationship information was also found in entries about Americans in Polish (proportionately even more than in Americans-English entries, relative to entry length, since the American-Polish entries were much shorter). Furthermore, more such information was reported in English than in Polish. The least amount of personal relationship information was found in Polish articles about Poles. Six Polish entries did not include any information about marital or extramarital relationships, and five of them had only minimal information, reporting the name of a spouse. See Figure 4.

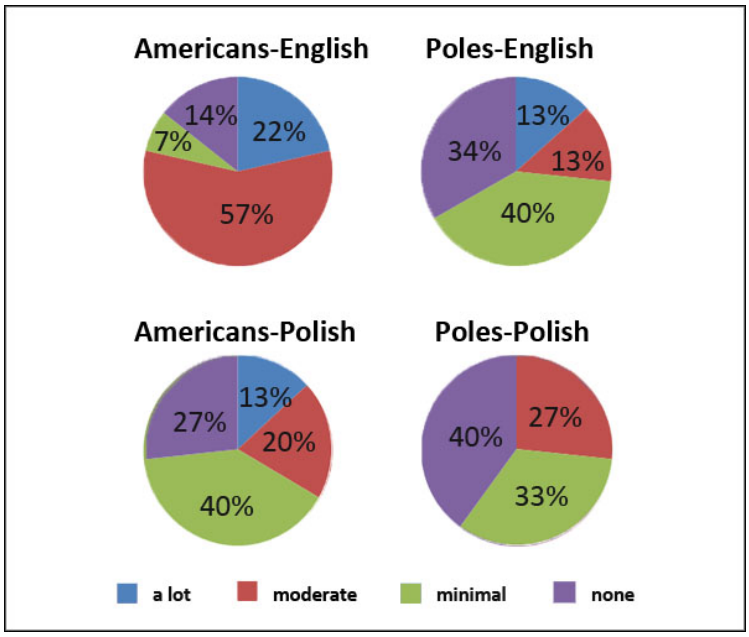

Figure 4. Mentions of spouses and/or romance

Family members

'Family members' in this analysis comprise parents, children, siblings, and other blood relations such as grandparents and uncles. Entries about Americans in English had the most mentions of these relations, although rarely was very much information provided about any of them (see Figure 5). When mentioned at all, parents were usually mentioned by name and occupation, and siblings and children were typically mentioned by name alone. Other relations were only mentioned if they played an important role in the subject's upbringing or career advancement.

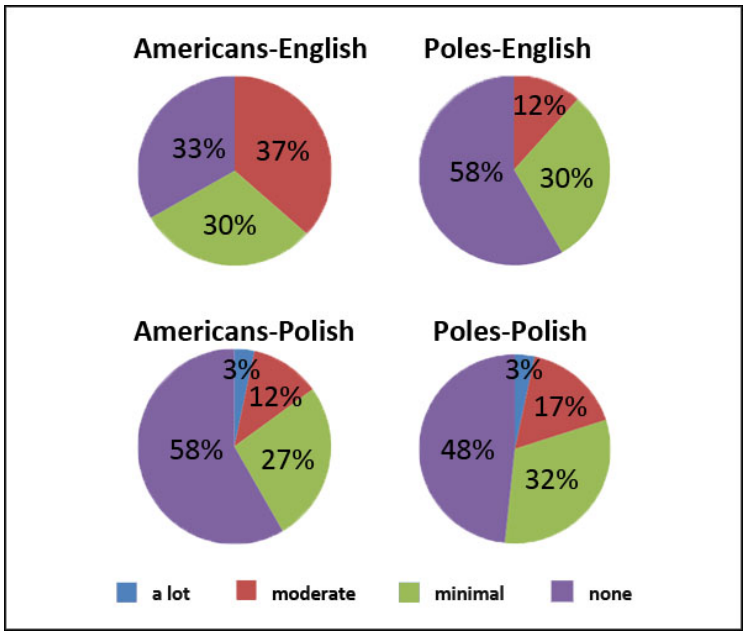

Figure 5. Mentions of family members

Other non-family

Other non-family personal information refers mostly to health problems and a subject's personal activities after his/her years of fame. By far the most information about other non-family is provided for Americans in English, but there is also a considerable amount in the Polish entries about Americans, especially when one considers their short length. See Figure 6.

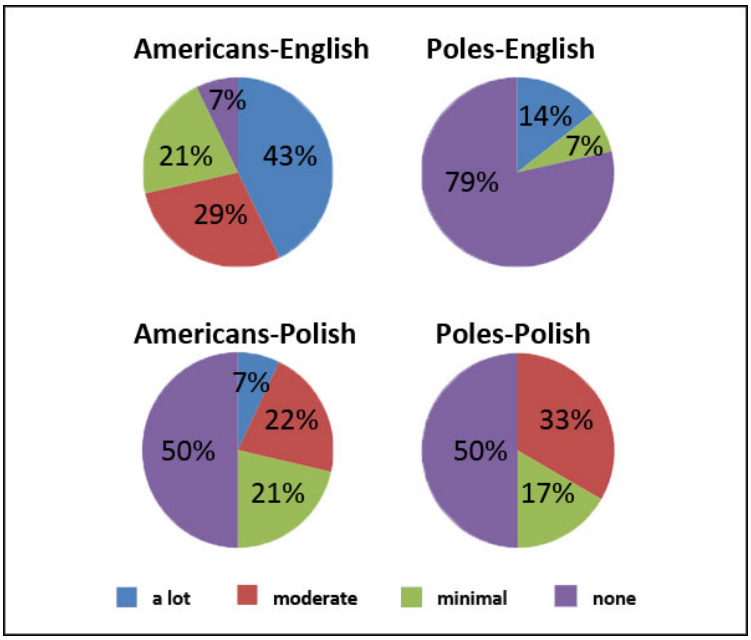

Figure 6. Mentions of other non-family

Education

Judging by the frequency with which it is mentioned, education is an important component in same-culture entries, especially for Poles in Polish. It is also frequently mentioned for Poles overall. The least amount of information about education was found in the Polish language entries about Americans, which are also the shortest. See Figure 7.

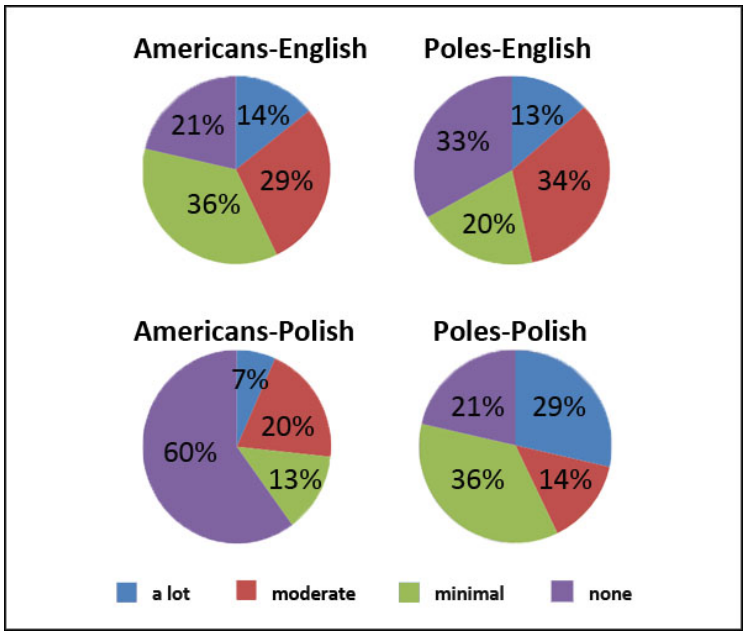

Figure 7. Information about education

Nationality

The subject's nationality was mentioned in all entries—unsurprisingly, since the entries are biographies of famous persons. However, entries about Poles have more mentions of nationality, especially proportional to their length, which is less than half of that of the Americans-English entries (see Figure 8). Interestingly, nationality is most emphasized in English language entries about famous Poles (other-culture), while entries about Americans in Polish (also other-culture) mention the subject's nationality the least.

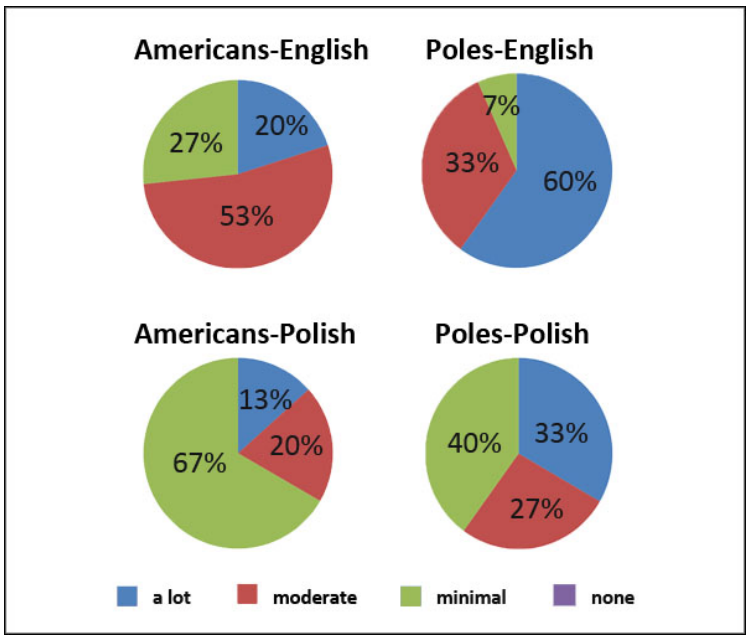

Figure 8. Mentions of nationality

Communism and Socialism

Mentions of communism and related concepts like socialism and class struggle are more frequent for Poles in both linguistic versions, especially for subjects involved in politics and the arts. In contrast, communist or socialist ideas are rarely mentioned for American subjects, although such themes are present in the English entries for Martin Luther King, Jr., Frank Sinatra, Linus Pauling, and Condoleezza Rice. See Figure 9.

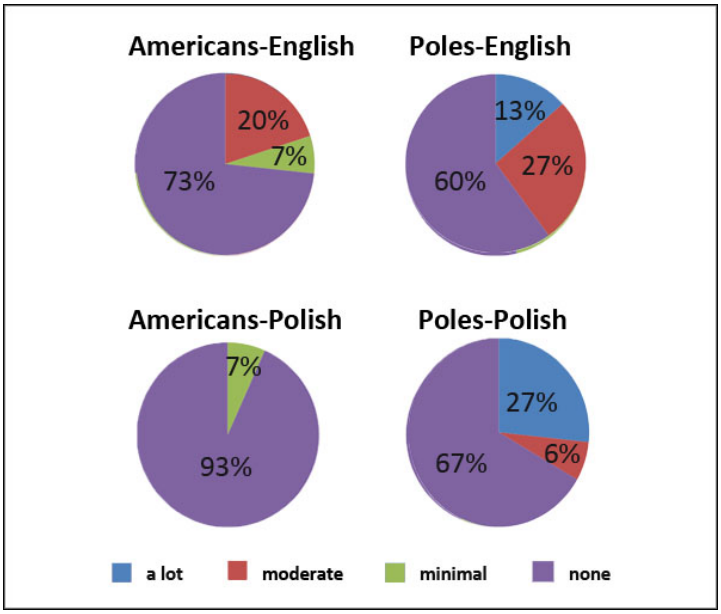

Figure 9. Mentions of communism and socialism

Democracy

Mentions of ideals related to democracy (e.g., freedom, equality, citizens' rights) are more common than themes related to communism or socialism In the study set, and were found in all four nationality-language categories (see Figure 10). However, they are most common in entries about Americans, especially in Polish. Democracy is also a relatively common subject in entries about Poles, especially juxtaposed with references to communism.

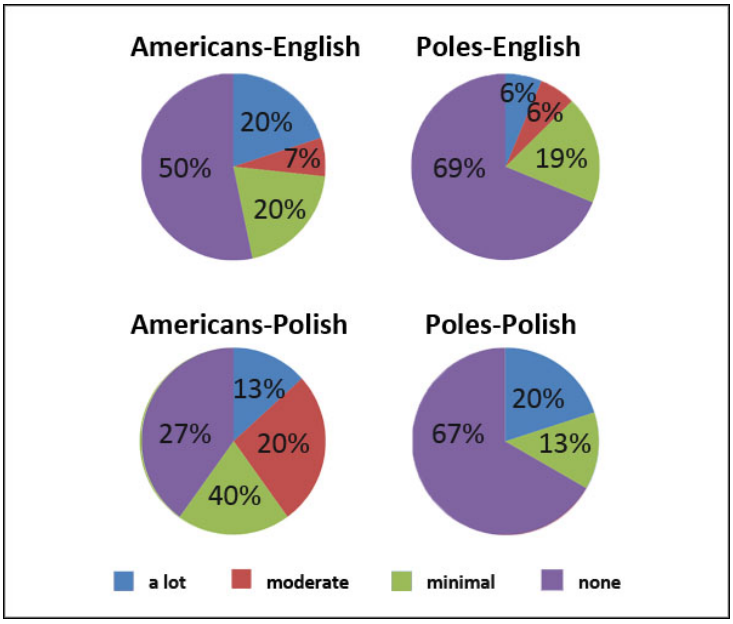

Figure 10. Mentions of democracy

Controversies and adversities

Controversies relating to career and personal life are a common theme in the biographical entries for most of the subjects. Some kind of controversy was noted in all entries in English about Americans, and in only three cases was the amount of information on the controversial issue 'minimal.' The frequency of controversies is also high in English entries about Poles, even higher than in the Polish versions about Poles. In almost 75% of the entries, the coverage falls in the 'a lot' or 'moderate' categories.

Mentions of controversies are less frequent in Polish. For Poles, although 53% of entries fall in the 'a lot' or 'moderate' categories, in one entry there is no negative information, and in six other entries a controversy is only casually mentioned. The entries for famous Americans in the Polish versions also mainly have 'minimal' or 'none' mentions of controversy. Overall, the English-language entries mention controversy more than the Polish-language entries do, especially when the categories 'a lot' and 'moderate' are combined. See Figure 11.

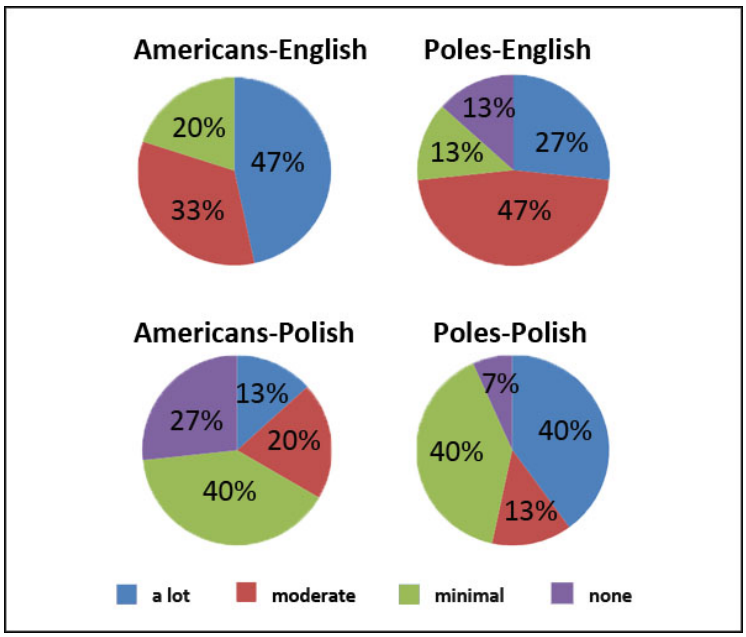

Figure 11. Mentions of controversy

Types of controversy

The controversies reported include extramarital affairs, problems with law enforcement, and politics (classified as 'personal'), career-related controversies (classified as 'career'), and controversial opinions (classified as 'career' when related to the subject's professional activities and 'personal' otherwise). A breakdown of the types of controversies reveals that personal controversies were mentioned more for Americans, while professional controversies were mentioned more for Poles, especially in English. However, career controversies were also mentioned often for Americans in English, and personal controversies were also mentioned often for Poles in Polish—that is, the same-culture entries are more likely than the other-culture entries to have entries of both types. These patterns can be seen in Figure 12.

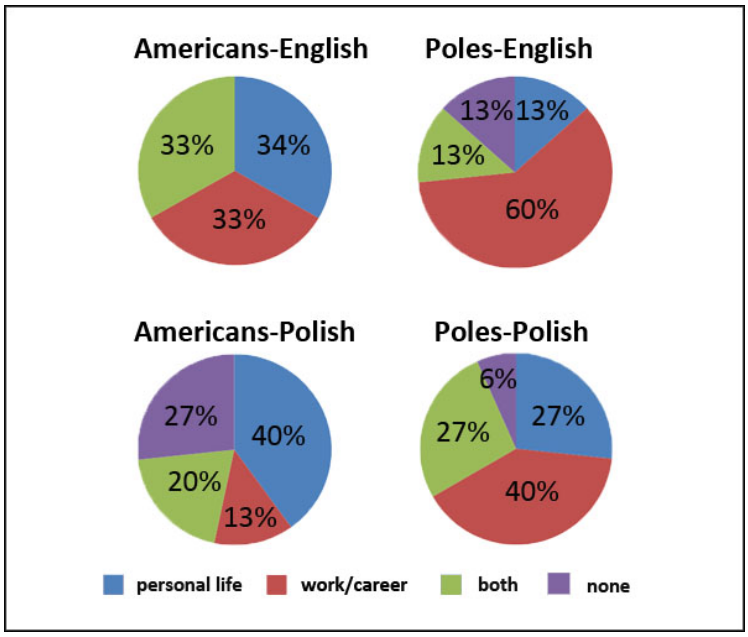

Figure 12. Types of controversy mentioned

Qualitative Analysis of Controversies, Life Adversities, and Omissions

In addition to controversy, mention of life adversities and the tone of language in which they are portrayed contribute to how famous persons are perceived by readers. The adversities emphasized in the entries analyzed in this study frequently relate to family poverty, death of a parent/sibling (especially when the subject was very young), imprisonment of the subject or family member, usually for political reasons, health troubles, and/or loss of fortune. The coverage of encyclopedia entries is also characterized by inclusion and/or omission of certain types of information. Omissions are difficult to analyze using quantitative content analysis techniques, and grouping content into categories can be reductionistic and result in nuances being overlooked (Bauer, 2000). In this section, controversies, life adversities, and inclusions/omissions are analyzed qualitatively to shed further light on the patterns identified in the quantitative analysis. Observations are grouped according to nationality; within each nationality, differences between the linguistic versions are noted.

Americans

It is difficult to generalize about the entries for Americans, since their English language entries are much longer than those in Polish. English versions overall mention more controversies and adversities than their Polish counterparts, due in part to the length differences. The Polish versions usually mention some of the controversies/adversities included in the English version, but in less detail. Only in rare cases is information provided in the Polish version that is not included in the English entry (e.g., the kidnapping of Frank Sinatra's son; the negative reception of Gershwin's *Porgy and Bess* by New York viewers). This pattern is evident throughout the corpus.

Life adversities. A difficult childhood, especially after the death of a parent, is sometimes mentioned in entries about Americans, albeit not as often as in entries about Poles (see the following section). However, detailed descriptions of health troubles are quite common in English language entries about famous Americans, both for the subjects and their families. Examples can be found in the entries for James Dean (bipolar depression), Lance Armstrong (testicular cancer), Michelle Kwan (injuries), George Gershwin (brain tumor), Frank Sinatra (heart attacks), Theodore Roosevelt (asthma), Silvia Plath (father's diabetes, Plath's mental problems, miscarriage), and Muhammad Ali (Parkinsons Disease). In some cases the illnesses are described with a level of detail that seems unnecessary in an encyclopedic entry, such as for example the illness and death of Linus's Paulings father: "Herman Pauling was suffering from poor health and had regular sharp pains in his abdomen. Lucy's sister, Abbie, saw that Herman was dying and immediately called the family physician. The doctor gave Herman a sedative to reduce the pain, but it only offered temporary relief." Most of these illnesses are also mentioned in the entries in Polish, albeit in less detail. Similarly, the death of a subject is sometimes described in considerable detail in the English version but simply reported in Polish. For example, there is a long section describing the death of James Dean in English, whereas the Polish version just mentions that he died in an automobile accident.

Romance and sexual lifestyle. Romances, extramarital affairs, and sexual orientation are common topics in entries about famous Americans in English. When included in the Polish versions, such topics are typically simply mentioned. Romances and affairs are reported for Lillian Gish (EN), Muhammad Ali (EN), Britney Spears (EN, PL), George Gershwin (EN, PL), Frank Sinatra (EN, PL), and Linus Pauling (EN). Unfulfilled romances are also mentioned for two subjects in the American set: The English version reports Linus Pauling's college love, while the Polish version reports Betsy Fauntleroy's refusal of George Washington's proposal of marriage and his platonic love for Sally Cary Fairfax. Sexual orientation is discussed for James Dean (EN, PL), whose homosexuality was controversial in his lifetime. Overall, romance and sexual lifestyle controversies are among the most commonly mentioned for Americans.

Career controversies. Professional controversies about famous Americans are more frequent in English. Examples include criticism of Britney Spears' projection of an overly-sexual style for young girls, Mohammad Ali's embarrassing fight against Alfredo Evangelista, in which both contenders seemed to refuse to fight each other, and the controversial "American beats out Kwan" headline on MSNBC after the Nagano Olympic competition. Some of the controversies, like Lance Armstrong's doping allegations, Frank Sinatra's possible connections to the Mafia, and Condoleezza Rice's refusal to testify before the National Commission on Terrorist Attacks upon the United States, are discussed in both language versions but again in more detail in English.

Ethnicity and nationality. Mentions of American nationality are quite frequent in English, but not as common in the Polish versions, and information about family heritage is also mostly limited to the English language entries. The only mention of a subject's

heritage in a Polish entry that is not included in the English version is Theodore Roosevelt's Dutch origins.

Political ideology. A common theme for American male subjects during the Vietnam War was their attitude toward the military draft. The avoidance of military service is reported in the English entries about several famous male Americans, including Mohamed Ali, who refused to be drafted, James Dean, who registered himself as homosexual (at that time classified as a mental disorder), Frank Sinatra, who claimed irrational fear of crowds, and Mel Gibson, whose father relocated to Australia in hopes that the "Australian military would reject his oldest son for the Vietnam War draft." This information is almost never mentioned in Polish. The Polish version of the last entry suggests that Gibson moved to Australia in protest against the Vietnam War, without mentioning the motivation of draft avoidance.

Interestingly, a subject's possible association with Communism is often mentioned in English entries but not in Polish (the sole exception is a short reference to Condoleezza Rice's Sovietology studies). The English entries discuss in detail accusations of communist sympathies for Martin Luther King, Jr., Frank Sinatra, and Linus Pauling. For example, the entry for Martin Luther King, Jr. records the actions of the FBI and J. Edgar Hoover in response to fears that Communists were infiltrating the Civil Rights movement, leading to extensive surveillance of King and his associates. The detailed descriptions of actions again King and King's assertion that "there are as many Communists in this freedom movement as there are Eskimos in Florida"–depict the political climate in the US during the Cold War era, as does the observation that "[t]he attempt to prove that King was a Communist was in keeping with the feeling of many segregationists that blacks in the South were happy with their lot but had been stirred up by 'communists' and 'outside agitators.'"

The lack of mentions of Communism in the Polish entries about Americans is somewhat surprising, especially in the case of Linus Pauling, who, as reported in English, "was awarded the International Lenin Peace Prize by the USSR in 1970." However, at only 202 words, the entry about Pauling is the shortest in the Americans-Polish set; it is possible that this entry was simply underdeveloped at the time we collected it for analysis.

Poles

The same topics relating to controversies and life adversities can be found in entries about famous Poles; however, the emphasis given to each topic differs from the American set, and some different issues are foregrounded.

Life adversities. Life adversities for Poles more frequently include a difficult childhood, as opposed to sickness and death. A difficult childhood due to family poverty or the death of a parent is mentioned in the entries for Pola Negri, Krzysztof Kieślowski, Jan Ignacy Paderewski, John Paul II, Józef Piłsudski, Marie Curie, and Tadeusz Kościuszko. Life adversities are emphasized in most of the entries equally in Polish and

English, which is understandable in that a number of entries seem like direct translations from Polish to English (as evident, e.g., from nonstandard grammar in the English versions). However, in some cases, contrary to the hypothesis that own-culture entries will be more informative, the English entries provide more detail, and, through the tone of the language used, the adversities are presented in a harsher light. An example is the English version of the entry for internationally-acclaimed film director Kieślowski, which includes the following passage:

> Leaving college and working as a theatrical tailor, Kieślowski applied to the Łódź Film School […]. He was rejected twice. To avoid compulsory military service during this time, he briefly became an art student, and also went on a drastic diet in an attempt to make himself medically unfit for service. After several months of successfully avoiding the draft, he was accepted to the Łódź Film School on his third attempt.

The Polish version reports these events simply as:

> In 1964-68, he studied at the Łódź Film School, to which he was accepted on his third attempt. [translation ours]

The greater elaboration in English is somewhat surprising, given that both the Polish and the English entries appear to have been written by Poles.

Romance and sexual lifestyle. Romances are also reported for Polish subjects, and similarly to adversities, the English language entries provide more details. Extra-marital affairs are mentioned in both languages for Pola Negri, Marie Curie, and Józef Piłsudski, but_the English version of the entry for Negri provides more information about her romances with Charlie Chaplin and Rudolf Valentino than does the Polish version. Unfulfilled romances are mentioned for Tadeusz Kościuszko and Marie Curie; the same information is presented in both versions for Kościuszko (the English seems like a direct translation from the Polish), but the English entry for Curie gives far greater detail.

There is no mention of a homosexual lifestyle for any of the Polish subjects in either linguistic version. The only references to homosexuality come from the entry on Lech Kaczyński, for his controversial actions and opinions on this issue.

Career controversies. Career controversies are frequently mentioned for Poles, even moreso than for Americans, but it is difficult to generalize about them because the range of careers represented in the corpus is very broad. The entry on John Paul II is one of the most interesting to compare between the two linguistic versions. Both versions include a *Criticism* section, but this section is considerably longer and more elaborate in English, where in adddition to criticism of the exclusion of women from the priesthood, his anti-conception stance, the celibacy of priests, the lack of stronger reactions by the Vatican to problems of pedophilia among priests, and his negative view toward theological freedom, the article mentions his support for Opus Dei, opposition to homosexuality and same-sex marriage, the use of charitable programs as a means to

convert people in the Third World to Catholicism, and acts like kissing the Quran in Damascus. In this case, since the Pope is an international personage, the English entry was likely written not just by Poles but by English speakers from various backgrounds; this could account for why it is longer and more detailed than the Polish version.

Ethnicity and nationality. Nationality and ethnicity are of frequent concern in entries about Poles. While entries about Americans often mention nationality, entries about Poles emphasize subjects' national and ethnic backgrounds (especially their identity as Polish) through tone and number of details included. Moreover, whereas entries in English tend to acknowledge all possible ethnic backgrounds of a subject, entries in Polish emphasize Polish nationality and tend to avoid mentioning other ethnic groups. For example, the ethnicity of John Paul II is described in English as Polish and Lithuanian, while the Lithuanian roots of John Paul II's mother are not mentioned in the Polish version. Similarly, the English entry describes Tadeusz Kościuszko as "a Polish, American, Belarusian, and Lithuanian national hero and general," while the Polish version mentions only his Polish and American connections. Moreover, the English version states that as a reward for his military service, Kościuszko was granted American citizenship, a piece of land, and a sum of money. While the Polish version mentions the material rewards, it states only that he also received a special letter of appreciation. Both versions mention his honorary French citizenship, but it is described in the Polish version only as a Title.

Interestingly, sometimes the English version mentions facts relating to the person's Polish ethnicity that are omitted in the Polish entries. For example, the English entry about Pola Negri reports that she refused to play a part in a German movie with an anti-Polish plot, and that she gave a large portion of her estate to Polish nuns. Similarly, patriotism to her country of birth is underlined in the entry for Marie Skłodowska Curie, who named one of the elements she discovered Polonium. These details are not mentioned in the Polish versions, for reasons that are unclear.

One type of controversy among Poles relates to individuals of actual or suspected Jewish descent. Although speculations about Negri's Jewish ancestry in relation to performing in German movies are mentioned in both entries, more details are provided in the English version. Similar speculations about Jewish ethnicity are reported in both versions of the entry on Curie, although the Polish version explicitly discredits this claim by providing a detailed account of her Polish/Catholic ancestry. In these ways, Polish Wikipedia authors seek to downplay Jewish ethnicity among their famous persons, claiming them to be simply Polish.

One exception is the entry on Irena Szewińska, who is described as of Jewish heritage in the Polish version and the victim of anti-Semitism during the infamous anti-Jewish attacks by the Communist government in 1968, which resulted in job losses and the emigration of the vast majority of Polish Jews. The English entry does not mention this incident or her Jewish heritage. Poles have reason to be sensitive around the topic of Polish-Jewish relations, not just from the treatment of Jews in Poland during the Second

World War, but from more recent history; this sensitivity may be expressed through omission or explicit rejection of that history.

Political ideology. Other controversies/adversities about Poles are related to the subjects' relationship to Communist Poland, either because they conformed to Communist rules or they participated in the opposition. Both linguistic versions of the entry for Lech Kaczyński mention his work in the opposition and his political imprisonment, and the Polish version also mentions his subsequent interactions with former Communists in the post-Communist era. Examples for artists deal with the lack of free expression in the era of censorship and the difficulty of finding a balance under a Communist regime. For example, both the English and Polish entries for the poet Wisława Szymborska reported that she initially accepted socialism, even though her early work was rejected for publication because it was not considered socialist enough, and that she denounced communism later in life. The English version of the entry for filmmaker Krzysztof Kieślowski reports both criticisms by colleagues for his cooperation with the government and his run-ins with censorship. Both languages document that the passport of popular singer Violetta Villas was withheld to prevent her returning to the US, and the Polish version also mentions other difficulties in her career resulting from the Communist regime. Overall, mentions of communism are more frequent in Polish than in English.

Summary of Results

In this section we summarize the results for each nationality-language category, as well as for nationality and language separately, triangulating the findings of the structural and thematic content analyses and the qualitative analysis.

Entries about *Americans in English* are the longest and contain the most outlines, references, and external links, as well as the most information in the sidebar. They also have the most diverse content and more mentions than the other three categories of personal information, especially about spouses and romance and health-related issues, and the most mentions of controversy and the Vietnam war. These entries were written by English speakers for English-speaking readers. Notwithstanding this potentially broad authorship and audience, given English's status as a global lingua franca, the patterns appear to reflect the cultural values and history of the US. These include the notion, promoted by the American mass media, that celebrities' private lives are of interest to average persons; a preoccupation with health; and a high tolerance for agonistic discourse (cf. Hara et al., 2010). The evidence in the entries of the first notion is especially compelling, in that it appears to contravene the English NPOV policy's explicit injunction that Wikipedia should not be a "vehicle for the spread of titillating claims about people's lives" (Wikipedia, 2011d, n.p.).

Entries about *Americans in Polish* are the shortest; however, they have the most positive content coverage and mentions of democracy, relative to their length. These entries were written by Poles for Polish readers, and their content may reflect what Poles find interesting about (or associate with) famous Americans. Many in Poland look

to the West, particularly the US, as a cultural and political model, especially since the end of communism.

Entries about *Poles in English* have the most mentions of education, nationality, and career controversy. The grammatical and stylistic evidence suggests that most of these were written by Poles, presumably for international readers of English. As such, they may reflect what the writers want outsiders to know about famous Poles: that they are well-educated, professional, and above all, Polish. However, the (long) entries on several international personalities who are ethnically Polish—John Paul II, Curie, Kieślowski—could have been written by other English-speaking nationalities, as well as by Poles. This could explain why more diverse ethnicities and nationalities are associated with them and why this category has more mentions of nationality overall.

Finally, entries about *Poles in Polish* have the most 'balanced' tone—which is to say, personal adversity is mentioned most often—as well as the most mentions of communist ideology and Polish nationality. They also seem to downplay multiple ethnicities and Jewishness. These entries are clearly written by Poles for Polish readers, and they reflect Polish history, values, and concerns. In particular, while the emphasis on personal adversity might appear to an American readership to undermine the biographical subject, to Polish readers, having overcome hardship is part of what makes the subjects worthy of admiration; it is a cultural value.

The results can also be summarized according to subject nationality:

Famous Americans overall receive more positive coverage, as well as being more associated with personal relationships, personal controversy (especially about romance and sexual lifestyle), non-family other information (e.g., about health), and democracy.

Famous Poles receive more balanced coverage (more mentions of adversity), as well as more mentions of nationality, career controversies, and communism.

Finally, there are differences relating to language version:

English language entries have more references and external links, as well as an overall more positive tone, a greater diversity of information, and more mentions of controversy. They also tend to be longer than Polish language entries.

Polish language entries are more likely to include information about professional accomplishments and personal life only, without mentioning other types of information.

Since the authorship of the last four categories is mixed, it is difficult to attribute any general explanations to the differences in content found, beyond that each nationality/language is associated with certain values, e.g., Americans and English language with "upbeat" coverage and (especially personal) controversy; Americans with democracy; and Poles with communism, national pride, and careers that overcome adversity. For both nationalities, the same-culture entries exhibit these associations

more strongly than the other-culture entries, further evidence that they are associated with the culture of each nationality. Finally, the differences in entry length and number of references and links can be attributed to the fact that the English Wikipedia is larger and more established than the Polish Wikipedia, and that it follows traditional encyclopedic norms (e.g., of attribution of sources) more closely (Emigh & Herring, 2005).

Hypotheses Revisited

This study asked whether there are differences between English and Polish versions of articles about the same famous persons in Wikipedia, and advanced a general hypothesis, based on previous literature, that systematic biases would be found. This hypothesis appears to be supported by our limited data; our analyses revealed a number of differences in coverage. However, there is no evidence that the resulting biases are intentional attempts to deceive or distort, as the word 'bias' may connote. Rather, they reflect the recent political and economic histories of the US and Poland, which shape the contributors' values in systematic ways, as summarized in the previous section.

We also asked if Wikipedia language versions favor "local heroes" in the amount and nature of their coverage, hypothesizing, following Kolbitsch and Maurer (2006) and others, that entries about Americans in English and about Poles in Polish (own-culture famous people; "local heroes") would have more content and more favorable coverage than articles about Americans in Polish and about Poles in English (other-culture famous people). This prediction was only partially supported. A same-culture advantage was found for numbers of main categories and subcategories in outlines; number of photos (to some extent); numbers of internal links and lists of accomplishments (including in sidebars); mentions of education; and in having a balance of controversies that are both personal and professional. However, these are not among the most important or revealing results of the analyses, and entries in the Americans-English and the Poles-Polish categories differ more than they resemble one another.

An explanation that accounts for more of the patterns is that there is an English-language and American nationality advantage, reflecting the fact that the English Wikipedia is based in the US and that it is larger and more active than the Polish Wikipedia (Ortega et al., 2008). It is also more diverse, because English is a global lingua franca, whereas Polish is a relatively small national language. Ultimately, these differences are part of a larger political reality, which is that the US is a powerful world power as compared to Poland's more limited influence and local situatedness. The asymmetries in the amount and nature of coverage in the English and Polish Wikipedia versions thus reflect larger asymmetries in the world (cf. Rask, 2007).

Implications and Recommendations

These findings challenge Wikipedia's NPOV policy, broadly construed, in that it is questionable whether content can be fair and balanced (cf. Reagle, 2005) when

systematic cultural biases exist. Objectivity normally requires that content be the same regardless of who reports it, and that different reports contain no notable omissions or elaborations, because these can have consequences for what the reader understands. Yet it is difficult to create perfect translation equivalents across languages—other cultures' normally-unarticulated assumptions would need to be spelled out, which could be awkward and would change the character of the entries; they would not be functionally equivalent.[3] Moreover, variations in content such as those analyzed in this study reflect real-world cultural differences; filtering them out to create homogenized content (as traditional print encyclopedias do) is artificial. Valuable information could be lost in an attempt to standardize across cultures. This suggests that strict "objectivity" should not be the primary measure of the worthiness of content for inclusion in Wikipedia. Different language versions may be equally "true," but present a subject from different perspectives.

Some researchers believe that automated translation tools will eventually be able to produce analogous knowledge in different languages, leading to a homogenization of Wikipedia content (cf. Adar et al., 2009). However, this approach is also problematic, in that it raises the question of whose version is "best" and deserves to be translated. Such an approach could be seen as dogmatic, even oppressive, and disadvantageous to smaller languages, especially given the proscription in the NPOV policy in the English Wikipedia, which is most likely to serve as the model from which content is copied into other languages, that the "views of tiny minorities should not be included [in Wikipedia] at all" (Wikipedia, 2011a). The Polish version of the Wikipedia NPOV policy explicitly stresses the liberating potential of the NPOV policy: "When we clearly tell readers that we do not expect that they must conform to certain concrete opinion, this means that we propose the execution of free choice; that is, we encourage the *intellectual independence* of our readers" (Wikipedia, 2011b, n.p., emphasis in original; our translation). Imposing translation equivalents from another language could be seen as contrary to this spirit and hence unlikely to be accepted by speakers from cultures such as Poland.

At the same time, there are legitimate reasons to be concerned about differences in Wikipedia language versions. Cultural differences in mentions of adversity, for example, could make famous Poles appear less successful and important to Americans, and differences in reporting personal controversy could make Americans appear more frivolous and scandal-prone to Poles and readers from other cultures. Such differences could breed or reinforce cultural stereotypes. Moreover, non-English speakers get less information than do English speakers from Wikipedia. Monolingual Poles, for example, would be disadvantaged when reading about Americans. It follows from this that bilinguals would get the most information from reading in English rather than their native language, at least on topics covered by both language editions. However, we would advise bilingual Wikipedia end-users to read articles both in their native language and in English, when available, to obtain the most integrated understanding of topical content.

Our view is that diversity across Wikipedia editions is acceptable, even desirable. Current trends suggest that the English language Wikipedia will continue to grow as a

general repository of "global knowledge," while Wikipedias in smaller languages will tend to retain their regional character and promote their "local heroes" and local values (Hara et al., 2010; Hecht & Gergle, 2009; Jones, 2009; Kolbitsch & Maurer, 2006; Stvilia et al., 2009). Rather than a weakness, this could become a strength, if good machine translation tools are available to translate from one language to another, and if readers approach Wikipedia editions in other languages with an awareness of cultural differences such as those identified in this study. In this scenario, multiple versions—including small, specialized Wikipedias—could exist alongside larger, more comprehensive ones, without denying anyone, monolingual or multilingual, access to the information contained in any version.

A recommendation that follows from this view is that Wikipedia content developers should allow linguistic editions to develop organically, rather than seeding them with content from other (especially the English) editions; instead they should focus on developing and providing ready access to accurate machine translation tools. While some borrowing across editions seems inevitable, rather than using machine translation to copy from one version to another, an alternative approach would be to provide a "translate" button on each Wikipedia page to allow readers to translate the page's contents dynamically on an as-needed basis. Such an approach would support a diversity of perspectives in Wikipedia as a whole, while respecting bodies of knowledge typical for the cultural regions of various language groups.

Limitations and Directions for Future Research

The strength and generalizability of our findings are limited by the fact that we analyzed only two language editions and one type of entry (biographies of famous persons); moreover, only 15 subjects in each nationality-language category were studied. Large-scale comparative studies are needed of languages and different content types to test further the hypotheses advanced in this article. Still, the results from the Polish-English comparison are suggestive and consistent with claims of previous research (e.g., Hara et al., 2010; Lipsch, 2009; Pfeil et al., 2006; Stvilia et al., 2009) that culturally-related differences exist in different language versions of Wikipedia.

Our study drew plausible inferences about the nationalities of authors based on reasoning from real-world circumstances and evidence of nonstandard language in the English entries about Poles. Analysis of edit histories would reveal who actually contributes, how many different contributors are involved for each article, and whether any of the same (e.g., bilingual English-Polish) people contribute to the different language editions (cf. Ortega et al., 2008; Stvilia et al., 2009). Such a study would also allow for comparison across cultures of Wikipedia editing strategies (cf. Pfeil et al., 2006).

In the present study, articles were selected for analysis that were of substantial length and found in both language versions. However, an important aspect of coverage bias concerns what topics are included, barely mentioned, or absent altogether from

different language editions (cf. Halavais & Lackaff, 2008). Previous research suggests that the majority of Wikipedia article topics are not shared across language editions (e.g., Hecht & Gergle, 2010; Oh et al., 2008). Further in-depth research is needed in this area.

A final, important limitation is that our study analyzed content only at one point in time. It would be interesting—and feasible, drawing on Wikipedia article edit histories—to analyze the evolution over time of articles in different languages/cultures. Anecdotally, we noticed that a later James Dean entry in Polish was much longer than the one we analyzed; the earlier version seemed to have been translated in large part from English. (Translation practices across Wikipedia editions, automated or otherwise, is another topic requiring further analysis.) It may be that the short, underdeveloped Americans-Polish versions were collected earlier in their developmental life-cycle; that they represented minimally-elaborated stubs translated automatically from English (Lih, 2004); and that they will catch up eventually. From what we observe at the time of this writing, however, it does not seem to be generally true that shorter Polish versions "catch up" with English versions over time (e.g., this is not the case for Pauling, Gershwin, Plath, or Gibson, although the entries for Ali, Dean, Kwan, and Spears are somewhat longer now). The extent to which other-culture entries continue to expand may depend on the extent to which articles are translated from one language to another and the enduring fame (or lack thereof) of the subjects in the other culture; this could be clarified through longitudinal analysis.

In the meantime, one conclusion seems clear: Wikipedia as a whole is a new and different kind of encyclopedia, one that incorporates cultural variability, regardless of whether or not that variability is intentional or endorsed by the Wikipedia community. The consequences of this fact require careful consideration by scholars, developers, and end-users of Wikipedia content.

Notes

[1] The word "Polonocentrizm" appears to be a translation from the English NPOV article, which cautions contributors to avoid "Anglocentrism." However, the nature of the risk of *-centrism* is quite different in each case, given the asymmetrical extent of Polish as compared to "Anglo" cultural influence.

[2] The following abbreviations are used in Tables 3-6: AM-EN (Americans-English), P-EN (Poles-English), AM-PL (Americans-Polish), and P-PL (Poles-Polish).

[3] The wording of the English NPOV policy itself seems to acknowledge the impossibility of perfect neutrality, even within a single language: "Editing from a neutral point of view (NPOV) means representing fairly, proportionately, *and as far as possible* without bias, all *significant* views that have been published by *reliable* sources" on the topic covered (Wikipedia, 2011a, emphasis added; the Polish entry expresses the same ideas). The hedge "as far as possible without bias" presupposes that there are limits to what is possible; in the case of different language editions, bias may be inherent in content in a given language. Moreover, what constitutes "significant views" and "reliable sources" may vary across cultures.

References

Adafre, S. F., & de Rijke, M. (2006). Finding similar sentences across multiple languages in Wikipedia. *EACL 2006 Workshop on New Text, Wikis and Blogs and Other Dynamic Text Sources,* 62-69.

Adar, E., Skinner, M., & Weld, D. S. (2009). Information arbitrage across multi-lingual Wikipedia. *WSDM '09*, 94-103.

Bauer, M. (2000). Classical content analysis: A review. In M. Bauer & G. Gaskell (Eds.), *Qualitative researching with text, image and sound* (pp. 131-151). Thousand Oaks, CA: Sage.

Chesney, T. (2006). An empirical examination of Wikipedia's credibility. *First Monday, 11*(11). Retrieved August 1, 2010 from http://www.firstmonday.org/issues/issue11_11/chesney/

Denning, P., Horning, J., Parnas, D., & Weinstein, L. (2005). Wikipedia risks. *Communications of the ACM, 48*(12), p. 152.

Emigh, W., & Herring, S. C. (2005). Collaborative authoring on the Web: A genre analysis of online encyclopedias. *Proceedings of the Thirty-Eighth Hawai'i International Conference on System Sciences.* Los Alamitos, CA: IEEE Press.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature, 438,* 900-901. Retrieved February 12, 2009 from http://bert.lib.indiana.edu:2146/nature/journal/v438/n7070/full/438900a.html

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine.

Halavais, A., & Lackaff, D. (2008). An analysis of topical coverage of Wikipedia**.** *Journal of Computer-Mediated Communication, 13*(2), 429-440.

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.

Hecht, B., & Gergle, D. (2009). Measuring self-focus bias in community-maintained knowledge repositories. *Proceedings of the fourth international conference on communities and technologies* (pp. 11-20). ACM.

Hecht, B., & Gergle, D. (2010). The tower of Babel meets Web 2.0: User-generated content and its applications in a multilingual context. *CHI 2010,* April 10–15, 2010, Atlanta, GA.

Hofstede, G. (1991). *Cultures and organizations—Software of the mind.* London: McGraw-Hill.

Jones, C. O. (2009). Look it up on Wicipedia. *Planet, 195*, 27-31.

Kittur, A., Chi, E. H., Pendleton, B. A., Suh, B., & Mytkowicz, T. (2007). Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *CHI 2007,* April 28-May 3, San Jose, CA.

Kolbitsch, J., & Maurer, H. (2006). The transformation of the Web: How emerging communities shape the information we consume. *Journal of Universal Computer Science, 12*(2), 187-213.

Lewis, M. P., Ed. (2009). *Ethnologue*, 16th edition. Dallas: TX: SIL International.

Lih, A. (2004, April). *Wikipedia as participatory journalism: Reliable sources? Metrics for evaluating collaborative media as a news resource.* Paper presented at the 5th International Symposium on Online Journalism, Austin, TX. Retrieved February 12, 2009 from http://staff.washington.edu/clifford/teaching/readingfiles/utaustin-2004-wikipedia-rc2.pdf

Lipsch, M. (2009). *National culture and the presence of experts on the online encyclopaedia Wikipedia.* Unpublished Master's thesis, Maastricht University. Retrieved October 15, 2010 from http://arno.unimaas.nl/show.cgi?fid=16743

Oh, J.-H., Kawahara, D., Uchimoto, K., Kazama, J. I., & Torisawa, K. (2008). Enriching multilingual language resources by discovering missing cross-language links in Wikipedia. *WI-IAT 2008*, 322-328.

Ortega, F., Gonzalez-Barahona, J. M., & Robles, G. (2008). On the inequality of contributions to Wikipedia. *Proceedings of the Forty-First Hawai'i International Conference on System Sciences.* Los Alamitos, CA: IEEE Press.

Pfeil, U., Zaphiris, P., & Ang, C. S. (2006). Cultural differences in collaborative authoring of Wikipedia. *Journal of Computer-Mediated Communication*, *12*(1), article 5. http://jcmc.indiana.edu/vol12/issue1/pfeil.html

Rask, M. (2007). The richness and reach of Wikinomics: Is the free web-based encyclopedia Wikipedia only for the rich countries? In *Proceedings of the Joint Conference of the International Society of Marketing Development and the Macro-marketing Society*. Retrieved February 24, 2009, from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=996158

Reagle, J. (2005). *Is the Wikipedia neutral?* Retrieved February 12, 2009 from http://reagle.org/joseph/2005/06/neutrality.html

Sanger, L. (2004, December 30). *Why Wikipedia must jettison its anti-elitism*. Retrieved March 27, 2008 from http://www.kuro5hin.org/story/2004/12/30/142458/25

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

Wikipedia. (2011a). *Wikipedia:Neutral point of view.* Retrieved March 20, 2011 from http://en.wikipedia.org/wiki/NPOV

Wikipedia. (2011b). *Wikipedia:Neutralny punkt widzenia (pełna wersja.)* Retrieved March 20, 2011 from http://pl.wikipedia.org/wiki/Wikipedia:Neutralny_punkt_widzenia_ (pełna_wersja)

Wikipedia. (2011c). *Wikipedia:Notability.* Retreived March 20, 2011 from http://en.wikipedia.org/wiki/Wikipedia:Notability

Wikipedia. (2011d). *Wikipedia:Notability (people).* Retrieved March 20, 2011 from http://en.wikipedia.org/wiki/Wikipedia:Notability_(people)

Wikipedia. (2011e). *Wikipedia:Encyklopedyczność.* Retrieved March 20, 2011 from http://pl.wikipedia.org/wiki/Wikipedia:Encyklopedyczność

Wikipedia. (2011f). *Wikipedia:Kryteria_umieszczania_not_biograficznych.* Retrieved March 20, 2011 from http://pl.wikipedia.org/wiki/Wikipedia:Kryteria_umieszczania_not_ biograficznych

Wikipedia. (2011g). *List of Wikipedias*. Retrieved March 20, 2011 from http://en.wikipedia.org/wiki/List_of_wikipedias