FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Comparison of Wikipedia articles in different languages

MASTERARBEIT

zur Erlangung des akademischen Grades

Master of Science

im Rahmen des Studiums

Wirtschaftsinformatik

eingereicht von

Dipl.-Ing. Nemanja Rajcic
Matrikelnummer 1329450

an der Fakultät für Informatik
der Technischen Universität Wien
Betreuung: ao.Univ.Prof. Mag. Dr. Wolfdieter Merkl

Wien, 25. April 2017

Nemanja Rajcic	Wolfdieter Merkl

FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Comparison of Wikipedia articles in different languages

MASTER'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Business Informatics

by

Dipl.-Ing. Nemanja Rajcic
Registration Number 1329450

to the Faculty of Informatics
at the TU Wien
Advisor: ao.Univ.Prof. Mag. Dr. Wolfdieter Merkl

Vienna, 25th April, 2017

Nemanja Rajcic	Wolfdieter Merkl

Erklärung zur Verfassung der Arbeit

Dipl.-Ing. Nemanja Rajcic
Oelweingasse 12/62, 1150 Vienna

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 25. April 2017

Nemanja Rajcic

Acknowledgements

First of all, I would like to show my gratitude to professor Dieter Merkl for suggested topic and professional help which made it possible for me to achieve one of my biggest goal and complete my master studies.

I would like to give special thanks to my parents Ružica and Branko including my brother Miloš who believed in me and gave me all the understanding and support I needed to reach this finish line.

Finally, I would like to say thanks to my friends and for their immeasurable support during the preparation and writing of master's thesis.

Kurzfassung

Diese Arbeit untersucht das Phänomen Wikipedia im Kontext der Mehrsprachigkeit und versucht einen Beitrag zu dem soziopolitischen Bild einer Region zu leisten, die in der jüngeren Vergangenheit stark durch den Krieg geprägt wurde. Die Forschung beschäftigt sich mit der Hypothese des 'lokalen Helden', welche davon ausgeht, dass der Informationsgehalt über Personen, in deren Sprache der Artikel verfasst wurde, den Informationsgehalt jener Artikel übersteigt, die über Personen aus anderen Ländern verfasst wurden. Außerdem analysiert sie die Qualitätssteigerung der Artikel, die durch erhöhte Teilnahme am Verfassen und durch hohe Besucherzahlen erzielt wird, also das Argument der 'kritischen Masse'. Auch die Verbesserung der Qualität im Zusammenhang mit der Bekanntheit der Person wird untersucht.

Die Masterarbeit befasst sich in diesem Kontext mit drei Artikeln, die über berühmte Personen aus Serbien, Bosnien, Herzegowina, Kroatien und Slowenien verfasst wurden. Dabei werden unterschiedliche sprachliche Versionen analysiert: serbisch, bosnisch, kroatisch, serbo-kroatisch und englisch. Die Untersuchung bezieht sich sowohl auf quantitative Daten (Wörterzahl, Größe in Bytes, Besucherzahlen etc) sowie auf qualitative Daten, wobei die Chico Richtlinien in diesem Zusammenhang als Evaluierungsintrument der Informationsqualität dienen.

Die Ergebnisse der Arbeit bestätigen sowohl die Hypothese des 'lokalen Helden' wie auch die der 'kritischen Masse'. Da bisher keine vergleichbare Forschung durchgeführt wurde, repräsentiert diese Arbeit einen neuen Forschungszugang innerhalb des Feldes und dient gleichzeitig als Ausgangspunkt für weitere multilinguale Studien zur Informationsqualität.

Abstract

This thesis examines the Wikipedia phenomenon in the context of its multilingual character, through the lens of a greater sociopolitical context of a recently war-torn region. The research tackles the hypothesis of the 'local hero,' which stipulates that content about persons from the country in whose language the article is written is going to be superior as opposed to content about persons from another country. Furthermore, it analyzes the improvement in quality of articles through greater participation and number of visitors, which is known as the 'critical mass' argument. Finally, the improvement of articles in relation to a person's importance is analyzed.

The thesis takes into account three articles about famous persons from Serbia, Bosnia and Herzegovina, Croatia, and Slovenia, across different Wikipedia language versions – Serbian, Bosnian, Croatian, Slovenian and Serbo-Croatian, as well as English. The articles are analyzed quantitatively (word count, length in bytes, number of visitors etc.) and qualitatively, using the Chico Guidelines for information quality evaluation.

The findings of the thesis confirm both the 'local hero' and 'critical mass' hypothesis. Seeing as a similar research has not yet been undertaken, this thesis represent a pioneering research in the field, and represents a strong starting point for multi-lingual information quality studies.

Contents

CHAPTER 1

Introduction

In the midst of what is termed as the *digital revolution of the 21st century*, lies the Internet. Singled out by many as the most transformative force since the Industrial Revolution, this complex structure consisting of an uncountable number of pages of text, images and video plays an indispensable role in the lives of most people on Earth. Housing an unimaginable amount of information, this recent phenomenon, although studied, continues to be the subject of a vast body of research.

One of the most influential pages on the World Wide Web is definitely Wikipedia, the free, online encyclopedia founded in 2001. As one of the most visited websites on the Internet, as well as a revolutionary form of knowledge creation and dissemination, this website exercises tremendous influence and is therefore the chosen subject of this research.

The aim of this thesis is to analyze, both quantitatively and qualitatively, the difference in breadth and quality of content about famous persons from former Yugoslav countries, examining simultaneously different quality indicators across the languages of the former republics. Although there have been significant findings confirming prevalence of neutrality and overall quality of content in the English Wikipedia, there has been no undertaking of such kind when it comes to Wikipedia's Serbian, Croatian, Bosnian, Slovenian and Serbo-Croatian versions.

Keeping in mind the tumultuous recent history of the Balkan region, this study can also be used as an indicator whether there is still perpetuation of cultural and national bias in these countries, and a test of geopolitical instability in the Balkan region. In a broader sense, this thesis is expected to contribute to the relatively small body of knowledge about what happens to Wikipedia's neutrality when the number of contributors and editors is not large enough to ensure that inaccurate, biased and subjective information is tagged and removed efficiently. It may serve as a starting point for researchers from a variety of fields, especially those concerning information quality in multilingual settings.

The scope of this thesis is multi-layered. On the one hand, this research will primarily address the broad differences between information available to Wikipedia visitors in different languages. Using the English Wikipedia articles as a starting point, the research will compare articles about famous persons from former Yugoslav republics in English, Serbian, Croatian, Bosnian, Slovenian and Serbo-Croatian languages, particularly taking into account the verifiability of the articles. On the other hand, the study will attempt to identify occurrences of bias and the exact situations wherein they occur through qualitative analysis. Ultimately, this research will analyze whether there is a correlation between "Wikipedia popularity" translated into number of monthly hits of a page, and the neutrality of the article, thereby testing the findings for English Wikipedia articles which suggest that article quality and neutrality inevitably improves with more contributors and visitors.

1.1 Problem definition

Wikipedia's slogan is "the free encyclopedia anyone can edit" [Wike]. As such, its policy of volunteer-produced content may, despite all efforts against it, lead to the perpetuation of inaccurate or incorrect information. Given the popularity and widespread use of the website as a reliable source of information, the potential impact of this problem becomes graver. Guidelines for Wikipedia articles stipulate that all articles should represent a neutral point of view (NPOV), wherein each article should cover all important information on the topic, all information must be correct and no bias or prejudice should be demonstrated [Wikm]. Many researchers analyzed the accuracy of information presented in Wikipedia articles. One of the most famous surveys, conducted only four years after the founding of Wikipedia, was published by the British magazine Nature. The researchers did a comparative analysis of two English-language encyclopedias [Gil05], the Britannica and Wikipedia. Among 45 analyzed articles, it was found that Wikipedia articles had an average of 4 errors per article, while Britannica articles had an average of 3 errors per article [Gil05]. Unfortunately, little attention was paid in the research to what extent Wikipedia articles upheld the NPOV, and the extent to which Wikipedia articles are balanced and fairly written.

The problem of bias and prejudice augments when the website is looked in terms of its multilingual character. Wikipedia articles are not always translated from English, but are often written by volunteers in their native language. Since volunteers may carry cultural or national prejudices of their own, there is a distinct possibility of Wikipedia articles in a specific language being colored by cultural or national bias or prejudice, particularly when the subject of the articles are events or people that are important in the volunteer's culture or nation. Pheil, Zaphiris and Ang investigated how national culture influence the style of contributions in four different language versions of Wikipedia [PZA06], while Denning, Horning, Parnas and Weinstein have identified six possible risks that Wikipedia's structure carries with it: the use of unapproved or unknown sources, inaccuracies, instability, insecurity and lack of knowledge about motives that drive volunteers to contribute to the encyclopedia's database [DHPW05].

This problem can be further exacerbated in situations where events or individuals covered in the articles are less popular or controversial. The social impact of this phenomenon is undeniable and can pose serious consequences: given that Wikipedia articles are often the first source of information for many individuals, and are thereby a strong influence on the formation of their opinion, incorrect or prejudiced information can perpetuate cultural prejudice and, in the long run, feed animosity, hatred and misinformation. A good example of this problem can be found in the comparison of two articles that cover the Bosnian War that took place between 1992 and 1995, written in Serbian and Bosnian. When describing the beginning of the war and the events that mark it, the Serbian article1[1] refers to the (alleged) shooting of a Serbian couple and their wedding guests by a Bosnian soldier in front of the Old Church in Sarajevo, while the Bosnian article2[2] refers to the demolishing of the village Ravno in the Herzegovina part of the country. The Croatian article3[3], on the other hand, singles out a military show of force from the Serb-controlled Yugoslav National Army in Polog, a village in Bosnia as the first war event.

An important piece of information is that more economically developed countries have a higher rate of participation and use of Wikipedia [Ras08], and are therefore more likely to host neutral and balanced articles, given the greater number of both contributors and editors. Hence, the problem of inaccurate and biased articles is likely to be more present in Wikipedia versions of less economically developed countries, particularly those whose recent history is plagued by war, conflict and dissension. For this reason, this research will look at articles written in the languages of the Balkan Peninsula: Serbian, Bosnian, Croatian, Slovenian and Montenegrin about famous people from these countries. The research will aim to establish the differences in amount of information, accuracy of information, type and number of sources used and correlation between the country in whose language the article was written and the role (or lack of role) that famous person played in the nation's history.

1.2 Aim of this work

The aim of this work is to examine the differences between the content, structure and neutrality of Wikipedia articles written in different languages of the countries of Former Yugoslavia, in order to establish whether there are significant differences in the data available to visitors of Wikipedia in different languages. The paper will aim to pinpoint whether there is a prevalence of bias or inaccurate information, and identify the circumstances in which such situations occur. Moreover, this research will compare information available in the Balkan languages with what is available in English, as

[1] https://sr.wikipedia.org/wiki/Rat_u_Bosni_i_Hercegovini. Visited: November 10, 2016

[2] https://bs.wikipedia.org/wiki/Rat_u_Bosni_i_Hercegovini. Visited: November 10, 2016

[3] https://hr.wikipedia.org/wiki/Rat_u_Bosni_i_Hercegovini. Visited: November 10, 2016

a means of establishing whether the situation is any different for the international Wikipedia. Finally, the research will analyze articles about famous persons from Balkan countries in terms of their "Wikipedia popularity," meaning the number of hits the page gets on a monthly basis, the number of contributors and the number of languages into which the page was translated. These findings will then be used to establish whether there is a correlation between "Wiki popularity" and objectivity and neutrality of the article, thus putting to a practical test the idea that Wikipedia articles improve with more participation and contribution from volunteers (also known as the critical mass argument).

1.3 Expected Results

There are a number of premises that this research will be based on. To begin with, the first premise of this research is that the quality and accuracy of information, as well as the overall neutrality of a Wikipedia article improves with its popularity, overall significance of the subject of the article, and the number of contributors to the article. Furthermore, this research is based on the premise that Wikipedia in languages of the Former Yugoslav countries is not nearly as popular as the English Wikipedia, which contributes significantly to the possibility that the abovementioned criteria of quality is not met in many articles. Finally, this research is based on the premise that articles in different languages of Balkan countries about subjects that tackle their mutual history are likely to perpetuate cultural and national myths and bias that are part of the nation's regular discourse.

Hence, the researcher expects to find that there are significant differences between the information available on different Wikipedia pages based on their language, and that neutrality is likely to be decreased as the controversy and importance of the subject of the article increase. The researcher also believes that there is a strong likelihood that Wikipedia's official criteria for neutrality will not be followed as strongly as in the English Wikipedia.

Given that there was little research done in this particular area, the relevance of this work lies in its pioneer role in this area of research. Specifically, this research will contribute to the knowledge available to the public about not only the correlation between Wikipedia popularity and bias, and therefore the efficiency of Wikipedia's volunteer-based model, but also in the field of knowledge about how Wikipedia of the Balkan countries helps or hinders the perpetuation of cultural myths and misinformation, and will be a solid ground for further research in the field of Wikipedia's social and cultural influence.

1.4 Methodical Approach

This research will be conducted through both a quantitative and qualitative analysis of Wikipedia content on famous persons from former Yugoslav republics, and a comparison of those articles in six Wikipedia language versions. The choice of famous persons will

be based on the Wikipedia articles *List of famous persons* from each of the respective countries, taking three famous persons from each of the lists, based on the length of the article. Taking into account the previous empirical research done on the subject of Wikipedia content quality, quantitative metrics of length of article, availability in other languages, number of references, number of external links, categories within the article, total number of views, daily average number of views, and total number of editors will be performed. The results across languages will then be compared. Through these metrics, the research will attempt to distinguish between higher- and lower-quality articles. The results will then be validated through the use of the *Cornell University Source Guidelines* in order to establish the qualitative value of the articles. Furthermore, the qualitative analysis through the *Cornell University* guidelines will attempt to examine the neutrality and balance of information in the articles, also taking into account omission and inclusion of information across language versions.

The data will be collected through the use of the Wikipedia Sandbox API and Wikidata, both of which are freely accessible tools through the Wikipedia website and Wikimedia Foundation. The extracted data in its raw form will be presented in the Methodology section, followed by the results of the analysis.

1.5 Structure of the Work

The first Section of the thesis will present, on the hand, the historical development of Wikipedia and its role in the greater context of the Open Source Movement and participatory journalism. A detailed overview of the development of the first Wikipedia, the English version, along with the history of the Wikipedia in the former Yugoslav region will be followed by vital statistics about its usage, including, most importantly, the size and scope of each of the Wikipedia versions. Furthermore, this section will describe in great detail the principles upon which Wikipedia is founded, including its most important *Five Pillars*, which will serve as a starting point for this study and the research questions it is trying to answer. Finally, it will present to the reader some of the most prominent criticisms that Wikipedia is facing, therefore painting a fuller picture of the context of the Wikipedia phenomenon. The historical overview will be followed by a literature review of previous work studying Wikipedia. The literature review will be divided into descriptive and empirical studies; descriptive studies will further examine the principles and mechanisms of Wikipedia, acquainting the reader with vital information about the inner workings of Wikipedia, while the empirical studies will be used as a foundation of the creation of the research methodology of this thesis.

The second Section of the thesis will describe in great detail the methodology that was decided upon after extensive research. It will outline the research process, including data collection and choice of the data sample, indicate the limitations of the research, and present the key tools that were used to collect and organize the data that is analyzed in this study. Furthermore, this section will introduce the concept of KPIs, *key performance indicators*, which will be used during the quantitative stage of the research process, as

well as the *Cornell University Source Guidelines*, which will be used as an instrument of the qualitative analysis of the subject of this research. In conclusion, this section will connect the knowledge derived from the research process and literature review with the actual subject of the study, therefore demonstrating the use of this knowledge in a practical sense.

Section Three of this thesis will present the analysis of the collected data using the methodology described in the previous section, followed by a discussion of the results. It will present the results that were achieved through the data analysis, representing graphically the findings of the study, along with a discussion regarding the outcomes of the research. In short, this section will answer the two research questions guiding this study:

1. What is the difference in availability and quality of data about historically important persons across different languages of Wikipedia, as well as in English?
2. To what extent, if any, does the objectivity, neutrality and overall quality of Wikipedia articles increase when the article has more contributions, and when the subject of the article is a person of global importance?

The fourth and final Section will present a conclusion of the study, which will, in addition, highlight the limitations of the study, areas of improvement, and suggestions for further research. Keeping in mind that this study is a pioneering work in the area, the author judges that this section is of vital importance for researchers seeking to examine the Wikipedia phenomenon from a multilingual perspective in the future, particularly in the case of disputed and/or controversial subjects.

CHAPTER 2

Theoretical background

2.1 History of the encyclopedia

Although the emergence of the "first" encyclopedia cannot be pinned down to a single date, the first encyclopedic work to have survived until modern times is the *Naturalis Historia* by the Roman statesman Pliny the Elder, which dates back to the 1st century AD [Lin92]. Although today it serves as an insight into the Roman way of living and a historical resource regarding Roman art and technology, it nevertheless embodies many entries that have been confirmed by today's research [Liv].

Until the 14th century, and the invention of the printing press, encyclopedias were rarely available to anyone but the lucky few; the small number of those that were hand-copied usually belonged to monasteries or wealthy families [Lin92]. A privileged number of scholars was allowed access to these, and it was not until the 1500s that it became commonplace for encyclopedias to be common companions to scholars.

It was, in fact, during the 15th century that the term encyclopedia was coined, on the basis of two Greek words *enkyklios paidea* being wrongly connected into one by scholars of that era [Lin92]. The words was introduced into the English language by the English physician and philosopher Sir Thomas Browne [SSL12] who included it in the preface to his work *Pseudodoxia Epidemica* [CC79].

The first encyclopedias to be written in the form we are familiar with today were those of 18th century scholars, such as Chambers' *Cyclopaedia* [CC79] or the world-famous Encyclopaedia Britannica[1]. The single scholar most commonly credited with the introduction of the alphabetical style and reference book format is John Harris, the author of the *English Dictionary of Arts and Sciences* (1708) [NC].

[1] `https://britannica.com`. Visited: December 20, 2016

With the arrival of the 19th century, the production of encyclopedias continued to flourish, including now not only the English, French and German language, but a variety of other languages as well [Wikg]. It was during this time that the Russian, Swedish, Danish, Norwegian and many other nations saw the first encyclopedias published in their language [Wikh].

2.2 The modern encyclopedia

During the second half of the 20th century, encyclopedias ultimately became household items that were not only affordable, but also easy to find and popular among the general public [Wikg]. Their greatest importance lies in the synthesis of existing knowledge in specific fields, as well as classification of such information for easier access and reference [ND]. Today, the "old" encyclopedias are often used as firsthand sources about the worldviews and attitudes of particular societies (or fractions of society) during a particular time [Lin92].

The emergence of the modern encyclopedia also introduced the model of employing a high number of employees – text writers – that are experts in their particular field [Wikh] thus enabling for encyclopedias covering general topics to provide sufficient expertise in each of the fields it addresses so as not to be discredited. This, of course, is not to say that there were not encyclopedias that address a narrow topic or field: in fact, it is estimated that today there is at least a single encyclopedia covering every academic discipline, including very narrow ones [atW].

2.3 The digital encyclopedia

Once personal computers became commonplace, the traditionally printed medium began to transition to a digital format. The first instance of such an attempt was the 1993 edition of Microsoft's *Encarta* [Wei], that has no printed edition even today. It was stored on a CD-ROM, and, apart from text entries, comprised of high-quality images, videos, and audio files [Wei]. Microsoft continued producing this multimedia edition of an encyclopedia until 2009, when the product was finally discontinued [Wei].

The reason for discontinuing the product may be related to the emergence of a somewhat similar phenomenon, which is the free online encyclopedia.

2.4 The free online encyclopedia

Throughout the 1990s, there were a number of projects and initiatives aimed at establishing a free, open-source, user-generated generic knowledge bank or encyclopedia. Some

notable examples include *Everything 2*[2], *Open Site*[3], and *GNUPedia*[4] and *Interpedia*[5].

The idea between the creation of such a medium is well represented in a message written on the forum Listserv by Interpedia's conceiver, Rick Gates:

> "The more I thought about this, the more I realized that such a resource, containing general, encyclopedic knowledge for the layman, would be an important tool for some types of research, and for the Net.Citizenry in general.
>
> Ahh.. but what about contributors... where will you find authors to write the short articles you need? Well, I'd first have to start out by finding some way of communicating with an extremely diverse set of people... everyone from linguists, to molecular biologists, from animal rights activists to zymurgists, and from geographers to gas chromotographers. Guess what? :-) The Net provides just such an arena! So I thought about it some more... ... and came to the conclusion that this is a good idea!" [KW08]

Unfortunately, although it attracted a great deal of discussion, the Interpedia remained in the planning stage for approximately half a year, until it finally died [RL12a]. The first similar project that succeeded in its realization, and, more importantly, one that continues to remain stable and thrive even today, is Wikipedia.

2.5 Wikipedia

The English version of Wikipedia (and the only one at the time) was conceived by Jimmy Wales and Larry Sanger in 2001 [Lih10]. The official beginning of Wikipedia's history is generally pinned down on 15$^{\text{th}}$ January, 2001, the date the website was launched [Lih10].

Wikipedia was based on the concept of a *wiki*, which was introduced by Ward Cunningham in 1993 [Lei14a]. A *wiki* is an online page created through the use of a *wiki* software or engine [Lei14a]. Although there are now many different wiki engines available, their unifying characteristic is that their structure enables text to be added and edited directly from the web browser [Lei14a].

Wikipedia began as an additional tool for its predecessor, Nupedia. The Nupedia was the first online encyclopedia that relied on volunteer contribution. However, the content was to be produced through exclusively expert contribution, and an elaborate process of peer-review [Lei14a].

It became clear to Sanger, one of Nupedia's employees who was later to be the founder of Wikipedia, that Nupedia's structure would take too long to develop content, and he

[2] `https://everything2.com`. Visited: December 20, 2016
[3] `https://en.wikipedia.org/wiki/Open_Site`. Visited: December 20, 2016
[4] `https://www.gnu.org/encyclopedia/encyclopedia.html`. Visited: December 20, 2016
[5] `https://en.wikipedia.org/wiki/Interpedia`. Visited: December 20, 2016

began to search for an alternative model. After attempting to embed the *wiki* structure into Nupedia, Wales and Sanger agreed that this separate website was to be housed on a different domain and therefore become, to a certain degree, detached from the Nupedia project [Lei14a].

The project skyrocketed, and it took slightly less than a month for Wikipedia to reach its 1,000 article milestone; it also took less than two months for the first non-English Wikipedia to be introduced: the German Wikipedia was created on 16th March 2001 [Wikh]. Figure 2.1 shows Wikipedia version by sizes.

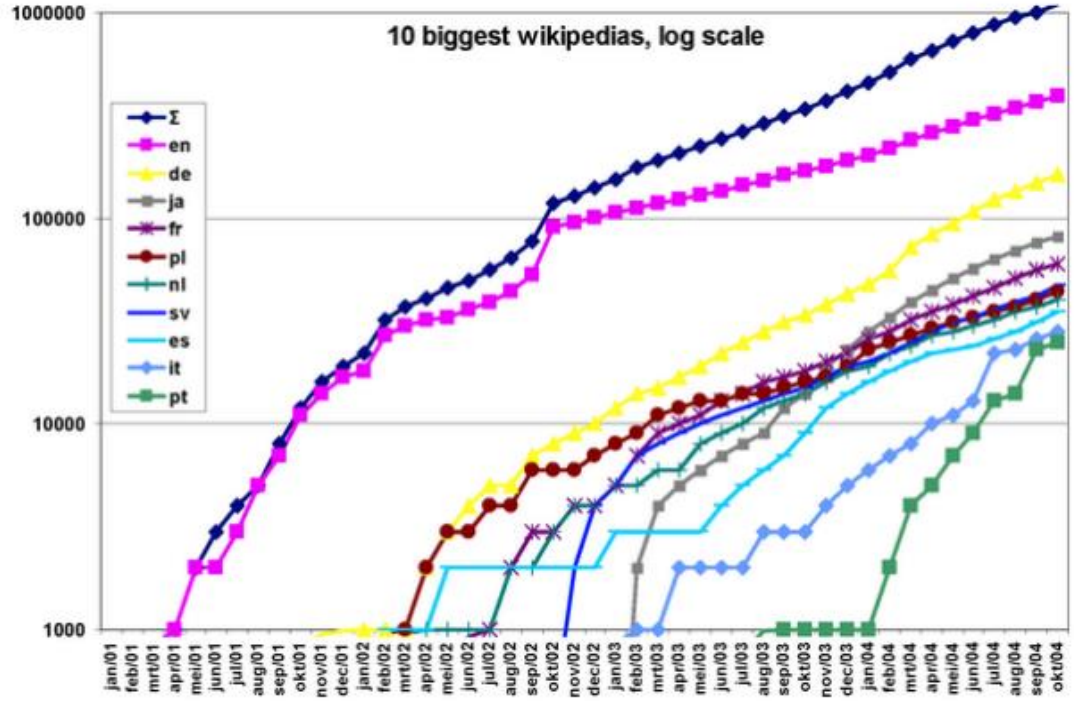

Figure 2.1: Wikipedia versions by sizes

2.6 Wikipedia today

According to *Alexa Internet*, Wikipedia is today the 6th most visited website on the World Wide Web, coming after search engines Google, Baidu and Yahoo, and social networks YouTube and Facebook [Ama].

The English Wikipedia today has approximately 5.2 million articles, approximately 40 million pages(see Figure 2.2), and nearly 29 million users. Internationally, it exists in 294 languages, 13 of which have more than one million articles each. Its English version is the most visited one, receiving more than 50% of the entire website's traffic [Wikf].

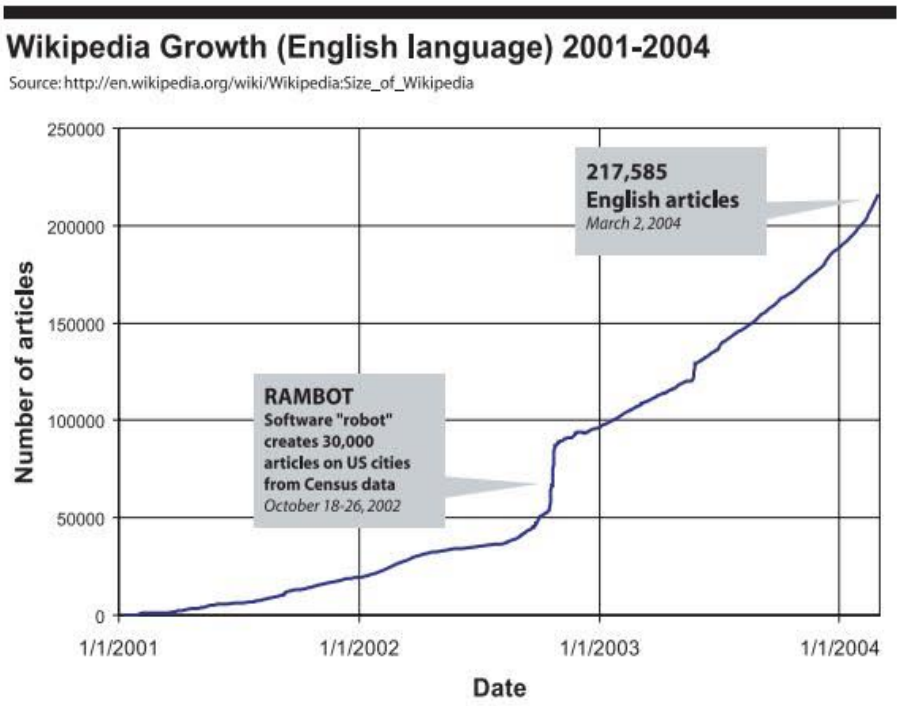

Figure 2.2: Wikipedia Growth (English language)

2.7 The "Yugoslavian" Wikipedia

The conception of Wikipedia in the countries of former Yugoslavia took place soon after the launch of the English Wikipedia website. Towards the end of 2002, the Bosnian, Slovene and Serbo-Croatian Wikipedias were founded, followed by the Serbian and Croatian Wikipedias in February of 2003 [Wikb].

The largest of all of them is the Serbo-Croatian version, with slightly more than 400,000 articles and nearly 50 million edits as of January 2017 [Wikb]. As such, it is the 21st largest Wikipedia version in the world. The Serbo-Croatian version is followed by the Serbian Wikipedia, ranking as 28th largest in the world with slightly less than 350,000 articles and almost 13 million edits (id.) The Croatian and Slovene version rank at 42nd and 45th place respectively, both hosting more than 150,000 articles and around 5 million edits (id.) The Bosnian version ranks lowest at 65th place with around 73,000 articles and around 3 million edits (id.). The total number of active users[6] for all these Wikipedias is 2,337. When it comes to traffic they receive, the Serbo-Croatian version counts 382,681 users, the Serbian version 195,011, the Croatian version 174,629, the Slovenian version 152,003, and the Bosnian version 93,511 as of January 2017 (id.)

[6] Active users are defined as "registered users who have made at least one edit in the last 30 days. [Wikb]

The Wikipedia community in the former Yugoslav region has not been very strong. Its strongest point is considered to be the Serbian wiki community, which facilitated 253 meet-ups as of September, 2013 [Wikc] The members of the community, whose official meeting space was the Belgrade Youth Center, founded the local chapter of the Wikimedia Foundation in December, 2005 (id.). Originally, its name was Wikimedia Serbia and Montenegro, and was renamed Wikimedia Serbia after the breakup of the Serbia and Montenegro union. The local chapter hosted a total of four regional conferences for Southeast Europe.

When it comes to controversies surrounding the Wikipedia of former Yugoslav countries, the Croatian version attracted international attention towards the end of 2013[7], when it was discovered that many of its articles embodied a fascist worldview, and there was a strong current of historical revisionism, particularly when it comes to the age of the Ustaše regime in Croatia.

2.8 Wikipedia's philosophy

The philosophy that guides Wikipedia's content creation is by many attributed to the FOSS (Free Open Source Software) Movement of the early 1990s, founded by Richard Stallman. Tired of the corporate model of software creation (also known as the proprietary model), which prevented its users from doing anything with it other than using it, Stallman fought for 'open-source' software, built on the same model of voluntary contribution that Wikipedia is based upon [O'S09]. In fact, one of the founders of the website, Jimmy Wales, confirmed the role of the Free Open Source Software Movement in the creation of Wikipedia in an interview in July 2006:

> "I spent a lot of time thinking about online communities and collaboration, and one of the things that I noticed is that in the humanities, a lot of people were collaborating in discussions, while in programming [...] they were working together to build things of value. [RL12b]"

As such, Wikipedia has served and continues to serve as a complete antidote to the traditional encyclopedic medium, whose editorial processes are strictly controlled and publishing cycles take a long time, as opposed to the ability of Wikipedians to instantly react to new breakthroughs, news and even "article vandalism" which is often taken care of within minutes [Jem14a].

2.9 The five pillars

Wikipedia's philosophy and vision are best reflected in the *five pillars* that constitute it, as outlined on the Wikimedia website [Wikd]. The *first pillar* refers to the website's

[7] https://www.dailydot.com/layer8/croatian-wikipedia-fascist-takeover-controversy-right-w
Visited: December 21, 2016

encyclopedic character, which "combines many features of general and specialized encyclopedias, almanacs, and gazeteers [Wikd]" but is not, on the other hand, "a dictionary, newspaper, or a collection of source documents [id.]." The first pillar, in short, clearly states the founders' purpose in creating the Wikipedia project, which "aims to create high-quality digital information products through the participation of large numbers of contributors, mostly volunteerse [Oko09]."

The *second pillar* stresses the importance of the neutral point of view and provides a basic outline for a balanced, acceptable Wikipedia article, noting that articles should "document and explain major points of view, giving due weight with respect to their prominence in an impartial tone [Wikd]." The second pillar also touches upon the importance of utilizing reliable sources in the creation of articles, emphasizing that "[articles] must strive for verifiable accuracy [...] especially when the topic is controversial. [Wikd]. As [Jem14a] notes from his own experience as a Wikipedia admin for many years, "[the] NPOV [...] is among the strongest norms of editing and one of three core content policies of Wikipedia [Jem14a, pp. 20]." [Jem14a] notes verifiability and "no original research" as the other two core content policies of the website [Jem14a, pp. 20].

Wikipedia's *third pillar* relates to its open-source character. It stresses that Wikipedia is "free content that anyone can use, edit and distribute [Wikd]" and emphasizes the concept of licensing the content of Wikipedia to the public, thus making it impossible for any author or editor to claim ownership of their work, or object their work being modified or distributed further. According to Wikipedia's policy on article ownership, "no one, no matter how skilled, and regardless of their standing in the community, has the right to act as if they are the owner of a particular article [Wikd]."

The *fourth Wikipedia pillar* is upholding the values of civility and respect within the Wikipedia community, particularly in the face of disagreement. As stated by [Jem14a], "one of the most important behavioral rules is assuming good faith (ASG) [Jem14a, pp. 19]," which translates into maintaining a respectful, open attitude towards any other member of the community regardless of indications that what was said (or done) could have been driven by negative intentions. Even in the face of disagreement or controversy, Wikipedia's policy strongly leans toward discussion and reaching a consensus, as opposed to open conflict. When it comes to dispute resolution, Wikipedians favor a consensus born from a discussion more than democratic means such as voting and polls [Jem14b], and believe that "having the option of settling a dispute by taking a poll, instead of the careful consideration, dissection and eventual synthesis of each side's arguments, actually undermines the progress in dispute resolution [Wikd]."

Finally, Wikipedia's *fifth pillar* underlines the project's openness. Although there are many policies, norms, guidelines and straightforward rules, "[they] are not carved in stone; their content and interpretation can evolve over time [Wikd]." Wikipedians are aware of the evolving nature of their community, and as such understand that the synergic character of the project may often induce changes or cause unforeseeable situations which may in turn naturally alter the structure and policies of the website.

- The NPOV

  The *Neutral Point of View* or *NPOV* policy of Wikipedia was first introduced in February, 2002 [Lei14b], and is defined by the organization as a core principle that translates to "representing fairly, proportionately, and [...] without editorial bias, all of the significant views that have been published by reliable sources on a topic [Wikm]." Although the policy stipulates that articles should be written in such a manner from the outset, the collaborative spirit of Wikipedia in which any anonymous or registered user can make a change in an article contributes to neutrality. Even if an article is not initially written from a neutral stance, the Wikipedia community relies on its visitors to correct this error. (S[Shi08] has the following observation regarding this phenomenon:

  > "In a system where anyone is free to get something started [...], a short, uninformative article can be the anchor for the good article that will eventually appear. [...] many more people are willing to make a bad article better than are willing to start a good article from scratch [Shi08, pp. 121-122]."

  The neutrality of a specific article can always be questioned on Wikipedia through the use of the article's Discussion page. With careful watchers consisting not only of readers, but also volunteers acting as administrators or editors, Wikipedia is known to strive for neutrality in the majority of user interactions. This practice, however, can often take an absurd turn, which is the case with a number of controversial topics or persons on the English Wikipedia; hence, the Wikipedia article on abortion[8] has experienced approximately 12,000 edits since its creation in November, 2001 [Wiki]. Similarly, the article on George W. Bush, the most edited entry on Wikipedia of all time [Poe06], surpasses it by more than 30,000, totaling in approximately 46,000 of edits since its creation in October 2001 [Wikj]. Among Wikipedians, this byproduct of upholding neutrality is known as *edit wars*, a phenomenon that will be addressed later in this research.

- Wikipedia and the participative journalism movement

  Also known as *citizen journalism or produsage*, the term participative journalism is described as content created by "the people formerly known as the audience [Ros06]," an order that challenges the traditional form of journalism (also known as the top-down approach) wherein content is curated by editors for the consumers(see Figure 2.3). It has eliminated the previously understood distinction between information providers and information receivers and significantly blurred the line between consumers and producers of information [KM06].

[8] https://en.wikipedia.org/wiki/Abortion

Type	Sources	Type	Time Scope
Wisdom	Books, academic journals	Research and analysis	Years, decades, centuries, ad inf.
Knowledge	Magazines, encyclopedias	Secondary source	Weeks, months, years
Information	Newspaper, magazines, TV news, news wire	Primary source	Minutes, hours, days, weeks
Data	Stock quotes, sports scores, election results, economic statistics, interviews, press conferences	Live feed	Instantaneous, seconds

Figure 2.3: Journalism traditional sources

With its community-driven approach, Wikipedia is a stellar example of a user-generated information repository that can be considered on par with the Encyclopedia Britannica [Gil05]. As [Ma06] concludes, this approach enables "active participants [to] project their demands through contribution, instead of waiting for someone to ask for what you want [Ma06, pp. 200]." It represents an entirely new model of information exchange on a global level, one that is not closely tied to the interests of mass media, and can therefore cater to anyone's needs and desires. Among the vast number of articles covering various concepts, ideas, persons, political movements and many others, every potential consumer of information can find what they are looking for, or participate in creating it without any obligation for further involvement, requiring only a basic understanding of computers and the Internet. Figure 2.4 shows Wikipedia's number of page view by languages.

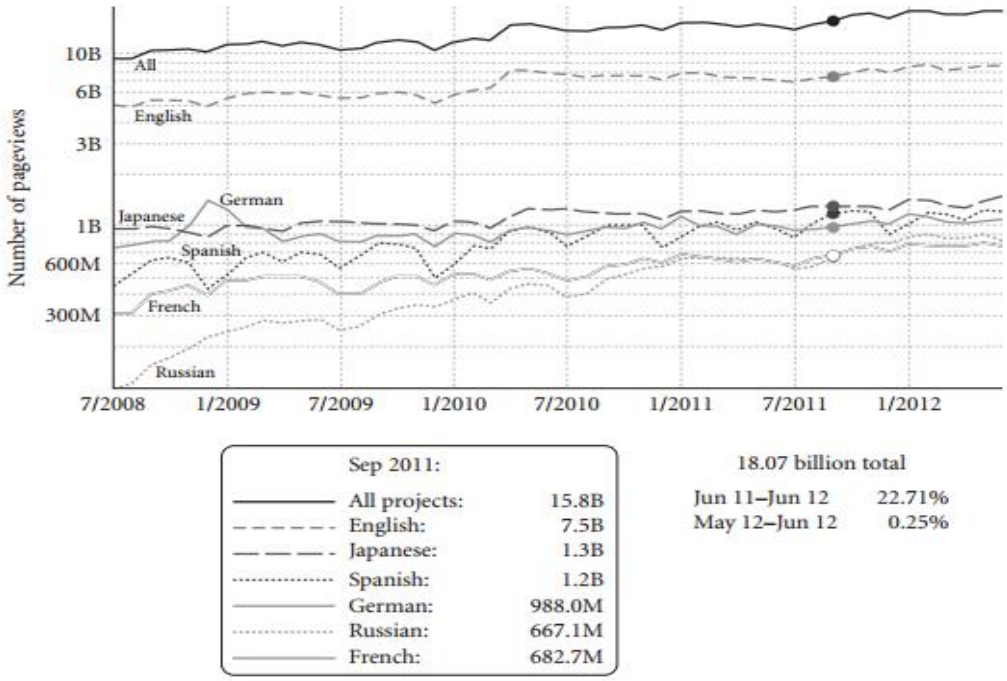

Figure 2.4: Number of page views by languages

- Volunteer participation

  Another equally important segment of the Wikipedia community is that its editors, contributors, administrators and any involved parties choose to participate for no monetary compensation. Furthermore, given the large number of anonymous contributions to Wikipedia articles (24% percent of the total edits [Ma06]) it appears that personal recognition and credit outside the Wiki community seem to play a relatively insignificant role. According to a survey whose results were published in Kuznetsov's *Motivations of Contributors to Wikipedia* (2006), almost half of Wikipedians choose to contribute to the project in order to "educate humanity/raise awareness", while almost a fifth of the participants responded that their primary motivation is the feeling of making a difference. Kuznetsov then concludes that the Wikipedia project has undoubtedly given birth to a tight community strongly driven by their mutual goal of building a high-quality, freely accessible online encyclopedia [Kuz06]. She further notes that the effectively ingrained principles of autonomy, freedom, altruism and collaboration are a strong driving force behind the Wikipedians' dedication to helping this project although their work is rarely rewarded financially. [Oko09] found that two additional strongly motivating factors were fun and the ideology behind the project. Given the relatively high retention rate of contributors [Jem14a], it can also be concluded that a strong sense of

community plays a role in Wikipedians choosing to contribute and stay with the project.

CHAPTER 3

Literature Review

3.1 Epistemological perspectives on the phenomenon of Wikipedia

During the early years of its conception, Wikipedia received little praise from the academic world for its encyclopedic features [Oko09]. The first peer-reviewed study of Wikipedia that was published in a peer-reviewed journal dismissed the concept and recommended Nupedia as a much better alternative [Oko09], while a number of researchers analyzed Wikipedia as an example of the general concept of Wikis [McF]. However, the concept attracted great attention from academics in the field of epistemology[1] or *theory of knowledge*.

Given the nature of this subject field, it comes as no surprise that majority of researchers were concerned with Wikipedia's paradigm-shifting approach to knowledge, which challenged the "classical knowledge model of knowledge creation by experts [Oko09]," and replaced it with that of "knowledge created by consensus of a community of contributors [Oko09]."

[Par08] summarizes the role of Wikipedia in the transforming of knowledge as such, particularly with relation to educational institutions:

> "We do a fundamental disservice to our students if we continue to propagate old methods of knowledge creation and archivization without also teaching them how these structures are changing [...]. No longer is an encyclopedia a static collection of facts and figures [...] it is an organic entity. [...] To

[1] According to the Stanford Encyclopedia of Philosophy, epistemology is the study of knowledge and justified belief, or, more broadly, the study of issues having to do with the creation and dissemination of knowledge in particular areas of inquiry.

> train students in old literacy seems to me to be the fundamentally wrong approach." [Par08]

In the epistemological sense, Wikipedia may be said to stand as a representative of the technological revolution in learning and knowledge creation, acquisition and dissemination. With the introduction of user-generated educational content, particularly when it comes to exclusively academic subjects, the centuries-old structure of academia was severely shaken. In the words of [OPSL15], "Wikipedia has created an environment that questions how scholarship was created, who created it and who owns it, [OPSL15]" and as such, it is one of the biggest challenges to the current system of education and learning.

3.2 Wikipedia as a Social Network

Wikipedia is by no means the most illustrative example of a social network, such as Twitter or Facebook. Nevertheless, the communal spirit of Wikipedia along with a strong sense of identity grounded in being a *Wikipedian* indicate that, to a certain extent, it must be treated as a social network as well as a community. As it has been noted above, it has been concluded that a strong sense of belonging influences the behavior of contributors to Wikipedia; in addition, although there is no direct ownership of articles on Wikipedia, a registered user's input leaves an accessible trace in the page information – since work of great quality will rarely be edited and contributors' names will be clearly visible [Ma06]. Furthermore, [Jem14a] lists other ways in which Wikipedia can act as a social network, mainly through Pages, Discussion and Talk functionalities of Wikipedia. In this respect, recognition by the Wikipedia community can also be a potential motivator in contributing high quality content and remaining active in the Wikipedia community.

3.3 Questions in reliability of Wikipedia

From the onset of Wikipedia, the novelty of the concept of *the people's encyclopedia* and the open disregard for credentials and expertise by the Wiki community have motivated many studies regarding the reliability of the website's information and the success of the model. The openness of the website is, according to [KM06] "one of the strengths of wikis but also one of their major drawbacks [KM06, pp. 193]." Although the lack of a central controlling authority is being cited as one of the main doubts about the Wikipedia model, [Lih10] argues that "authority is not absent, just dispersed, in an online encyclopedia." This is also the case with [Jem14a] who notes in great detail the various roles of volunteers that help the encyclopedia not only stay as accurate and verifiable as possible, but also generally promote the values that motivated Wikipedia's creation in the first place.

The majority of questions related to Wikipedia contribution stem from the lack of understanding of motives of contributors. In an economically driven world, wherein most actions are supposed to have an economic incentive as the primary motivating factor,

researchers struggle to understand not only the reasons behind Wikipedia contribution, but open-source contribution in general. Wikipedia's case poses even more interest given the lack of recognition that is also associated with it along with lack of financial compensation.

One of the most commonly cited studies when it comes to Wikipedia errors and accuracy is that of Jim Giles, published in the journal *Nature*, who concluded that when compared to Encyclopaedia Britannica, Wikipedia is nearly equal in terms of article errors and omissions [Gil05]. Furthermore, an information quality analysis performed by [STGS05] suggests that the importance of quality, embedded as a value in the Wikipedia community, plays a role in Wikipedia being more reliable than expected from an open-source encyclopedia. Finally, [ASW09] introduce the concept of critical mass as being crucial to the quality of Wikipedia articles; as the authors state, by allowing open-source content creation, Wikipedia can reach a *critical mass* of users (that is, increase the number of contributors dramatically), whereby the quality will naturally increase due to the large number of readers and contributors on the website.

Overall, it has been found by majority of researchers that the quality and reliability of Wikipedia articles tends to be at least satisfactory, with *Featured Articles* setting the benchmark for outstanding quality Wikipedia entries. It has nevertheless been noted by [HL08] that both the quality and breadth of entries on Wikipedia appear to be closely related to the most popular interests of its contributors [HL08, pp. 436].

3.4 Criticism of Wikipedia

[DHPW05] name six potential risks when it comes to using Wikipedia:

- Motivation
- Accuracy and sources
- Uncertain Expertise
- Volatility
- Coverage

According to the authors, lack of knowledge about what motivates contributors to create or edit an entry poses a serious risk in terms of the quality of information presented in the article. Although a number of studies indicate that the model of "good faith collaboration" appears to mostly yield positive, constructive edits, and that *edit wars* are easily resolved. Aaccording to [STGS05], obscene edits were sometimes resolved in as short as 1.7 minutes. However, it still goes without saying that *trolls*, as the Wikipedians refer to individuals whose purpose of contribution is ill-meaning, undoubtedly exist and facts found on Wikipedia should be verified further. There have also been instances of

public personas or institutions attempting to remove incriminating information or reverse public opinion through campaigning on Wikipedia (e.g. the example of representatives of U.S. Congress [Jem14c].

Another concern is that there is no accurate way of knowing whether the information presented in the Wikipedia article is accurate. However, in light of Wikipedia policies on verifiability and reliable sources, any claims proposed in an entry should be substantiated by a reliable, verifiable peer-reviewed source, and no original research or opinion should be presented in the article. Furthermore, studies quoted in sections above have concluded that the accuracy of Wikipedia is comparable to that of Encyclopedia Britannica when it comes to the English Wikipedia. Ultimately, readers must use critical thinking skills in approaching the source, particularly if the subject of the article is a controversial one (or, alternatively, an unimportant one). The question of quality evaluation will be discussed even further in subsequent sections of this research.

Closely tied to this concern is that of uncertain expertise. While the expertise of authors and editors can rarely, if ever, be verified, the policies of verifiability and reliable sources are there to prevent dissemination of false or inaccurate information, and inclusion of rumors, hearsay and speculation in the articles. Ultimately, the openness of Wikipedia enables any knowledgeable individual to correct any shortcomings intentionally or unintentionally included in the Wikipedia article.

The openness of Wikipedia, however, raises an additional concern. The information represented in entries is subject to volatility: that is, frequent changes in the content of the article due to disagreement between editors and/or authors. Though this is a legitimate concern, studies quoted above have demonstrated that Wikipedia's NPOV policy along with consensus building tends to lead to objective reporting of issues and clear indication if there are differing/opposing views on a particular subject.

Moreover, authors raise concern about the potential lack of coverage of particular topics, given that the diversity of entries present on Wikipedia is demand-driven and dominated by the interests of its contributors. In the English Wikipedia, the sheer number of articles means that even the topics of the least interest will be represented to a certain degree, but it remains clear that Wikipedia is lacking in some areas (e.g. law, medicine) while being abundant in others (e.g. popular culture) [HL08].

Finally, one of Wikipedia's most vocal opponents is author and entrepreneur Andrew Keen. In his book, "The Cult of the Amateur: How Today's Internet is Killing our Culture (2007)", Keen criticizes the entire Web 2.0 movement, with Wikipedia as one of its most prominent examples. Although the work's reception is divided, it has nevertheless been labeled as a "much-needed Web 2.0 reality check [San] by one of the people considered to be Wikipedia's founders. The author most prominently underlines the *democratization* of media, brought about by the Web 2.0 "revolution" as a concept that is "undermining truth, souring civic discourse, and belittling expertise, experience, and talent. [Pet09, p. 15]. On top of that, Keen draws upon the concept of *ownership* or the lack thereof, as one of the major dangers of a Web 2.0 world. In an illustrative example, Keen notes

that "the value once placed on a book by a great author is being challenged by the dream of a collective hyperlinked community of authors who endlessly annotate and revise it, forever conversing with each other in a never-ending loop of self-references." Throughout the book, Keen lists example after example of ways in which the truth, which he emphasizes as the most important currency of humanity, was distorted, misinterpreted or blatantly ignored on the Internet, and the dangers that such behavior poses to the future of the entire world. He raises legitimate concerns about the lack of antidefamation and libel laws which are an indispensable part of traditional media structures, and presents real-life examples of situations where these laws were openly disregarded and broken. As a conclusion, Keen suggests alterations of traditional media business models to accommodate for changes brought about by Web 2.0, as seen in the example of the British newspaper Guardian. He calls that *digital utopianism*, which is according to the author an omnipresent force in Silicon Valley, be exchanged for digital pragmatism: an understanding that digital technology has tremendous benefits, but needs to be kept in check with rules and regulations, since "we [humans] are easily seduced, corrupted, and led astray." The author finishes off with a strong point that,

> "Our real moral responsibility is to protect mainstream media from the cult of the amateur. We need to reform rather than revolutionize an information and entertainment economy [...] Once dismantled, I fear that this professional media—with its rich ecosystem of writers, editors, agents, talent scouts, journalists, publishers, musicians, reporters, and actors—can never again be put back together. We destroy it at our peril." [Pet09]

Ultimately, it is important to analyze the criticisms described above in light of the topic of this thesis. Given that this research tackles comparatively smaller language editions of Wikipedia (Serbian, Bosnian, Croatian and Slovenian), the issues outlined by the researchers will be kept in mind and analyzed in terms of Wikipedia sites specific to these languages. Seeing as the number of users (both readers and contributors) is much lower in these language-specific Wikipedias, this is especially important keeping in mind the argument of critical mass to which many researchers attribute Wikipedia's success and reliability of its content.

3.5 Empirical studies of the Wikipedia phenomenon

As an innovative and unique approach to encyclopedic knowledge, and a challenge to the traditional knowledge formation and dissemination model in a broader sense [Eij10], Wikipedia has attracted much attention from a variety of angles in the academic world. The majority of empirical studies regarding Wikipedia tackle the question of content quality, which is, in turn, also the primary concern when it comes to the free, volunteer-based model the site utilizes. In addition, [PZA06] analyzed the cultural aspects of collaborative authoring and its impact on Wikipedia content, while Ma (2006) broadly analyzed the economic, cultural, and social implications of the website. Furthermore,

researchers such as [KM06] and [KK08] tackled the concept of collaborative authoring and wisdom of the crowd. Finally, [VWKvH07] analyzed the administrative structure that serves as the backbone of Wikipedia.

When it comes to quality of content, studies range from those analyzing the existing content on the site, to those proposing different methods and models that will be able to handle the task of quality assurance in Wikipedia. The conventional wisdom and popular perception is that Wikipedia articles represent a sea of errors, inaccuracies, and are not fit for information gathering and knowledge acquisition. However, study after study has proven that the quality of Wikipedia content, particularly in the English Wikipedia, can be measured against that of traditional encyclopedias, such as the Encyclopedia Britannica [Gil05]). [Lih10] concludes that the sheer number of visits and edits to the majority of general interest articles warrants quality and reliability of information, meaning that articles which receive more traffic and edits are, ultimately, of better quality. A similar conclusion has been reached by [LST10], by ranking Wikipedia articles for quality through evaluating article length, revision numbers and author reputation measured by their editing history.

Before providing an overview of empirical studies, it is important to note that the first step towards quality assurance practices is embedded in the Wikipedia philosophy itself. As noted in the previous section, the five pillars of Wikipedia provide a foundation for the set of guidelines that authors and editors alike are reminded to adhere to on a regular basis. In a study by [VWKvH07], the role of Wikipedia's *Talk* pages, meant to be for discussions on a particular article, is thoroughly analyzed. The authors conclude that these pages have an indispensable role in quality assurance, but also in "fostering civil behavior and community ties." Furthermore, these authors conclude that "administrative and coordinating elements seem to be growing at a faster pace than the bulk of the articles in Wikipedia [VWKvH07]," which is further reinforced by the findings of [KK08], who have found that "nearly 40% of all edits in Wikipedia involve indirect work, e.g. communication, consensus building, development of policies and procedures." [VWKvH07] go on to conclude that "*Talk* pages serve a variety of important functions in the maintenance of articles, ranging from strategic planning of editing activities to the enforcement of Wikipedia policies and conduct guidelines."

On top of that, Wikipedia's success is largely attributed to the positive impact of harnessing "the wisdom of crowds." The term "wisdom of crowds" is thought to be introduced as early as the beginning of the 20^{th} century by Galton, who conducted an empirical study that averaged the estimations of the weight of an ox at a county fair provided by a number of independent observers, and concluded that the average of these judgments was more accurate than individual judgments of experts [Gal07]. The Wikipedia, among many other representatives of Web 2.0, operates under similar assumptions. On the other hand, critics of Wikipedia, often assert that "Wikipedia is nothing more than an unusually unvarnished avatar of the marketplace of ideas, in which there is no evidence, only hope, that good ideas will drive out bad [Lei14a]". Andrew Keen, one of Wikipedia's most notable opponents, describes it as "the blind leading the

blind—infinite monkeys providing infinite information for infinite readers, perpetuating the cycle of misinformation and ignorance [San]." He attributes this to the emergence of the "cult of the amateur," which "worships the creative amateur: the self-taught filmmaker, the dorm-room musician, the unpublished writer [...] and suggest that everyone—even the most poorly educated and inarticulate amongst us—can and should use digital media to express and realize themselves." This dimension of the problem is summarized in the study by [Eij10], who states that "the real concern is about the *form* of Wikipedia as a new knowledge construction process and, by extension, as the iconic representative of new and uncontrollable Web 2.0+ collaborative knowledge production environments," adding also that "for many academics, Wikipedia has become a symbol of resistance against traditional academic power-knowledge arrangements." He concludes that Wikipedia's controversial role and standing in the academic world is largely owed to its potential to drastically transform higher education in terms of research methods and pedagogic practices, therefore threatening institutionalized learning (id.).

Nevertheless, the majority of the vast body of work exploring Wikipedia is not dedicated to taking sides or dismissing it is a "scrappy, chaotic, dilettantish, amateurish, upstart free-for-all [San]," but rather to admitting the undeniable influence this website holds and, by extension, suggesting ways to improve it through establishing methods, practices and models that will be enable both its readers and administrators to judge its quality and act upon their judgment. For the purposes of this study, these studies are particularly important, specifically since there is no work done on the Wikipedia language versions that this research tackles.

[HL08] analyzed the English Wikipedia in terms of topical coverage as a potential marker of content quality. Their research is based on a random sample of 3,000 English Wikipedia articles that were then coded according to the Library of Congress coding system and compared to books in print. The findings of the study clearly indicate areas in which Wikipedia falls behind in relation to printed material, but also those in which Wikipedia has richer content than printed books. Among the many categories analyzed, the most significant findings are that of Military, American History, General History, Political Science, Physical Science, Music, and Geography, in which Wikipedia leads in terms of amount of content, as well as Social Sciences, Technology, Philosophy, Education, Literature, Law, and Medicine, where Wikipedia falls behind. The researchers particularly emphasize the radical lacking demonstrated in the fields of Law and Medicine, both traditionally associated with higher expertise. In the secondary area of their study, these researchers concluded that certain topics, such as Physical Science, and Popular Culture, experience more rapid expansion than, for example, National Poetries or Prosodies. Overall, the study concluded that Wikipedia's coverage of topics is "more limited than that of the printed, expert-created encyclopedia," and that the website is "not well-organized but covers a broad range of topics."

Another approach, utilized by many researchers, was to judge content quality on the basis of editorial history of the article, such as in [WH07] and [KK08].[KK08] used a longitudinal approach to analyze how contributor numbers and the coordination methods

they use influence the changes in article quality within a particular timeframe. In that way, these researches avoided the trap of reverse causation, since higher quality articles tend to attract more editors rather than the other way around, and this can potentially misdirect the results of the analysis. Both studies by [KK08] and [WH07] found that an increase in the number of contributors tends to improve the quality of the article. However, their findings also underline an important principle: a greater number of contributors is only beneficial to article quality if the efforts are efficiently coordinated. In fact, these authors underline that "simply adding more contributors may incur coordination costs and process losses," thereby emphasizing the importance of structured collaboration in the process of utilizing crowd wisdom. Furthermore, they found that improvements in quality were greater when a small number of editors performed majority of the work, with additional contributors playing a less central role, as opposed to evenly distributed work among all contributors. This is linked to performance studies of other types of open-source collaboration arenas, such as the open-source software development field, in which similar mechanisms operate. Therefore, [KK08] provide valuable insight into the mechanics of not only Wikipedia, but any collaborative environment, which is that effective coordination and communication is crucial to successful collaborative work in any field.

A similar approach was employed by [Lih10] , who proposed that the quality of articles be judged on the basis of what he called *rigor* and *diversity*. According to [Lih10], rigor refers to the overall number of times the article has been edited, with the hypothesis that "more editing cycles on an article provides for a deeper treatment of the subject or more scrutiny of the content." Diversity, on the other hand, is described by [Lih10] as the overall number of unique edits, supposing that "with more editors, there are more voices and different points of view for a given subject." The articles examined were those cited in the press, and the benchmark metrics for it were based on a general interest topic list derived from the Dorling Kindersley e.encyclopedia print edition. [Lih10] found that his hypothesis, which stipulates that article quality improves with greater number of edits and greater number of unique edits is true. This was particularly effectively demonstrated due to the data sample used in the study: press-cited Wikipedia articles attracted more traffic as a result of being broadcast in public, which in turn contributed to the quality of those articles.

An even simpler approach to measuring Wikipedia's content quality was proposed by [Blu08], who proposed that the number of words in an article is a viable metric to determine the quality of the said article. The justification for using such approach was its simplicity. [Blu08] argues that methods proposed by other researchers are too complicated in the sense that they tend to: 1) require information that cannot be obtained easily, and 2) operate with parameters and results that cannot be understood by the average user of the website. The data sample used in this study consisted of a total of 11,067 articles, out of which 1,554 were featured articles and 9,513 were randomly selected "clean" Wikipedia articles. The 1,554 featured articles were taken as a proxy for quality articles. The results of the study suggest that word count is an accurate predictor of whether an article will be

featured. The author notes, however, that the limitation of this study is the assumption that *featured equals high-quality*, and calls for further research on the subject.

Wöhner and Peters point out a very simple drawback to this approach. If word count is accepted as a method of measuring article quality, the possibility for exploitation is large, since the score can be augmented by merely inserting additional text into the article. Furthermore, they point out that non-featured articles, which are used as a benchmark for low-quality articles are not necessarily low-quality; this benchmark, therefore, can produce misleading results. Hence, they propose an approach that is based on measuring not merely the number of edits and editors and the word count of the article, but one that qualifies different edits according to the time they spend as part of the article. Hence, they divide contributions into persistent and transient, the former defined as those that remain part of the article through multiple edit cycles (periods), and the latter as those that are quickly removed and therefore do not make it to the subsequent version of the article. The article versions are identified on the basis of a one-month cycle, since the authors found that anything less than that provides data too chaotic to be analyzed, while anything more leaves out valuable information and, by extension, valuable insight. By using this approach, the authors were able to pinpoint typical lifecycle-based patterns that are characteristic of high-quality and low-quality articles. For high quality articles, the authors have found that there is much more intensive editing with persistent contributions throughout the entire lifecycle of the article, with transient contributions increasing once the article has reached the end of the lifecycle. This suggests that once the articles reach high quality, there is a tendency to refuse new contributions within the community. Low quality articles, on the other hand, have much less intensive editing, with a maximum of 65 words per lifecycle. Their results were evaluated according to Wikipedia's existing user-based ratings (*Article for Deletion, Good Article, Featured Article*) and were found to be accurate in comparison to the user evaluations.

Wikipedia's content was also studied from the perspective of its most common quality flaws. [AS12] acknowledge the crucial role played by Wikipedia in today's knowledge acquisition, and focus on exploring the *types* of quality flaws, their distribution, and extent among content of the English Wikipedia. In other words, seeing the vacuum in empirical research that aims to pinpoint the type of quality flaws that most often occur, [AS12] set out to identify all types of quality flaws that can occur, where they occur, as well as methods to quantify them. Drawing upon principles of the study of information quality, these two researchers used Wikipedia's embedded clean-up tags to classify 388 quality flaws into 12 flaw types. The results of the automatic mining study reveal that the most common flaws are connected with verifiability of Wikipedia's content, present in 19.46% articles that were studied, along with the fact that almost a third of all articles on the website are tagged with at least one quality flaw. This number, however, may be even higher, as the authors note that, owing to Wikipedia's size and amount of content, it is expected that many flawed articles are not adequately tagged by the Wikipedia community.

Content quality was also examined from the perspective of the Wikipedia community

by [STGS05], who examined the way information quality in Wikipedia is influenced by the discussion mechanisms of the website embedded in the Talk function, as well as the types of roles present on the website, and the concept of Featured Articles. Drawing upon previous research in the field of information quality, the authors identified ten types of quality issues in a data sample that includes both Featured and Random articles data sets. While their findings in terms of quality issue frequency were similar to those of [AS12], an important takeaway from this study is the qualitative analysis of the attitude of the Wikipedia community towards quality assurance. As the authors conclude, "although anyone can participate in editing articles, the results are carefully reviewed and discussed, in ways very similar to open source programming projects." Furthermore, [AS12] propose that the wiki software, which allows "disputing sides to obliterate each other's contributions easily," in fact expands the chances of resolving disputes via consensus as opposed to open confrontation, which is further motivated by the fact that if the community wants to promote a certain article to *Featured*, there has to be agreement about the nomination of the article.

On the other hand, [WL15] have shown that efficient consensus building does not always take place on Wikipedia, particularly in the case of politically controversial subjects. Taking the example of the English Wikipedia entry on acid rain, to whom they themselves were contributors, the authors tracked the edit wars that took place within the article, and decided to analyze this phenomenon from a broader perspective. The study compared articles on three politically controversial subjects with four that are not considered politically controversial, and analyzed the editing history of all those in order to determine if there is any correlation between the editing frequency and the controversy of the topic. Although limited by the sample size, the study demonstrated that the average edit size and number of edits was significantly higher for controversial than non-controversial topics, thus highlighting the concerns of accuracy that are central to Wikipedia's criticism.

3.6 Cross-lingual empirical studies of Wikipedia

Among the body of research dedicated to Wikipedia, a significantly smaller portion is concerned with multi- or cross-lingual studies. The author of this thesis supposes that the reason for this is the number of constraints (physical, technical) that such an attempt would entail. The most comprehensive analysis of Wikipedia that spans across languages is the study by [ZCR15], which analyzes the neutrality in 30 different Wikipedia articles addressing different wars that took place since 1945 across all languages in which the article exists. [ZCR15] performed a sentiment analysis of Wikipedia's content by extracting only those parts of articles that were a clear sentiment expression. The analysis was performed through the use of a lexicon of about 6,800 words that express positive and negative opinion. The authors found that, at least in terms of war-related subjects, Wikipedia's content is not neutral. Furthermore, the study demonstrated that, across language versions of the website, the strength of the sentiment changes (in the case of war, this was mostly found to be related to the involvement of a particular country in

the war described in the article). Finally, they concluded that the varying distribution of sentiment demonstrates how the points of interest regarding a war change across different languages and therefore different peoples.

The differences among Wikipedia language versions have also been demonstrated by [PZA06]. In their study, *Cultural Differences in Collaborative Authoring of Wikipedia*, [PZA06] used the concept of Hofstede's cultural dimension[2] to demonstrate how cultural differences that exist in the real world translate onto Wikipedia and the Internet as a whole. Although not too specific in their conclusions, the authors suggest that understanding and accommodating for cultural differences in online collaborative work can hold an important role in better online collaboration. By using the information from Hofstede's cultural dimension to predict the behavior of members of an online community, [PZA06] theorize that "if we understand the way people behave in online communication, the effectiveness of this communication or work can be increased and misunderstandings and problems may be minimized." Thus, the authors underline an important dimension of communities such as Wikipedia: cultural backgrounds that govern our behavior in everyday life are equally potent in online communities, which is why communities should be designed in such a way to accommodate for those differences.

[KM06], in their analysis of Web 2.0 applications, closely looked at the influence of Wikipedia's multilingual character on both the quality of its content as well as its neutrality. The authors point that "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." The authors concluded, therefore, that neutrality and balance of a Wikipedia article is a direct result of the balance and professionalism that its authors and demography exhibit.

This issue of cultural differences was also addressed by [HG10] in much greater detail in their study *Measuring Self-Focus Bias in Community-Maintained Knowledge Repositories*. They introduce the concept of self-focus, which they define as the occurrence when "contributors to a knowledge repository encode information that is important and correct to them and a large proportion of contributors to the same repository, but not important and correct to contributors of similar repositories." In other words, [HG10] look into a phenomenon that has already been identified in the various analyses of Wikipedia's topical coverage, but taking into account also the degree of prominence of articles that tackle certain topics within the network of Wikipedia's content. Drawing upon the research of [DHPW05] about Wikipedia risks, the authors propose that self-focus is introduced as an additional risk. On the one hand, self-focus plays an important determining role in all of the other identified risks, while on the other, it can be a strong and important shaping factor in itself. Its strength was comically illustrated by Stephen Colbert in The Colbert Report, where he referred to it as "wikiality" – a type of bias wherein contributors to a user-generated knowledge repository recreate the world in accordance with their personal view of it. By analyzing Wikipedia versions in 15 different languages, the authors have

[2] `https://geert-hofstede.com/national-culture.html`. Visited: December 13, 2016

shown that each demonstrates, to a smaller or larger degree, the presence of self-focus in it that the main focus of the content of the website is directed at the country wherein the language is spoken. Out of the 15 analyzed language versions, 12 of them demonstrated that the primary focus of the Wikipedia content was related to the home country of the language, while for three of them, the content related to the home country of the language came in second place.

This topic of research was taken even further by [HG10] in their study The Tower of Babel Meets Web 2.0: *User-Generated Content and Its Applications in a Multilingual Context (2010)*. Although focusing on Wikipedia, this work explored the role of language in the context of user-generated content, particularly the way in which it acts as a barrier for world knowledge diversity. In order to do so, they questioned the global consensus hypothesis, which stipulates that "every language's encyclopedic world knowledge representation should cover roughly the same set of concepts, and do so in nearly the same way [HG10]. The authors took both levels of this hypothesis, the one stipulating that the same knowledge should be covered, and the one stipulating that the knowledge covered should be presented in the same way, and examined its truthfulness through an analysis of 25 Wikipedia language versions. Their findings absolutely dismantled the global consensus hypothesis: the authors found that more than 74% of concepts they analyzed were present only in one language, while only 4.5% of concepts appear in more than six language versions of Wikipedia. They further examined the correlation between the size of the Wikipedia and the presence of the analyzed concepts. Supposing that the lack of coverage may be due to lack of content on other Wikipedias, the researchers analyzed only 3 largest Wikipedias, but the results were not different: 80% of concepts were present only in one language, and 7% were present in all three of the biggest Wikipedias. Ultimately, the researchers found that a mere 0.12% of concept, which equals to 6,966, were present in all of the 25 language versions that were analyzed in the study. When it comes to what the authors refer to as *sub-concept diversity*, which is the second level of the global census hypothesis that stipulates that knowledge should be presented in the same way, the authors found that there is some overlap in the way knowledge is presented. They attribute the differences in terms of knowledge presentation to *self-focus*, since Wikipedia language versions are, once again, more likely to refer to knowledge that is locally-relevant when discussing or describing a particular subject.

Another cross-lingual study that is closer to the research objectives of this study is *Cultural Bias in Wikipedia Content on Famous Persons* by [CH11]. Starting from the findings of [KM06] that Wikipedias in different languages highlight "local heroes" and therefore do not present a balanced view, these two researchers set out to analyze the differences in content of the Polish and English Wikipedias about famous persons. More specifically, these researchers examined the neutrality and balance of articles in both languages by performing a both quantitative and qualitative analysis. Their study is composed of four data sub-sets: famous Polish people in English, famous Polish people in Polish, famous American people in English, and famous American people in Polish, further defined as same-culture and other-culture persons, wherein same-culture persons

are the "local heroes" (Polish-Polish and American-English) and other-culture persons are those that are not a national of the country in whose language the article is written (Polish-English and American—Polish). In total, they analyzed 60 entries, 15 from each nationality-langugage pairing, meaning 30 individuals from each of the countries and 30 articles in each of the two languages. On the one hand, the data was analyzed structurally in terms of article length, frequency of article outline, references, external links, lists, photographs and sidebar presence. On the other hand, the authors looked at the data thematically, analyzing favorableness of coverage as well as inclusion of personal information, nationality, education, controversy, and political ideology. Finally, the authors focused on the previously defined controversial aspects of the famous persons' lives, and analyzed whether such information was included or omitted from the entries in both languages. The findings of their research were mixed. To begin with, the author concur that their findings demonstrated the existence of systematic bias in articles in both language versions, though pointing out that the bias seems not to be the product of intentional deception, but rather an accurate reflection of recent histories of the two countries. Furthermore, when it comes to the hypothesis of 'local heroes' proposed by [KM06], the authors concluded that although partially supported, it appears that the English Wikipedia leads in terms of overall amount and quality of content. According to the authors, these results demonstrate the overall superiority of the English Wikipedia due to its greater user base, and show that "these differences are part of a larger political reality, which is that the United States is a major world power as compared to Poland's more limited influence and local situation." Overall, the authors concluded that users would "get the most information from reading in English rather in their native language, at least on topics covered by both language editions."

[SAFJ09] examined cross-lingual characteristics of Wikipedia through the lens of information quality. Through the analysis of Korean, English, and Arabic Wikipedias, the authors studied the concept of information quality as such, and the difficulties of communicating a single standard of information quality across languages, and, by extension, across cultures. In order to answer this question, the Stvilia, Al-Faraj and Yi formalized quality models used in the three language versions of Wikipedia, and then analyzed their mutual relationships in order to determine whether quality standards are transferrable across different language versions of Wikipedia, and whether different versions that have similar socio-cultural characteristics also utilize similar quality standards. In their findings, the researchers demonstrated that in their sample of three different language versions of Wikipedia, there is significant difference in quality models utilized. The differences were present, on the hand, in what the members of the language-specific community identified as virtues that determine articles of high quality, and, on the other, the importance of particular virtues in upgrade to featured status. Similar findings were reported for the ways in which editing processes are judged. Particularly relevant to this research is the finding that quality measurement practices, which were hitherto very often applied to the English Wikipedia but not to other language versions, were applicable regardless of the language of the website. Hence, quality measurement metrics, such as number of edits, number of unique editors are equally applicable to the English Wikipedia as

they are to other language versions. Finally, when it comes to the research question of socio-cultural similarity between countries and its impact on Wikipedia similarity, the authors' hypothesis was not proven to be true – the similarities among countries in terms of Hofstede's cultural dimensions did not influence the similarities between the different language versions of Wikipedia.

[BB14] looked at the issues associated with cross-lingual Wikipedia collaboration in a very specific example: the contested Republic of Kosovo. By analyzing the processes of editing, collaboration and conflict management in a highly controversial subject, the authors were able to tap into the background of these interactions, which is the identity of the contributors and the influence of this identity on content creation. The Republic of Kosovo, formerly an autonomous province within Serbia, has declared independence from Serbia in 2008. Although recognized by many countries globally, its secession is not accepted by Serbia and a number of other countries. With the majority of its population being ethnic Albanians, and a minority of Serbs living in the northern areas, it has been striving for independence for the past three decades, and is right now under provisional rule of the United Nations, while engaging in less than fruitful discussions with the Government of Serbia. The authors analyzed the articles on this topic in three Wikipedia language versions: English, Serbian and Croatian. Apart from articles, the authors also looked into the Talk pages for the purposes of discerning the motivation behind contributions and pinpointing the identities that are embodied in the editors. The authors concluded that three types of identities are present among the editors of all three language versions. The encyclopedic identity, present among a minority, upholds the values that Wikipedia itself promotes, arguing that the purpose of Wikipedia (or any other encyclopedia) is to provide neutral and balanced information, and warn of the dangers of basing encyclopedic entries on current news and biased media content. The territorial and language identities are different, carrying within them a complex web of socio-cultural predispositions, which the authors did not attempt to dismantle. They have, however, concluded that these identities play a major role in the dynamics of both the articles and the *Talk* pages, and warrant attention as such. Overall, the authors found that the editing discussions do not seem to be aimed at establishing consensus, but rather deepen the conflict, as editors address each other with sarcasm and cynicism, holding directly opposing positions, unwilling to discuss openly. More importantly, they have found that in the Serbian and Croatian Wikipedia, there is no third-party arbitration action (which is the case in the English article about Kosovo), thereby further hindering the possibilities of conflict resolution. Finally, the authors concluded that the dynamics and conflicts occurring in the *Talk* pages and edit wars taking place in the articles on the Serbian and Croatian Wikipedia are a mere representation of what is taking place offline, but underline the importance of keeping in mind such painful, controversial topics as phenomena to watch closely, as they seriously impede the neutrality and objectivity that the Wikipedia community is attempting to achieve.

CHAPTER 4

Methodology and data collection

4.1 Data sample

The dataset for this study is based on the largest articles for famous persons from Serbia[1], Croatia[2], Bosnia[3], Montenegro[4] and Slovenia[5], in English, Serbian, Croatian, Bosnian, and Serbo-Croatian languages, accessed through articles from the English Wikipedia listing famous persons from these countries. From the lists accessed from the English Wikipedia, three largest articles for each of the five countries in each of the six languages will be selected and analyzed in terms of *key performance indicators*. The largest articles will be chosen on the basis of length in bytes, retrieved from Wikipedia's Article Information.

pages.

In total, the sample will contain 90 articles. Out of the 90 articles, 18 will be dedicated to famous persons from each of the countries listed above, grouped by language into six groups(see Table: 4.1). Therefore, the total number of famous persons will be three from each of the countries.

[1] https://en.wikipedia.org/wiki/List_of_Serbs. Visited: October 15, 2016
[2] https://en.wikipedia.org/wiki/List_of_Croats. Visited: October 15, 2016
[3] https://en.wikipedia.org/wiki/List_of_Bosnian_and_Herzegovinian_people. Visited: October 15, 2016
[4] https://en.wikipedia.org/wiki/List_of_Montenegrins. Visited: October 15, 2016
[5] https://en.wikipedia.org/wiki/List_of_Slovenes. Visited: October 15, 2016

Country	1st Name of person	2nd Name of person	3th Name of person
Serbia	Jelena Jankovic	Zivojin Misic	Stepa Stepanovic
Croatia	Franjo Tudjman	Stjepan Radic	Aloysius Stepinac
Bosnia	Ivo Andric	Zdravko Colic	Emir Kusturica
Montenegro	Peko Dapcevic	Petar II Petrovic Njegos	Slobodan Milosevic
Slovenia	Robert Kranjec	Anton Martin Slomsek	Zoran Music

Table 4.1: List of famous persons from each of the countries per length

The data sample will contain the following information: length of article in bytes, availability in other languages, references, external links, categories, total number of views between November 1st 2016 and December 31st 2016, number of average daily views, and number of edits. This information will be used as *key performance indicators*, a concept which will be further described below, in order to judge the quality of the articles and compare them across languages.

The motivation behind the choice of famous people comes from multiple reasons. To begin with, in order to ensure the greatest overlap of data, the researcher has chosen to use famous persons as the research subject. This topic, particularly keeping in mind the mutual history of the countries whose language versions are studied, is expected to be at least somewhat represented in each of the language versions. Furthermore, biographies of famous persons are a common encyclopedic entry, and are therefore expected to be one of the most easily accessible topics in any encyclopedia, Wikipedia included. Moreover, through the selection of famous persons' biographies for analysis, the researcher, who comes from Serbia and is familiar with Yugoslav history, was able to judge that all of the persons included in the lists accessed from the English Wikipedia are indeed notable in the history or current affairs of the country that they come from. Finally, given the relatively small global significance of each of the countries from the former Yugoslav region, the researcher chose a subject that would also warrant at least some global recognition, therefore ensuring that articles about the persons from the sample are also available in the English Wikipedia. Keeping in mind the overall higher quality standards in the English Wikipedia, demonstrated many times over in previous studies of Wikipedia, it was important to try to minimize the number of articles on famous persons that are not included in the English Wikipedia, in order to have a provisionary benchmark, particularly for qualitative analysis.

4.2 Methodology

4.2.1 Wikidata

Wikidata is another project of the Wikimedia foundation, which, unlike Wikipedia, is a data repository that collects "structured data to provide support for Wikipedia, Wikimedia Commons, the other Wikimedia projects [Wikk]. Like Wikipedia, it is free

and published under the Creative Commons licence. However, unlike Wikipedia, it is a data repository, meaning it collects only data in a structured form, and therefore can be reused and even understood by computers. The advantage and motivation behind choosing Wikidata for this study is that it is not tied by language, which allows the data required to be accessed to be seen as it appears on different Wikipedia language versions simultaneously. Furthermore, it is automatically kept updated, which means that if a change occurs in a Wikipedia article in a particular language, this change is automatically tracked in the Wikidata page for the topic of the article. It is partially fed information by bots, but also maintained and improved by Wikidata editors, whose administrative role is to define guidelines and rules of management and content creation for Wikidata [Wikk].

Wikidata information about a particular topic is housed on a Wikidata repository. This can be considered the equivalent of a homepage for the concept in question, providing all information from all Wikipedia language versions through the system of *statements and properties*. A *statement* is, basically, a recording method for Wikidata that allows the creation of information about a concept through connecting with its *properties*. Depending on the type of concept represented, properties can be anything from occupation (if the repository is about a person) through population (if the repository is about an inhabited geographic location), to color (if the repository is about an object). The properties are expanded through sources, ranks, and qualifiers, allowing for the data to be contextualized and explained further, thereby improving the quality information provided by the repository. Another strong quality of Wikidata is its interlinking. Through the abovementioned properties, *items* (concepts) can be linked to Wikidata repositories describing the particular property.

Ultimately, the aim of Wikidata is to provide a centralized overview of all information available on Wikipedia, regardless of the language in which it was written, including even conflicting and contradicting data. Because it is a secondary source, a Wikidata repository will always link to the source of the information included in the repository, therefore allowing for easy reference and evaluation of the information, and serving as an indispensable tool in research.

For the purposes of this research, Wikidata will be used to scrape information about the articles on famous persons from former Yugoslav republics. By allowing the researcher to access quantitative information about the articles, Wikidata will provide all the necessary knowledge required for this thesis.

SPARQL Query Service

The data for this thesis has been collected from Wikidata, through the use of two tools: the Wikipedia Sandbox API, and the Wikipedia SPARQL query service. Both of the methods are pre-developed functions embedded in the structure of Wikidata, allowing a user to retrieve any type of information they require through the use of specific instructions ("actions") and queries [Med].

The data retrieval process consists of using a query "action" followed by the input of particular queries (in SPARQL) that return data sets corresponding to criteria identified in the query. SPARQL is an RDF query language utilized to return data for complex queries, consisting of a subject, a predicate, and an object [Wikl].

Each query consists of several integral parts that, combined together, provide the result of the query. Because of the architecture of the SPARQL query tool, the queries can be very specific or very broad, depending on the construction of the query.

```
#defaultView:Map
SELECT ?item ?itemLabel ?coord WHERE {
 ?item wdt:P27 wd:Q225.
 ?item wdt:P20 ?place.
 ?place wdt:P625 ?coord.
 SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
}
```

Figure 4.1: SPARQL Query of the Map death places for Bosnian famous people

```
SELECT ?item ?itemLabel ?placeLabel WHERE {
 ?item wdt:P27 wd:Q403.
 ?item wdt:P19 ?place.
 SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
}
```

Figure 4.2: SPARQL Query of the list birthplaces for Serbian famous people

The first part of the query is the SELECT instruction, which notifies the query that it needs to select *something*, which will be described in the code that follows. SELECT is followed by `?item`, which is a numerical representation of the concept that the query should look into (see Figure 4.1). According to Wikidata's website, numerical code that is associated with every concept is a way of overcoming differences in language and avoiding that the name used in the query is in English [Wika].

Once the item code is put in, it is followed by `?itemLabel`, which informs the query what the actual name of the concept is in the language of the person inputting the query (see Figure 4.2). In simpler terms, this is a translation from the numerical code to an actual language, and this is the way in which the item's name will be labeled in the results of the query.

This is then followed by the clause WHERE, which defines what goes in the placeholders, which is once again followed by the `?item` placeholder, which defines the actual concept that the query is addressing. Then, a predicate is introduced, which defines what needs to

be looked for, also represented in numerical code. A very common predicate is 'instance of'. Finally, the object of the query is introduced.

4.2.2 Wikipedia Sandbox API

Due to the great scope and multilingual character of this research, the Wikipedia Sandbox API will be used to scrape necessary data from Page Information for particular pages, and therefore equip the researcher with the data to be studied.

After extensive research, the researcher has decided to use the Wikipedia Sandbox API as its tool of choice due to it being the most efficient tool for scraping data from Wikipedia.

4.3 KPIs

KPIs, or *key performance indicators*, are measures used across industries to quantify successes or failures in a particular endeavor [Par07]. Used most often in business, these metrics help managers and executive officers keep track of the progress of projects and stay on top of quality assurance and process management practices.

[Par07] stresses the importance of developing appropriate *key performance indicator measurements*, highlighting that without an adequate choice of metrics to be included, the results tend to be skewed and have little to no effect in improving performance.

In the context of this research, key performance indicators were developed on the basis of previous studies on Wikipedia, taking into account empirically substantiated claims about the influence of certain metrics on article and information quality in the English Wikipedia (see: chapter 3).

Literature on this subject suggest there are alternative, more detailed methods of analyzing information quality in Wikipedia, such as through edit and revision history of specific articles, as well as quality evaluation algorithms developed particularly for this purpose. However, the researcher concludes that implementing these measures falls outside the scope of this thesis, its research questions, and the researcher's expertise.

This thesis will implement the following KPI measurements:

- Length of article in bytes
- Number of languages in which the article is available
- Number of references
- Number of external links
- Number of categories
- Number of views between November and December 2016

- Average daily number of views between 01.11 and 31.12.2016
- Number of edits
- Number of unique editors

In the Data Analysis Section of this work, all of the data derived from the KPIs will be combined in order to arrive at conclusions regarding article quality. The data from KPIs will be used for the first phase of the research, the quantitative analysis.

4.3.1 Length in bytes

Drawing upon the research of [Blu08] as well as others (see: chapter 3), the primary KPI used in this study is the length of the article in bytes. Blumenstock's study suggests that there is correlation between greater word count and quality of the article, thus providing a starting point for judging article quality. When looking at *Page Information* on a Wikipedia page, this number is the first and most prominent metric that the website provides, which further highlights its importance.

4.3.2 Availability in other languages

Although unrelated to previous research on the subject—it has not been studied in any of the previous researches—this KPI is in direct alignment with the subject of *Wiki popularity* that is closely addressed in this research. Furthermore, the author believes that this measurement can be used as a supplementation for the *critical mass argument* that is present in many of the researches regarding Wikipedia, and is often defined as one of the most important factors in Wikipedia's success.

The author stipulates that the existence of an article in multiple languages warrants that the subject of the article is a topic of interest for a wider audience. By extension, this signifies that the article receives a large amount of attention from a wide audience, and plays directly into the argument that with greater number of visitors and contributors, the quality of an article improves significantly.

4.3.3 References and external links

These two indicators serve the role of examining verifiability, which is one of the most prominent quality flaws present throughout Wikipedia, regardless of the language of the website. Although no benchmark number of references and external links has been set by the researching community, it is accepted that with a greater number of references and external links, verifiability improves. Since verifiability is one of the metrics used to determine information quality, this KPI will be used as yet another method of judging an article's quality.

4.3.4 Categories

In many of the studies of information quality in Wikipedia, as well as comparative studies of Featured and Non-Featured articles, it has been found that the inclusion of more *Categories* within the article is connected with the quality of the article. Furthermore, keeping in mind that one of Wikipedia's standards of quality is also organization of information, the author has decided to use this measurement as an additional metric for judging the quality of information.

4.3.5 Number of views between November 1st and December 31st 2016

One of the postulates of the *critical mass argument* is the number of visitors to an article, which is correlated with article quality. Because it is expected that an article will improve with a greater number of visitors and contributors, this KPI is used as another method of measuring whether critical mass in terms of the language versions of Wikipedia studied in this research is achieved. Furthermore, this KPI plays an important role in the secondary research question of the thesis, which tackles the concept of *Wiki popularity*.

4.3.6 Number of edits

Similarly, it has been proven by many researchers that article quality improves with greater number of edits. Although this view has been contested due to the existence of the phenomenon of *edit wars*, this research will still utilize this metric as a proven method of measuring article and information quality in the various Wikipedias it is studying.

Previous research on the subject has demonstrated that article quality improves with greater number of edits, which is once again related to the *critical mass argument*, meaning that with the greater number of edits, the bad is "weeded out" and replaced by content of higher quality.

Furthermore, within this argument is the argument of unique editors. In the findings presented in the Literature Review Section of this work, researchers have found that the quality of content in Wikipedia is also determined by a number of unique editors. It has been found that article quality is greatest when the majority of work is performed by a smaller body of editors, with other contributors playing a minor role in the editing process.

4.3.7 Thematic analysis

Once the quantitative analysis is performed, the second stage of the research will comprise of a thematic analysis, looking into the dates of birth and death (if applicable) of famous persons, place of birth and death, and occupation.

The thematic analysis will play a vital role in the research, as it will enable the answering of the secondary research question of this thesis, which contrasts the quality and neutrality

of an article with the "famousness" and global importance of the subject of the article. The discoveries of the thematic analysis will enable the researcher to assess the importance of the famous person whose biography is being judged to quality, as well as if their global fame and relevance.

On top of that, the thematic analysis will allow the researcher to evaluate the topical coverage of Balkan Wikipedias and compare the results to those found in the studies of English Wikipedia (see: chapter 3) which stipulate that Wikipedia has greatest coverage in terms of history, politics, military and popular culture, but lacks in social sciences, law, and medicine.

Ultimately, the findings from the thematic analysis will be combined with the findings of the qualitative analysis using *Cornell University Source Guidelines*, to arrive at definitive conclusions about the relationship between global popularity and quality of Wikipedia articles.

4.3.8 Qualitative analysis

For the purposes of a qualitative content analysis, this thesis will use adapted *Cornell University Source Guidelines*[6] . Given that the guidelines are aimed at evaluating books and journal articles, the author will adapt the guidelines checklist to accommodate for the differences between the printed sources and Wikipedia article. This adapted checklist will retain most of the measurements present in the guidelines, with the exception of those applicable to books/journals, such as the name and reputation of author, which is not an accessible piece of information in Wikipedia.

Each of the articles in each of the languages will be scored against the Cornell University Guidelines. Hence, the qualitative analysis will comprise of an analysis of six articles in total, chosen on the basis of their performance in terms of KPI measurements. Given that each of the questions in the Guidelines list is a "yes or no" question, articles will be scored as follows: one (1) point for a "Yes" answer, and zero (0) points for a :No" answer. The scores will then be added up and articles for one of each famous person from respective countries in all languages will be compared in order to answer the research questions of the thesis.. An article with the score of 3 will be judged as a quality article, while articles scoring less than that will be considered as low-quality articles.

The adapted questionnaire will comprise the following questions(see Table: 4.2):

The questionnaire above is expected by the researcher to provide a most basic overview of the quality of content present in the various Wikipedia language versions. Although this method can undoubtedly be expanded to include more criteria, and even automatized to a certain degree, the author of this thesis judges that this questionnaire is sufficient for a preliminary qualitative analysis, which can provide a starting point for further research. The motivation behind choosing to qualitative analyze one famous person for each of the

[6] `https://www.library.cornell.edu/research/introduction#2Findingbooks, articles, andothermater`. Visited: November 5, 2016

	Questions
1.	**Is the intention of informing an audience clear from the text?**
2.	**Is the level of information appropriate for an encyclopedic article?**
3.	**Is the language of the article objective and free of emotion-rousing words?**
4.	**Is the article organized logically, easy to read, with main points clearly presented?**
5.	**Does the information appear to be well-researched and is it supported by evidence?**

Table 4.2: Cornell University Guidelines Questionnaire

countries is based on the physical limitations of this study, explored further in Section Six of this thesis.

CHAPTER 5

Quantitative data analysis

The dataset that was gathered was narrowed down to three largest articles for each of the countries analyzed. Due to the fact that this comprehensive research is performed by only one person, and keeping in mind the physical and time constraints of the research, the researcher judged that this is the best way to preserve the research questions that drive the thesis on the one hand, and create a viable research design on the other hand.

5.1 Famous Serbs

The three famous persons from Serbia, chosen on the basis of the length of their biographical articles are **Zivojin Misic**(see Table: 5.1), **Stepa Stepanovic**(see Table: 5.2), and **Jelena Jankovic**(see Table: 5.3). A "-" in the Table representing data below indicates a missing article.

	LB	**LA**	**R**	**EL**	**C**	**V**	**DA**	**E**	**UE**
Serbian	382.412	18	139	25	18	10.545	173	7	3
Croatian	675	18	0	0	3	407	7	0	0
Bosnian	2.168	18	0	1	3	240	4	0	0
Slovenian	1.713	18	0	16	27	81	2	0	0
Serbo-Croatian	123.303	18	6	2	9	1.465	24	1	1
English	17.906	18	1	11	33	82	1	0	0

Table 5.1: Zivojin Misic

	LB	LA	R	EL	C	V	DA	E	UE
Serbian	316.244	14	71	13	15	5.676	93	1	1
Croatian	-	-	-	-	-	-	-	-	-
Bosnian	192.860	14	66	9	16	249	4	0	0
Slovenian	1.116	14	0	4	10	35	1	0	0
Serbo-Croatian	5.702	14	0	1	4	679	11	0	0
English	45.146	14	4	7	26	1.710	28	1	1

Table 5.2: Stepa Stepanovic

	LB	LA	R	EL	C	V	DA	E	UE
Serbian	283.197	48	102	161	15	2.291	38	7	6
Croatian	7.210	48	0	1	2	328	6	0	0
Bosnian	-	-	-	-	-	-	-	-	-
Slovenian	-	-	-	-	-	-	-	-	-
Serbo-Croatian	6.073	48	0	2	4	196	3	0	0
English	104.175	48	63	45	35	13.146	216	6	3

Table 5.3: Jelena Jankovic

The data above clearly demonstrates the advantage of the Serbian language version for all Serbian famous people, compared to all other versions and in the metrics of all key performance indicators except average number of categories, thus supporting the *local heroes hypothesis* introduced by Kolbitsch & Maurer (2006).

	ALB	AR	AEL	AC	AV	ADA	AE	AUE
Serbian	327.284	104	66	16	6.171	101	5	3
Croatian	4.280	0	0.5	3	368	7	0	0
Bosnian	193.944	33	5	10	245	4	0	0
Slovenian	2.271	0	10	19	58	2	0	0
Serbo-Croatian	45.206	2	4	6	780	13	0.3	0.3
English	97.777	23	63	31	4.799	82	2	1

Table 5.4: Average KPI metrics for all three famous persons from Serbia in each of the languages

In terms of article length in bytes, Serbian language articles(see Table: 5.4) are drastically longer than articles in all other languages, with 327.284 bytes average length of article, followed by Bosnian and English Wikipedia with an average length in bytes of 193.944 and 97.777, respectively. In terms of average number of references, Serbian is unmatched

by any of the other language versions, with Bosnian and English Wikipedia's once again coming second and third, with 33 and 23 as an average number of references respectively. Presence of external links is most prominent in the Serbian articles, with an average number of 66, followed closely by the English version at 63 references on average, and Slovenia with an average of 10 references per article. In terms of number of categories, English comes first in terms of organization, with an average number of 31, followed by the Slovenian Wikipedia at 19, and Serbian coming third at 16. The Serbian and English Wikipedia also receive the greatest average number of visitors, at 6.171 and 4.799 respectively, followed by the Serbo-Croatian Wikipedia at 780 average views in November and December of 2016. When it comes to the daily average number of visitors, the Serbian Wikipedia articles attracted on average 101 visitors daily, the English attracted on average 82 visitors, with the Serbo-Croatian Wikipedia coming third at almost a tenth of this number, with 13 visitors. The third most visited Wikipedia is the English version, with 10 average daily number of visitors. In terms of editors, it appears that even the Serbian Wikipedia articles on average have merely 5 edits, followed by the English and Serbo-Croatian version at 2 and 0.3 editors on average, respectively. Finally, in terms of unique editors, the Serbian language articles are once again first, receiving on average 3 unique edits, followed by the English language versions at an average of 1 unique editor, and Serbo-Croatian at an average of 0.3 unique editors.

5.2 Famous Croats

The three famous Croats chosen on the basis of article length are **Aleysius Stepinac**(see Table: 5.5), **Franjo Tudjman**(see Table: 5.6), and **Stjepan Radic**(see Table: 5.7). An "-" in the Table presenting the results indicates a missing article.

	LB	**LA**	**R**	**EL**	**C**	**V**	**DA**	**E**	**UE**
Serbian	52.900	24	32	24	13	1.767	29	0	0
Croatian	60.885	24	29	27	7	4.900	80	3	3
Bosnian	46.439	24	12	12	11	112	2	6	5
Slovenian	13.490	24	2	12	15	234	4	0	0
Serbo-Croatian	19.784	24	3	3	12	721	12	0	0
English	115.022	24	53	80	37	5.630	92	6	4

Table 5.5: Aloysius Stepinac

	LB	LA	R	EL	C	V	DA	E	UE
Serbian	29.015	47	19	29	28	2.727	45	1	1
Croatian	78.304	47	30	36	9	12.919	212	1	1
Bosnian	32.249	47	5	4	11	355	6	2	2
Slovenian	1.343	47	2	11	18	324	5	2	1
Serbo-Croatian	38.596	47	39	28	11	1.392	12	0	0
English	96.069	47	19	91	51	14.528	238	20	6

Table 5.6: Franjo Tudjman

	LB	LA	R	EL	C	V	DA	E	UE
Serbian	13.139	24	1	7	13	1.156	19	1	1
Croatian	66.581	24	76	29	11	7.469	112	2	2
Bosnian	-	-	-	-	-	-	-	-	-
Slovenian	1.727	24	3	11	15	112	2	0	0
Serbo-Croatian	4.555	24	0	0	4	545	9	0	0
English	26.257	24	4	27	33	2.913	48	6	5

Table 5.7: Stjepan Radic

The data above demonstrates that for Croatian famous persons, the *local hero hypothesis* is not supported. On the contrary, this dataset seems to demonstrate results equal to those from the study of Callahan & Herring (2008), wherein the English version of Wikipedia was superior to all other languages in the majority of KPIs that were examined.

	ALB	AR	AEL	AC	AV	ADA	AE	AUE
Serbian	31.685	17	20	18	1.883	31	1	0.7
Croatian	68.590	45	31	9	8.429	135	2	2
Bosnian	39.349	9	8	11	234	4	4	4
Slovenian	5.520	2	11	16	223	4	0.7	0.3
Serbo-Croatian	20.978	21	16	9	886	11	0	0
English	79.116	25	66	40	7.690	126	11	5

Table 5.8: Average KPI metrics for all three famous persons from Croatia in each of the languages

When it comes to length in bytes, the findings demonstrate that, on average, content of the English Wikipedia is longer, followed by the Croatian and then Bosnian Wikipedia, with an average of 68.590 and 39.349 average length in bytes, respectively(see Table: 5.8). However, Croatia is the most prominent in terms of number of references, with an

average value of 45, followed by the English articles with 25 references on average, and Serbo-Croatian with 21 references on average. On the other hand, content from the English Wikipedia comes first in terms of number of external links, with 66 on average, with Croatian articles behind it at 31 references, followed by the Serbian version with 20 external links on average. Similarly, the English Wikipedia articles have the greatest average number of categories (40), followed by the Serbian and then Slovenian articles, averaging at 18 and 16, respectively. The Croatian Wikipedia articles still attract the greatest number of visitors, with average views between November and December 2016 equaling 8.429, followed by the English articles at 7.690 average views. The Serbian Wikipedia comes third at a great difference, with 1.883 average views between November 1st and December 31st 2016. These standings are also reflected in average daily average views, with Croatian Wikipedia coming first at 135 average daily views, and the English Wikipedia second with 126 average daily view. The greatest number of edits was evident in the English Wikipedia, with an average of 11 edits and 5 unique editors, followed by the Bosnian Wikipedia at an average of 4 edits and 4 unique editors, and Croatia with an average of 2 edits and 2 unique editors.

5.3 Famous people from Bosnia and Herzegovina

The three famous persons from Bosnia and Herzegovina chosen on the basis of article length are **Ivo Andric**(see Table: 5.9), **Emir Kusturica**(see Table: 5.10), and **Zdravko Colic**(see Table: 5.11).

	LB	**LA**	**R**	**EL**	**C**	**V**	**DA**	**E**	**UE**
Serbian	74.076	75	45	53	23	22.138	363	7	6
Croatian	41.479	75	31	22	9	11.235	184	0	0
Bosnian	40.405	75	23	22	18	2.549	42	0	0
Slovenian	12.578	75	3	18	28	494	8	0	0
Serbo-Croatian	16.622	75	22	24	19	4.607	76	12	3
English	69.112	75	84	19	36	12.164	199	9	4

Table 5.9: Ivo Andric

	LB	LA	R	EL	C	V	DA	E	UE
Serbian	56.286	56	58	69	14	6.891	113	6	6
Croatian	12.269	56	2	17	5	1.900	31	3	3
Bosnian	20.179	56	11	17	10	1.132	19	0	0
Slovenian	5.312	56	2	15	15	340	6	0	0
Serbo-Croatian	19.634	56	12	16	6	1.484	24	0	0
English	53.240	56	77	109	54	29.291	480	13	11

Table 5.10: Emir Kusturica

	LB	LA	R	EL	C	V	DA	E	UE
Serbian	47.049	16	6	19	10	14.727	241	30	11
Croatian	20.228	16	1	5	3	4.107	67	1	1
Bosnian	22.061	16	2	5	6	3.625	55	0	0
Slovenian	3.859	16	0	3	7	533	9	0	0
Serbo-Croatian	25.173	16	1	4	5	1.804	30	0	0
English	24.162	16	9	19	31	11.062	181	22	11

Table 5.11: Zdravko Colic

The data for famous people from Bosnia and Herzegovina indicates, once again, similar findings as those of Callahan & Herring (2008). Apart from average article length, which is greatest in the Serbian Wikipedia articles, the English version is superior to all other versions. In the majority of KPIs, the English Wikipedia is followed by the Serbian version, with the native language version of these famous persons coming third.

	ALB	AR	AEL	AC	AV	ADA	AE	AUE
Serbian	59.137	36.3	47	16	14.585	239	14	8
Croatian	24.659	11	15	6	5.747	54	1	1
Bosnian	27.548	12	15	11	2.435	39	0	0
Slovenian	7.250	2	12	17	456	8	0	0
Serbo-Croatian	20.476	12	15	10	2.632	43	0	0
English	48.838	57	49	40	17.506	287	15	9

Table 5.12: Average KPI metrics for all three famous persons from Bosnia and Herzegovina in each of the languages

On average, the length of Serbian language articles about famous persons from Bosnia and Herzegovina is 59.137 bytes, followed by the English versions of these articles at 48.838 bytes on average, with the Bosnian version behind at 27.548(see Table: 5.12). In terms of

the average number of references, English articles on average have 57 references, Serbian articles have 36.3, and Bosnian articles are equal with Serbo-Croatian articles, with an average of 12 references per article. Furthermore, the English and Serbian language articles link to 49 and 47 external links respectively, while the Bosnian, Croatian and Serbo-Croatian version have an average of 15 external links. In terms of categories, the average number of categories in the English language articles (49) is nearly 2.5 times greater than in the Bosnian one, while the Serbian language versions come second at 47 external links on average. Surprisingly, the English language articles are also the most visited ones, with an average number of visitors between November and December 2016 at 17.506, followed closely by the Serbian language articles at 14.585. The author attributes this to the international recognition that two of the three famous persons from Bosnia and Herzegovina enjoy, Ivo Andric being a Nobel-winning writer, and Emir Kusturica a world-famous movie director. The Bosnian language articles come fourth in terms of number of visitors, with an average of 2.435, with Croatia attracting an average of 5.747 visitors in the two-month period. When it comes to number of edits and editors, the English version is once again the most edited, with an average of 15 edits and 9 unique editors, followed by the Serbian language articles with 14 edits and 9 unique editors. The Bosnian language articles has an average of 0 edits and 0 unique editors for the three articles.

5.4 Famous persons from Montenegro

The three famous persons from Montenegro chosen on the basis of article length are **Petar II Petrovic – Njegos**(see Table: 5.13), **Peko Dapcevic**(see Table: 5.14), and **Slobodan Milosevic**(see Table: 5.15). An "-" in the Table presenting the data indicates a missing article.

	LB	**LA**	**R**	**EL**	**C**	**V**	**DA**	**E**	**UE**
Serbian	88.306	19	47	44	15	23.013	377	16	11
Croatian	14.928	19	4	10	6	2.879	47	0	0
Bosnian	42	19	0	6	5	523	9	0	0
Slovenian	2.032	19	1	12	18	257	4	0	0
Serbo-Croatian	19.859	19	12	15	6	3.735	61	0	0
English	95.339	19	102	37	32	6.012	99	2	2

Table 5.13: Petar II Petrovic - Njegos

	LB	LA	R	EL	C	V	DA	E	UE
Serbian	35.611	10	6	7	27	2.867	47	4	2
Croatian	-	-	-	-	-	-	-	-	-
Bosnian	-	-	-	-	-	-	-	-	-
Slovenian	2.673	10	0	5	23	80	1	0	0
Serbo-Croatian	20.828	10	3	2	24	1.731	28	0	0
English	2.816	10	0	6	25	1.229	20	1	1

Table 5.14: Peko Dapcevic

	LB	LA	R	EL	C	V	DA	E	UE
Serbian	87.834	72	9	34	22	8.954	147	14	5
Croatian	28.336	72	1	4	10	4.114	67	0	0
Bosnian	20.600	72	1	4	7	713	12	0	0
Slovenian	3.792	72	3	13	20	510	8	0	0
Serbo-Croatian	13.700	72	14	16	8	1.870	31	0	0
English	102.548	72	142	79	38	96.748	1.586	26	15

Table 5.15: Slobodan Milosevic

The dataset for famous persons from Montenegro also follows the results of Callahan & Herring (2008). With the exception of average article length, which is greatest for the Serbian language versions of the articles, the English language versions have greater averages for each of the KPIs.

	ALB	AR	AEL	AC	AV	ADA	AE	AUE
Serbian	70.584	21	28	21	11.611	190	11	6
Croatian	21.632	3	5	8	3.497	57	0	0
Bosnian	10.321	0.3	3	6	618	11	0	0
Slovenian	2.832	1	10	20	282	4	0	0
Serbo-Croatian	11.862	10	1	13	2.445	40	0	0
English	66.901	81	41	32	34.663	568	10	6

Table 5.16: Average KPI metrics for all three famous persons from Montenegro each of the languages

The key performance indicators for famous persons from Montenegro(see Table: 5.16) lean in favor of the English language versions of the article. Apart from average article length, where the Serbian Wikipedia articles average at 70.584 bytes and English articles average at 66.901, the English language articles are superior in terms of average number

of references, with an average of 81, followed by the Serbian language version at 21 and Serbo-Croatian at 10 and in all other KPIs. When it comes to average number of external links, the English articles house an average of 41 external links, while the Serbian comes second at 28 on average, and Slovenian language versions come third at an average of 10. Furthermore, the average number of categories is greatest in the English language versions, followed by the Serbian and Slovenian versions, at 41, 21, and 20, respectively. The English Wikipedia articles also attract the greatest number of visitors. On average, 34.663 people visited the English Wikipedia articles in the recorded two-month period, while 11.611 visited the Serbian Wikipedia, and 3.497 visited the Croatian version. Apart from the Serbian and English article versions, no other language versions have any edits, while the Serbian version comes first with an average number of 11 edits and 6 unique editors, followed closely by the English version at an average of 10 edits and 6 unique editors.

5.5 Famous persons from Slovenia

When it comes to Slovenian-origin famous persons, the three chosen persons are **Robert Kranjec**(see Table: 5.17), **Zoran Music**(see Table: 5.18), and **Anton Martin Slomsek**(see Table: 5.19). An "-" in the Table presenting the data below indicates a missing article.

	LB	**LA**	**R**	**EL**	**C**	**V**	**DA**	**E**	**UE**
Serbian	10.714	17	0	2	10	76	2	0	0
Croatian	4.338	17	1	3	1	60	1	0	0
Bosnian	18.617	17	4	6	10	80	2	1	1
Slovenian	50.438	17	69	73	16	1.182	19	2	2
Serbo-Croatian	-	-	-	-	-	-	-	-	-
English	11.077	17	3	4	17	1.459	24	7	2

Table 5.17: Robert Kranjec

	LB	**LA**	**R**	**EL**	**C**	**V**	**DA**	**E**	**UE**
Serbian	-	-	-	-	-	-	-	-	-
Croatian	9.200	10	2	5	3	132	2	0	0
Bosnian	-	-	-	-	-	-	-	-	-
Slovenian	40.994	10	2	23	25	528	9	1	1
Serbo-Croatian	9.312	10	2	5	4	48	1	0	0
English	13.794	10	6	26	32	1.282	21	9	6

Table 5.18: Zoran Music

	LB	LA	R	EL	C	V	DA	E	UE
Serbian	-	-	-	-	-	-	-	-	-
Croatian	3.520	9	2	2	4	70	2	1	1
Bosnian	-	-	-	-	-	-	-	-	-
Slovenian	31.971	9	1	33	24	1.758	29	1	1
Serbo-Croatian	-	-	-	-	-	-	-	-	-
English	3.782	9	2	12	20	388	6	2	1

Table 5.19: Anton Martin Slomsek

In the case of Slovenian famous persons, the *local hero hypothesis* is confirmed. Slovenian articles are superior in terms of KPI results in each of the categories, with an evident lack of coverage of these famous persons in the Serbo-Croatian, Bosnian and Serbian Wikipedia's. Second to Slovenian Wikipedia is the Bosnian Wikipedia, followed by the Serbian version. It is particularly important to note that, out of 18 expected articles, 6 articles in various languages were missing, indicating a strong gap between the information available for famous persons from all the other countries and information available for famous persons coming from Slovenia.

	ALB	AR	AEL	AC	AV	ADA	AE	AUE
Serbian	10.714	0	2	10	76	2	0	0
Croatian	5.686	2	3	3	87	2	0.3	0.3
Bosnian	18.617	4	6	10	80	2	1	1
Slovenian	41.134	24	43	22	1.156	19	1	1
Serbo-Croatian	9.312	2	5	4	48	1	0	0
English	9.551	4	14	23	632	17	6	3

Table 5.20: Average KPI metrics for all three famous persons from Slovenia each of the languages

On average, there is a significant superiority of KPI results for the Slovenian language Wikipedia(see Table: 5.20) in comparison with other language versions. With an average article length of 41.134 bytes, Slovenian language articles are nearly 2.5 longer than Bosnian, which ranks second in terms of KPI results at 18.617 bytes on average, and nearly four times longer than the Serbian language versions, which come third at an average length of 10.714 bytes. Similarly, the average number of references for the Slovenian Wikipedia articles is 24, with English and Bosnian articles coming second at six times less, with an average of 4 references per article. In terms of external links, Slovenian articles on average link to 43 external links, followed by 14 average external links for the English version, and 6 for the Bosnian version. The English version is slightly superior in terms of KPI performance for the average number of categories, with an average of 23 categories, followed by the Slovenian Wikipedia at 22, and Bosnian and

Serbian Wikipedia's at 10. The Slovenian Wikipedia also attracts the greatest number of viewers, with an average for November and December of 2016 standing at 1.156 visitors, followed by the English and Croatian language versions, with 632 and 87 average number of visitors, respectively. Finally, the edit frequency is greatest in the English language article versions, with an average of 6 edits and 3 unique editors, followed by 1.3 for both KPIs in the Slovenian Wikipedia, and 1 edits and 1 unique editor on average for the Bosnian language version of the examined articles.

In order to interpret the two separate trends occurring in the KPI measurement results, it is important to keep in mind the type of recognition and fame the chosen famous persons from each of the countries enjoy. The superiority of the English version of Wikipedia, which disproves the local hero hypothesis, has been demonstrated for all of the articles which are concerned with biographies of *globally recognized individuals*. In the KPI performance measurements for famous Bosnian people, the English language version was superior to all other languages in terms of articles on two globally-acclaimed artists; Ivo Andric is a Nobel-winning author with books translated into many world languages, while Emir Kusturica is a world-acclaimed film director whose movies have made appearances at the prestigious Cannes Film Festival. Similarly, in the case of Croatian famous people, the biographies of two world-known figures, Franjo Tudjman and Aloysius Stepinac, perform better in terms of KPI measurements in the English versions of their biographies than in Croatian. Tudjman, a political leader of Croatia during the recent Yugoslav Wars, is a global political figure, while Stepinac was a Catholic cardinal in the 20^{th} century and is a prominent Catholic figure, having been proclaimed a Saint in 1998. Moreover, in the case of famous people from Montenegro, the ethnically Montenegrin former president of Serbia, Slobodan Milosevic, who was also a prominent figure in the Yugoslav Wars and was trialed in the Yugoslav Court for War Crimes in the Hague performs drastically better in terms of KPI measurements in the English version than in any other one.

However, when it comes to persons without global recognition or fame, particularly when analyzed in the context of average KPI performance for articles about other-country personas, the *local hero hypothesis* still appears to be supported. Below is a breakdown of each of the KPI performances analyzed for each of the language versions and their treatment of famous persons from other countries.

	ALB	**AR**	**AEL**	**AC**	**AV**	**ADA**	**AE**	**AUE**
Famous Serbs	327.284	104	66	16	6.171	101	5	3
Famous Croats	31.685	17	20	18	1.883	31	1	0.7
Famous Bosnians	59.137	36	47	16	14.585	239	14	8
Famous Montenegrins	70.584	21	28	21	11.611	190	11	6
Famous Slovenians	10.714	0	2	10	760	2	0	0

Table 5.21: KPI performance for Serbian Language Wikipedia content on other-country persons

Once again, taking into account the global recognition of two of three famous Bosnians, who are undoubtedly persons of interest in the Serbian community, as well as the famous Montenegrin Slobodan Milosevic who served as the president of Serbia, the data from the Serbian version Wikipedia clearly indicates better KPI performance for *local heroes* than for other-country famous persons (see Table: 5.21). The average length of article for Serbian famous persons, in comparison to other language versions is drastically higher, which is also the case for number of references and external links. However, the data indicates that visitors of Wikipedia—which are a different category from contributors—are more interested in Bosnian and Montenegrin famous people, which can be attributed to the abovementioned global acclaim of these individuals.

	ALB	**AR**	**AEL**	**AC**	**AV**	**ADA**	**AE**	**AUE**
Famous Serbs	4.280	0	0.5	3	368	7	0	0
Famous Croats	68.590	45	31	9	8.429	135	2	2
Famous Bosnians	24.659	11	15	6	5.747	54	1	1
Famous Montenegrins	21.632	3	5	8	3.497	57	0	0
Famous Slovenians	5.686	2	3	3	87	2	0.3	0.3

Table 5.22: KPI performance for Croatian Language Wikipedia content on other-country persons

For the Croatian Wikipedia, the *local hero hypothesis* proof is even stronger (see Table: 5.22). With content on Croatian famous persons leading in terms of every one of the KPI measurements, it is evident that the content on the Croatian Wikipedia is strongly focused on local heroes, with other-country persons receiving little attention. Once again, the only exception, which comes closest in terms of number of views to Croatian famous persons, is Bosnian famous persons. This can, again, be attributed to the world recognition that the two Bosnian persons enjoy. However, when analyzing the data as a whole, the *local hero hypothesis* is strongly proven in the case of Croatian Wikipedia.

	ALB	**AR**	**AEL**	**AC**	**AV**	**ADA**	**AE**	**AUE**
Famous Serbs	193.944	33	5	10	245	4	0	0
Famous Croats	39.349	9	8	11	234	4	4	4
Famous Bosnians	27.548	12	15	11	2.435	39	0	0
Famous Montenegrins	10.321	0.3	3	6	618	11	0	0
Famous Slovenians	18.617	4	6	10	80	2	1	1

Table 5.23: KPI performance for Bosnian Language Wikipedia content on other-country persons

The KPI performance of the Bosnian Wikipedia needs to be contextualized within its geopolitical and demographic position (see Table: 5.23). As it is evident from the table above, there are KPI measurements wherein the Bosnian language Wikipedia performs

better for content related to famous persons from Serbia and Croatia than people from Bosnia. However, keeping in mind the mixed demographics, which include a significant portion of ethnic Serbs living in the Republic of Srpska, which is an autonomous province of the country, as well as a significant portion of ethnic Croats living in the Herzegovina part of the country, it may still be interpreted that the Bosnian language Wikipedia is, to a certain extent, adhering to the *local hero hypothesis*. However, without more insightful data which would be able to pinpoint the demographic structure of the Bosnian Wikipedia community, the author is unable to draw any substantiated conclusions. It is worth noting, however, that the data from the Bosnian Wikipedia still demonstrates a tendency to focus on local heroes, as its content on Bosnian famous persons is performing better in all KPI measurements in comparison to persons from other countries with the exception of famous Serbs and Croats.

	ALB	**AR**	**AEL**	**AC**	**AV**	**ADA**	**AE**	**AUE**
Famous Serbs	2.271	0	10	19	58	2	0	0
Famous Croats	5.520	2	11	16	223	4	4	4
Famous Bosnians	7.250	2	12	17	456	8	0	0
Famous Montenegrins	2.832	1	10	20	282	4	0	0
Famous Slovenians	41.134	24	43	22	1.156	19	1	1

Table 5.24: KPI performance for Slovenian Language Wikipedia content on other-country persons

The *local hero hypothesis* is also proven in the case of Slovenian language articles (see Table: 5.24). In perhaps the greatest discrepancies in the entirety of this research, the average length in bytes of an article regarding a Slovenian famous person in Slovenian language is 43.134, while the second longest are articles about famous Bosnians at almost 5 times less, with an average length of article of 7.250 bytes. The superior performance in terms of KPI measurements is proportionately larger for all of the measurements, with enormous differences between articles about famous persons that are performing second in terms of key performance indicators

5.6 Overall presence of articles regarding other-country famous persons

The data above indicates the findings of measuring the total number of articles about famous persons for each of the language versions and each of the specific countries. With the findings confirming the *local hero hypothesis* noted above, these results seem surprising (see Table: 5.25). The *local hero hypothesis* is, in this case, not proven, since the majority of the language versions feature the greatest number of articles for famous people from Serbia.

In order to interpret these results, the sizes of respective Wikipedia versions and the

	Famous Serbs	Famous Croats	Famous Bosnians	Famous Montene-grins	Famous Slovenian
Serbian Wikipedia	1.686	420	265	189	98
Croatian Wikipedia	986	809	126	112	131
Bosnian Wikipedia	1.121	654	322	98	42
Slovenian Wikipedia	684	183	57	56	234
Serbo-Croatian Wikipedia	1.280	512	156	148	101
English Wikipedia	1.910	915	423	218	267

Table 5.25: Number of articles for other-country famous persons in each of the Wikipedia languages

amount of activity of their community needs to be taken into account. As it was explored in the Theoretical Background Section of this work, the Serbian Wikipedia community is the most active among all of the Balkan region Wikipedias. Furthermore, in terms of size, the Serbian Wikipedia is the second-largest, preceded only by the Serbo-Croatian versions (which was established before the Serbian version, which may explain the difference in size). Hence, it can be speculated that these results suggest that the Serbian Wikipedia community, concerned with sourcing knowledge of local persons throughout the region, is active outside the Serbian Wikipedia itself. However, this hypothesis cannot be confirmed until further examination of contributors, and perhaps the *Talk* pages of the articles of famous persons, in an effort to determine the nationality of the contributors to those pages.

Discussion

The first stage of this research provides a large amount of insightful information that will, on the one hand, be utilized in the subsequent stages of the research, and, on the other, be one of the foundations of judging and discussing article quality in the Wikipedia versions of the former Yugoslav republics.

To begin with, the data derived from the first stage of the research strongly indicates the existence of the *local hero* phenomenon defined by Kolbitsch & Maurer (2006). Through the analysis of the performance of each of the language versions of Wikipedia, it is evident that the difference in amount of information available about famous persons that are native to the country in whose language the article is written is great. In fact, with the exception of Bosnian Wikipedia, whose problematic context has been explained in the analysis of the results of its KPI measurements, all of the country-specific Wikipedia's have demonstrated a dramatically better performance for each of the KPI measurements.

In the case of these Wikipedia language versions, keeping in mind the recent history of the region, the author concurs that the *local hero* prevalence in the content holds greater

importance. The civil war that has plagued the region took place a mere three decades ago, and as such still causes great divide and controversy among the peoples now living in separate countries. Hence, the author believes that, apart from demonstrating the tendency to focus on local heroes, the large discrepancies in the amount and quality of content present in the different language versions may have to do with these conflicts, and, in short, paints a picture of the cross-national relationships and prejudice towards members of other nations that takes place *offline*. Although the scope of this research does not include an analysis of the Talk pages related to these articles, previous work on the subject by Bilic & Bulian (2014) that analyzed the discussions taking place regarding the disputed Republic of Kosovo representation suggests that if the *Talk* pages were to be studied, similar occurrences of conflict (as opposed to consensus building) would be found.

Moreover, the first stage of this analysis has pointed towards the lack of editing and general collaboration taking place in all of the Wikipedias. With the number of edits and unique editors often being zero in all of the different Wikipedia versions, the author concludes that the extensive process of collaboration and consensus building, which is an inevitable part of the Wikipedia community, is severely lacking and rarely taking place in the Wikipedia versions that were analyzed. This may point towards the possibility that, with the number of active users and active contributors, these Wikipedia versions have not yet reached the *critical mass* necessary for the creation of high-quality content that is characteristic of the largest Wikipedia websites, such as the English and Japanese. This insight will be important to keep in mind in the subsequent phases of the research, particularly in the qualitative analysis that will judge the content quality on the basis of the *Cornell University Guidelines*.

In addition, the author has discovered in the first stage of this research that verifiability, i.e. the use of references and external links to substantiate claims made in Wikipedia articles, has not yet been fully adopted by the community of content creators of Wikipedia from the former Yugoslav region. Combined with the abovementioned lack of editing and revising of the articles, this poses a serious threat to the success of Wikipedia in the Balkans. As it has been mentioned in the Literature Review, the problem of verifiability, or lack thereof, has most often been highlighted as the greatest challenge facing Wikipedia, and identified as a possible cause of its demise in the future. If the practice of substantiating claims with accessible, reliable sources does not catch on in the Balkan Wikipedia community, there is little hope that the Wikipedia's of the Balkan region will ever be on par with its more developed, larger counterpart, and that the website will ever succeed in its mission of providing objective, neutral information about "anything and everything" to its readers.

On top of that, the "Yugoslav Wikipedia" appears to also be facing challenges in terms of organization of content. The key performance indicator called *Number of Categories* was included in the research in order to judge the quality of organization practices on the Wikipedia language versions that were studied. However, the data indicates that the English versions of the article are very often ahead of the local Wikipedia's when

it comes to the inclusion of Categories. This discrepancy is particularly evident in the case of English version biographies of globally famous persons, such as the Bosnians Ivo Andric and Emir Kusturica.

Finally, it is important to underline the comparatively small amount of traffic that is directed towards the articles that were analyzed, which points to the lack of participation in Wikipedia and the consequent inability to reach critical mass of users that would ensure the production of high-quality content and the upholding of the values of neutrality and objectivity that serve as pillars of the international Wikipedia community.

Since this portion of the analysis is concerned with examining the differences in quality of articles for other-country and same-country persons, Serbo-Croatian Wikipedia is not included on the list. The reason for this is because the Serbo-Croatian Wikipedia is, as its name suggests, in a language that spans over two separate countries. Hence, the researcher judges that including this language version in the analysis would produce inaccuracies in the results of the research.

5.7 Thematic analysis

The second stage of the research consists of the thematic analysis of place of birth and occupation of famous persons, data which will later be utilized in answering the secondary research question of this thesis.

5.7.1 Distribution of occupations – topical coverage

The distribution of occupations of the largest articles about famous persons from each of the countries will be categorized in order to allow for easier reference. The exact occupation will be scraped from the Wikidata repository for each of the persons, and then assigned (if necessary) to a broader category.

Name	Occupation	Category
Zivojin Misic	Field Marshal	Military
Stepa Stepanovic	Commander	Military
Jelena Jankovic	Tennis player	Sports

Table 5.26: Distribution of occupations for Serbian famous persons

Name	Occupation	Category
Aloysius Stepinac	Cardinal	Religion
Stjepan Radic	Politician	Politics
Franjo Tudjman	Politician	Politics

Table 5.27: Distribution of occupations for Croatian famous persons

Name	Occupation	Category
Ivo Andric	Writer	Arts
Emir Kusturica	Film director	Arts
Zdravko Colic	Pop singer	Arts

Table 5.28: Distribution of occupations for Bosnian famous persons

Name	Occupation	Category
Petar II Petrovic – Njegos	Prince-Bishop	Politics
Peko Dapcevic	General	Military
Slobodan Milosevic	Politician	Politics

Table 5.29: Distribution of occupations for Montenegrin famous persons

Name	Occupation	Category
Robert Kranjec	Ski jumper	Sports
Zoran Music	Painter	Arts
Anton Martin Slomsek	Bishop	Religion

Table 5.30: Distribution of occupations for Slovenian famous persons

The distribution of occupations of the famous persons from each of the countries complies with previous findings on the subject of topical coverage. Out of the 18 famous persons analyzed, all of them fall into one of the five categories of **Politics**, **Military**, **Religion**, **Sports**, and **Arts**. In the previous studies, the English Wikipedia was found to be strongest in these subjects. The greatest number of famous persons belong to either Politics (4) or Military (3), which are known as Wikipedia's strongest topics. This is followed by Arts (4), and Sports (2).

5.7.2 Birth-death places

When it comes to place of birth and death (if applicable), the data is concentrated around capital cities of the respective countries. It is important to note that for historical persons, the place of birth and death may not correspond to today's capital (e.g. in the case of Petar II Petrovic – Njegos, the place of death is Cetinje, which was at the time the capital city of Montenegro) but was nevertheless the capital city at the time.

When it comes to famous persons with international recognition, there are also some instances wherein they have died outside their native country, such as for instance Slobodan Milosevic, who died while on trial in The Hague.

For famous persons from Serbia, the most common birth (see Figure 5.1) and death (see Figure 5.2) place is the country's capital city, Belgrade (1135 and 321 people, respectively), followed by its second-largest city, Novi Sad. Other most common places of death and birth all belong to the Republic of Serbia, with the exception of Paris and Vienna, which represent places of death for 10 and 6 people, respectively (see Table: 5.31).

Rank	Birth place	Number	Death place	Number
1	Belgrade	1335	Belgrade	321
2	Novi Sad	234	Novi Sad	21
3	Kragujevac	136	Paris	10
4	Nis	134	Nis	7
5	Cacak	98	Vienna	6

Table 5.31: Most popular Birth-death places Serbian famous persons

Figure 5.1: Distribution of Birthplace for Serbian famous persons

Figure 5.2: Distribution of Death place for Serbian famous persons

When it comes to Croatia, the results follow a similar trend. The most common place of death (see Figure 5.4) and birth (see Figure 5.3) is Croatia's capital, Zagreb, with 706 births and 289 deaths. Zagreb is followed by Split, another large city in Croatia, with 387 births and 26 deaths. Other most common places of death and birth are smaller Croatian cities, with the exception of Belgrade, which is the place of death for 12 persons, and Rome, which is the place of death for 9 persons (see Table: 5.32).

Rank	Birth place	Number	Death place	Number
1	Zagreb	706	Zagreb	289
2	Split	387	Split	26
3	Rijeka	181	Dubrovnik	13
4	Osjek	140	Belgrade	12
5	Zadar	101	Rome	9

Table 5.32: Most popular Birth-death places Croatian famous persons

Figure 5.3: Distribution of Birthplace place for Croatian famous persons

Figure 5.4: Distribution of Death place for Croatian famous persons

Among Slovenian famous persons, the majority of births (see Figure 5.5) took place in Ljubljana (1193), Maribor (304), and Kranj (245). Interestingly, however, the most common places of death (see Figure 5.6) for famous persons from Slovenia are Ljubljana (351), Maribor (28) and Kralj (13), which are Slovenian cities (see Table: 5.33).

Rank	Birth place	Number	Death place	Number
1	Ljubljana	1193	Ljubljana	351
2	Maribor	304	Maribor	28
3	Kranj	245	Trieste	13
4	Celje	208	Kranj	11
5	Jesenica	113	Vienna	9

Table 5.33: Most popular Birth-death places Slovenian famous persons

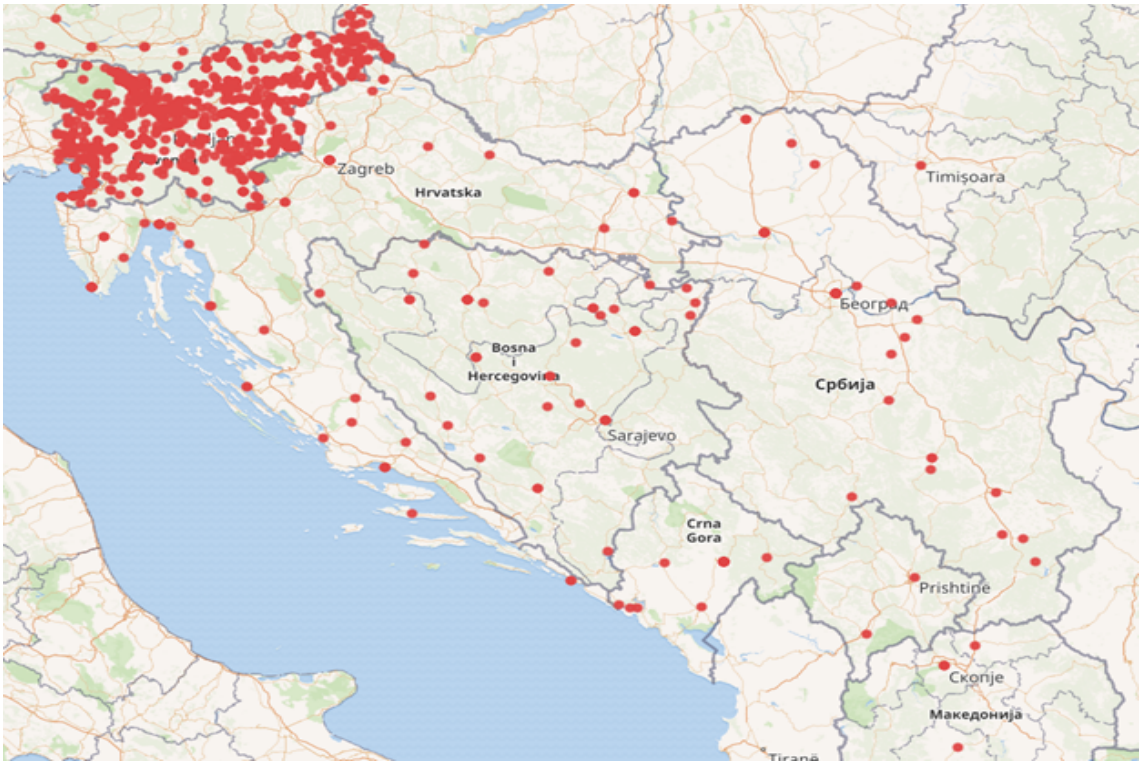

Figure 5.5: Distribution of Birthplace place for Slovenian famous persons

Figure 5.6: Distribution of Death place for Slovenian famous persons

For Bosnian famous persons, the most common birth place (see Figure 5.7) is Sarajevo, with 263 persons born there, followed by Banja Luka (76) and Mostar (73). In terms of places of death (see Figure 5.8), most famous persons from Bosnia are tied to Bosnian cities, but there is also a significant portion of famous persons whose death places are scattered throughout the former republics of Yugoslavia, including most commonly Serbia (Belgrade), Croatia, and Slovenia (see Table: 5.34).

Rank	Birth place	Number	Death place	Number
1	Sarajevo	263	Sarajevo	33
2	Banja Luka	76	Belgrade	13
3	Mostar	73	Zagreb	11
4	Tuzla	67	Banja Luka	3
5	Zenica	43	Mostar	3

Table 5.34: Most popular Birth-death places Bosnian famous persons

Figure 5.7: Distribution of Birthplace place for Bosnian famous persons

Figure 5.8: Distribution of Death place for Bosnian famous persons

The most common birth places for Montegrin famous persons are, once again, the country's capital city, Podgorica, with 168 births (see Figure 5.9), followed by Niksic with 73 births and Cetinje with 54. On the other hand, the most common death place (see Figure 5.10) of famous people from Montenegro is Belgrade (5), followed by Podgorica, Cetinje and Budva, with 3, 3, and 2 deaths respectively (see Table: 5.35).

Rank	Birth place	Number	Death place	Number
1	Podgorica	168	Belgrade	5
2	Niksic	73	Podgorica	3
3	Cetinje	54	Cetinje	3
4	Kotor	29	Budva	2
5	Bijelo Polje	26	Paris/Vienna	2

Table 5.35: Most popular Birth-death places Montenegrin famous persons

Figure 5.9: Distribution of Birthplace place for Montenegrin famous persons

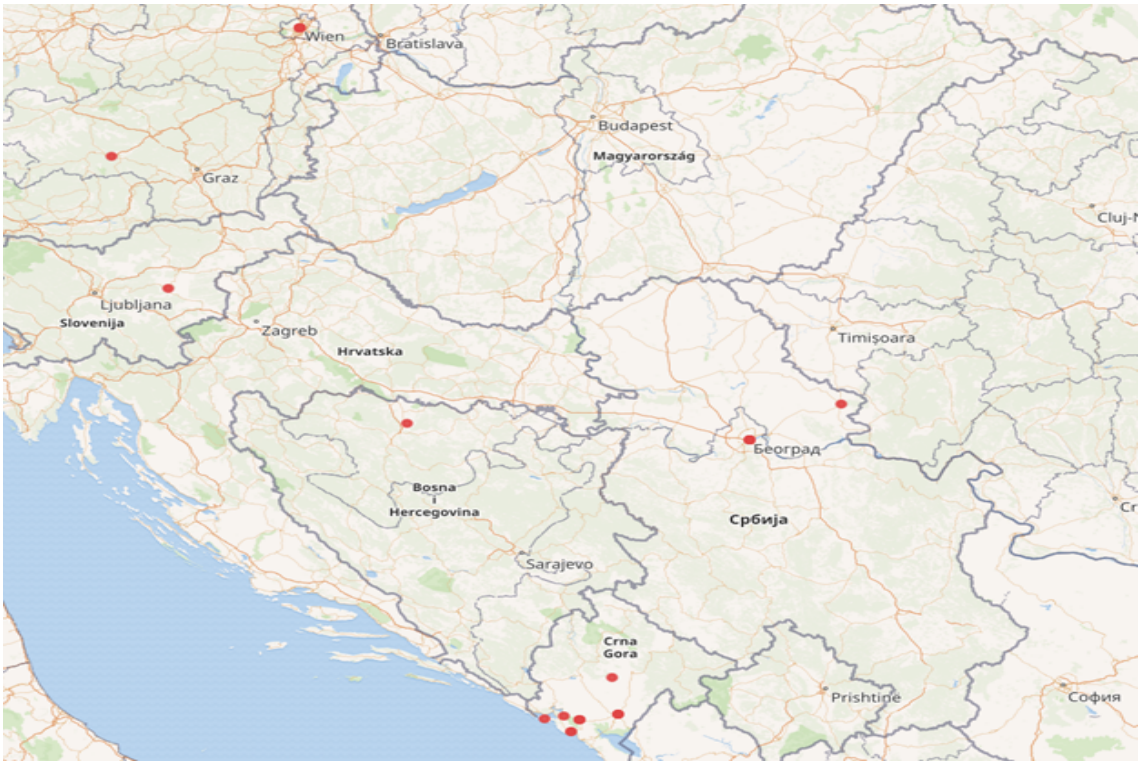

Figure 5.10: Distribution of Death place for Montenegrin famous persons

Overall, the results of this segment of data analysis demonstrate that both birth and death places of famous persons are most commonly associated with capital cities of their respective countries. When it comes to birth, the most common places are exclusively linked to capital cities; in terms of death places, there is some variation. Interestingly, although there are no famous persons born in other countries, Rome, Vienna, Paris, Prague and many other European cities emerge as death places of the analyzed famous persons. This occurrence is easily explained by the fact that a number of famous persons are of international prominence, which implies that at least some part of their life was spent in another country. In the case of smaller countries, such as Slovenia and Montenegro, the occurrence of death places being outside the country is more common, while for larger countries such as Croatia and Serbia, the most common places of death are also the capital cities of the respective countries.

Discussion

The findings of the thematic analysis of the content on famous persons confirmed that Wikipedia of the Balkans is close to the English Wikipedia in terms of topical coverage. Among the 18 famous persons, there were none that were related to social sciences, law, and medicine, which is also the case for the English Wikipedia. Similarly, majority of their occupations were related to Wikipedia's most popular topics: military, politics, arts,

and sports. In this case, the arts and sports famous persons can also be considered to be part of popular culture: in the case of the Bosnian Wikipedia, Emir Kusturica and Zdravko Colic, who are both still alive, are popular persons whose work is enjoyed by the living population of the country. Similarly, both of the sports-related famous persons, Jelena Jankovic and Robert Kranjec, are active in their respective sport, and globally acclaimed for their successes. In short, famous persons that belong to the Arts and Sports category are very much celebrities in today's society of their respective countries and further.

5.8 Qualitative analysis

In order to determine the articles to be chosen for qualitative analysis, a breakdown of the highest quality articles is required for each of the languages(see Table: 5.36). Given that the *local hero hypothesis* has been confirmed in most of the language versions, the qualitative analysis of content for the language versions whose famous persons are analyzed will be the same as their language for majority of the versions, except for the Bosnian. In other words, the highest quality Serbian article will be an article describing a Serbian famous person, the highest quality Croatian article will be an article describing a Croatian famous person, and so forth.

Language	**Best article**
Serbian	Zivojin Misic
Croatian	Franjo Tudjman
Bosnian	Stepa Stepanovic
Slovenian	Robert Kranjec
Serbo-Croatian	Zivojin Misic
English	Aloysius Stepinac

Table 5.36: Breakdown of highest quality articles for each of the languages according to KPI performance

5.8.1 Serbian article qualitative analysis

The qualitative analysis of the highest quality Serbian article, the biography of Field Marshal Zivojin Misic, scored 5/5 on the qualitative analysis questionnaire(see Table: 5.37). The researcher found that the overall tone and language of the article is neutral and objective, with appropriate language, free of emotion-rousing words. Furthermore, the article primarily provides a historical overview – without interpretation—and therefore adheres to the standards of encyclopedic content. The article is logically organized into chronological sections of Misic's life. With 139 references and 25 external links, the article is supported by evidence throughout, while also enabling to reader to navigate to other pages mentioned in the article. Finally, with a clear outline, a content-dense sidebar

Language	Best article
1. Is the intention of informing an audience clear from the text?	Y
2. Is the level of information appropriate for an encyclopedic article?	Y
3. Is the language of the article objective and free of emotion-rousing words?	Y
4. Is the article organized logically, easy to read, with main points clearly presented?	Y
5. Does the information appear to be well-researched and is it supported by evidence?	Y

Table 5.37: Serbian Language Article Evaluation Questionnaire

noting the most important information, and 25 categories, this article does well in terms of organization. In conclusion, this article is a high-quality article that scores well in terms of all common Wikipedia quality flaws and is comparable to high-quality articles of the English Wikipedia; according to the author's conclusions, it deserves the Featured Status it was awarded on the Serbian Wikipedia.

5.8.2 Croatian article qualitative analysis

Language	Best article
1. Is the intention of informing an audience clear from the text?	Y
2. Is the level of information appropriate for an encyclopedic article?	Y
3. Is the language of the article objective and free of emotion-rousing words?	N
4. Is the article organized logically, easy to read, with main points clearly presented?	Y
5. Does the information appear to be well-researched and is it supported by evidence?	N

Table 5.38: Croatian Language Article Evaluation Questionnaire

The Croatian article about its former president Franjo Tudjman does somewhat worse in terms of objectivity and neutrality of the article, as well as its verifiability through use of references (see Table: 5.38). Although structurally organized as a high-quality Wikipedia article, this biography lacks sufficient references for the various statements it is making, and also appears to often present opinions and present disputed facts as confirmed (e.g. it mentions the concept of "Great Serbian Aggression," which is not a historically appropriate term for referring to the Yugoslav Civil War). Furthermore, the article often employs glorifying terminology when discussing the role of former President in the country's

recent history, and overall seems to exhibit a nationalist leaning. Keeping in mind the accusations towards the Croatian Wikipedia of historical revisionism and nationalist outlook, mentioned in the Theoretical Background of this thesis, the author concludes that, although the article scores 3/5 in terms of the questionnaire, it nevertheless requires significant amount of work and editing before it fulfills the requirements for the Featured Article status which it was awarded on the Croatian Wikipedia. An important distinction to also keep in mind is that, unlike the article on the Serbian Wikipedia, whose subject is a historical person from the 20^{th} century, is that Franjo Tudjman is a recent political figure who continues to cause controversy and divide among the peoples of the former Yugoslav republics.

5.8.3 Bosnian article qualitative analysis

Language	**Best article**
1. Is the intention of informing an audience clear from the text?	Y
2. Is the level of information appropriate for an encyclopedic article?	Y
3. Is the language of the article objective and free of emotion-rousing words?	Y
4. Is the article organized logically, easy to read, with main points clearly presented?	Y
5. Does the information appear to be well-researched and is it supported by evidence?	Y

Table 5.39: Bosnian Language Article Evaluation Questionnaire

The Bosnian article, describing the Serbian military General Stepa Stepanovic also scored well in all categories of the questionnaire, receiving a score of 5/5 (see Table: 5.39). However, it is important to note that on the Wikipedia page of the article, there is a notice stating that the article was originally written for the Serbian Wikipedia version, and transcribed only in terms of dialect to be fit for Bosnian language. Nevertheless, the article demonstrates encyclopedic quality in terms of all criteria, including verifiability. With 66 references, it does comparatively well. However, in terms of organization, it lacks enough categories, which can be attributed to the lack of Bosnian Wikipedia articles for the concepts introduced or mentioned in the article. Moreover, it is important to keep in mind that the Bosnian Wikipedia is amongst the smallest and most recent in size of all the Balkan Wikipedia's, which makes the high quality of this article laudable. Finally, the article steers clear of presenting any opinion, and it structured well, divided into chronological sections of Stepanovic's life.

Language	Best article
1. Is the intention of informing an audience clear from the text?	Y
2. Is the level of information appropriate for an encyclopedic article?	Y
3. Is the language of the article objective and free of emotion-rousing words?	Y
4. Is the article organized logically, easy to read, with main points clearly presented?	Y
5. Does the information appear to be well-researched and is it supported by evidence?	Y

Table 5.40: Slovenian Language Article Evaluation Questionnaire

5.8.4 Slovenian article qualitative analysis

The Slovenian article that describes Robert Kranjec, a Slovenian ski jumper, scores very well in terms of all the criteria (see Table: 5.40). With a clear, objective tone throughout the article, the biography covers the jumper's career in a chronological fashion, from the beginnings to his current standings, including the medals won and a list of competitions that the ski jumper participated in. The sidebar content, presence of photos and a clear, logical outline also determine a high score in terms of organization, with more than enough references and external links to provide a substantiated, balanced view of Kranjec's career. The article does slightly worse in terms of organization of paragraphs (a criterion that is not scored in this questionnaire), but without hindering readability and understanding of the article.

5.8.5 Serbo-Croatian article qualitative analysis

Language	Best article
1. Is the intention of informing an audience clear from the text?	Y
2. Is the level of information appropriate for an encyclopedic article?	Y
3. Is the language of the article objective and free of emotion-rousing words?	Y
4. Is the article organized logically, easy to read, with main points clearly presented?	Y
5. Does the information appear to be well-researched and is it supported by evidence?	Y

Table 5.41: Serbo-Croatian Language Article Evaluation Questionnaire

The highest quality article on the Serbo-Croatian Wikipedia is that of Zivojin Misic, the

Serbian Field Marshal (see Table: 5.41). Before an evaluation of its score, it is important to note that the author concludes that this article is an adaptation of the previously analyzed Serbian article about Zivojin Misic, which is indicated by the same number of references and external links, and the same organization of the article. Nevertheless, this article receives a 5/5 score on the evaluation questionnaire, as it appears to embody all of the qualities expected from an encyclopedic article, including objectivity, verifiability and breadth of information. The article somewhat lacks in the number of categories, which can be attributed to the underdevelopment of the Serbo-Croatian Wikipedia as a whole, and not necessarily to the article as such. The article is also further improved with an average-quality sidebar showcasing basic information about Misic's life, and a logical, easy to read organization of the different sections within the article.

5.8.6 English article qualitative analysis

Language	**Best article**
1. Is the intention of informing an audience clear from the text?	Y
2. Is the level of information appropriate for an encyclopedic article?	Y
3. Is the language of the article objective and free of emotion-rousing words?	Y
4. Is the article organized logically, easy to read, with main points clearly presented?	Y
5. Does the information appear to be well-researched and is it supported by evidence?	Y

Table 5.42: English Language Article Evaluation Questionnaire

The highest-performing English article is the one covering the biography of Aloysius Stepinac, a Croatian Cardinal who has been Blessed by Pope John Paul II towards the end of the 1990s (see Table: 5.42). Upholding the standards of the best articles in the English Wikipedia, the article on His Eminence Blessed Dr. Aloysius Stepinac scored 5/5 on the quality evaluation questionnaire. The article provides an encyclopedic, objective overview of Stepinac's life, refraining from expression of opinion or presenting controversial and disputed views, therefore adhering to Wikipedia's neutrality standards. Furthermore, the article is well-organized, with a content-dense sidebar that provides all the vital information about the life of Stepinac, as well as a clear outline that separates the article content into thematic wholes, corresponding to a chronological order. With 53 references and 80 external links, the article is strong in terms of verifiability and provides an excellent starting point for further research. Overall, the qualitative analysis of the article indicates its quality in all respects.

Discussion

The results of the qualitative analysis indicate that, at least in terms of the largest and best-performing articles, the standards of the Balkan Wikipedia versions do not differ greatly from those of the English and other large Wikipedia versions. Overall, the articles that were analyzed performed well in all respects, with some of them particularly strong in terms of verifiability—the Serbian and Serbo-Croatian article on Field Marshal Zivojin Misic featured 139 references and 25 external links, which is stronger than many English articles.

Furthermore, neutrality appeared to be compromised in only one of the six articles, that of Franjo Tudjman, the former president of Croatia. As it has been noted in the analysis, this can, on the one hand, be partially attributed to the previously identified nationalist and revisionist tendencies of the Croatian Wikipedia community; on the other hand, it can be attributed to the fact that Franjo Tudjman, unlike any of the other persons featured in the articles, is a controversial historical person, with disputes and divisions still rampaging both the common and academic community of the former Yugoslav republics. On top of that, unlike the historical persons analyzed in articles from other versions, Tudjman belongs to a recent history, which does not allow for the benefit of hindsight to set in and the academic community to arrive to a consensus about his role and importance in Croatian history. Nevertheless, the author concludes that a possible source of worry for the overall quality of the Croatian Wikipedia is the fact that, regardless of the fairly obvious breaches of the objectivity and neutrality principles of Wikipedia, the Croatian article about Franjo Tudjman was awarded a Featured Status on the Croatian Wikipedia, thus causing the researcher to question the overall quality standards of the Croatian Wikipedia.

The qualitative analysis of the Bosnian language article also underlines another issue: the issue of underdevelopment of this Wikipedia, which the author believes should be addressed in the future. The article regarding Stepa Stepanovic, a Serbian military General, was in fact adapted from the Serbian Wikipedia, which points towards the lack of activity and participation in the Bosnian Wikipedia, whose KPI scores for persons from their own country were consistently low throughout this thesis. These findings, however, are in alignment with the usage statistics of the Wikipedia, as well as the size and length of existence of this Wikipedia version. Reasons for lack of usage fall outside the scope of this research, but the author believes that looking into this phenomenon could shed light and reveal interesting insights about reasons for lack of participation in Wikipedia, not only in Bosnia and Herzegovina, but throughout the world.

When it comes to the English Wikipedia, whose highest-performing article was that of Aloysius Stepinac, the author attributes this to the supra-national role of Stepinac. Although a Croat by ethnic belonging, Stepinac is primarily known for his role in the Catholic Church, particularly following his Sainthood that was established in 1998. It is important to underline that the quality of this article, which can be used as a benchmark due to the stronger enforcement of Wikipedia policies in the English Wikipedia than in its Balkan counterparts, is very similar to that of other languages, thus once again pointing to the strong upholding on Wikipedia values in the Wikipedia versions from

former Yugoslavia.

Ultimately, the author believes that this qualitative analysis, although definitely not applicable to the entirety of the Balkan Wikipedia versions, indicates that among the respective Wikipedia communities there is genuine understanding of the principles of Wikipedia, which is an important factor for the future development of these Wikipedia versions. Although articles of much lower quality undoubtedly exist in all of the Wikipedia versions, the demonstrated understanding of Wikipedia principles is reassuring; it signifies that the quality can and possibly will be achieved in the future, once the critical mass of users is reached. In the case of the Croatian Wikipedia, the author of this thesis concludes that the Tudjman article (and its Featured Status) demonstrate a lack of this kind of understanding in the Croatian Wikipedia. The reasons for the lack of understanding, and their socio-cultural roots fall outside the scope of this thesis, but are certainly a factor to be considered in future research.

CHAPTER 6

Results

The primary research question of this thesis was as follows:

"What is the difference in availability and quality of data about historically important persons across different languages of Wikipedia, as well as in English?"

The results of the research have confirmed the premise and hypothesis of this research that, across the different languages of the former Yugoslav republics, there are significant discrepancies in the amount and quality of data available. To begin with, one of the most insightful findings was that for a number of the chosen persons, articles were not available in all of the languages, particularly in the case of coverage of Slovenian famous persons. In fact, out of 18 articles in total, there were 6 articles that did not exist, while others were drastically smaller than those in the Slovenian version of Wikipedia.

On top of that, the results of this thesis support the *local hero hypothesis* almost exclusively. The availability of information about same-country persons seems to be much higher in the native language of the person being described than in other languages, thus pointing towards necessary amendments to all Wikipedia versions to provide a more inclusive picture. Similarly, articles about same-country persons perform better in all other KPI measurements, thus underlining the truthfulness of the *local hero hypothesis* and confirming the premise of the research that there is much more available information about local persons in their native language than in any other languages.

When it comes to the coverage of historically important persons from the Balkans in the English language, there is a varied amount of coverage. Primarily, it is important to note that all of the chosen famous persons have an article on the English Wikipedia, even those whose historical importance is linked exclusively to the local community, with little international recognition. However, in the case of globally known famous persons—such as Jelena Jankovic (Serbian tennis player), Slobodan Milosevic (former President of Serbia trialed in The Hague for war crimes), Ivo Andric (Nobel-winning Bosnian novelist), Emir Kusturica (award-winning Bosnian film director), Aloysius Stepinac (prominent

Catholic Cardinal blessed by Pope John II in 1998), Robert Kranjec (Slovenian ski jumper with international success)—the coverage in English articles is much greater than for those that are important only in terms of local history. Hence, it can be concluded that the availability of information in English is sufficient for a basic understanding and information about all of the famous persons, but increases greatly when the person is also of interest to the international community.

In terms of quality of data, the results indicate several important flaws that are encountered by the Balkan region Wikipedia's. One of the most prominent examples of quality flaws is the severe lack of editing across all languages, regardless of the importance of the person described in the article. Among the Wikipedia versions from the Balkans, there is a significant number of articles with zero edits and unique editors, with only a couple that have experienced more than 10 edits. Hence, the collaborative principle, which is one of the five pillars of Wikipedia, is rarely enforced and weakly utilized in the context of Wikipedia's from the Balkan region. Interesting findings regarding the online behavior with respect to Hofstede's theory of cultural dimensions (see: chapter 3) may explain this phenomenon, but this would require further research with a different outlook.

In terms of verifiability, which is another important indicator of information quality, the results are varied. The analysis of the data indicates that, for articles of great importance, referencing and external linking are very prominent, while this is not the case for articles regarding other-country persons. This may be due to lack of reliable sources in the language of the article regarding other-country persons, which may point toward a gap in knowledge in a wider context, or the unavailability of those sources in an online/electronic format. However, to explore the causes of this phenomenon, further research is also required.

Moreover, in terms of reaching critical mass, the results indicate that there is relatively little activity on the Balkan region Wikipedia's, with the most popular articles reaching less than 15.000 views in the two-month period (November-December 2016) that was analyzed. With some articles receiving as little as 35 views in 60 days (Slovenian article on Stepa Stepanovic), and many of them falling short of a thousand views in two months, there is an evident lack of audience for these Wikipedia versions. Hence, in order to reach the critical mass that is one of the major reasons for Wikipedia success, Wikipedia traffic in all of the Wikipedia countries needs to be increased.

When it comes to organization, there seems to be a good understanding of the importance of categories included in the article, and the importance of article organization as a whole. Cross-linking indicates a certain level of dedication of the Wikipedia community from the respective countries to ensure quality practices. Nevertheless, there is once again a significant gap between the number of categories included in articles for same-country persons, as opposed to number of categories included in articles for other-country persons. In conclusion, the structural principles of all the Wikipedia's are performing relatively well; naturally, there is space for improvement.

The quantitative results presented above were fully supported by the qualitative analysis

performed on a sub-set of six articles from each of the language versions of Wikipedia. Five out of six articles demonstrated a thorough understanding of all the Wikipedia principles, thus testifying to the upholding of Wikipedia values in the highest-quality article category. Although it cannot be accurately extrapolated that this is, therefore, the case for all articles in the respective Wikipedia's (which is also a question for further research), the results of the qualitative analysis of the highest-performing article underline that there is great potential in these Wikipedias—the author concludes that with reaching critical mass of users that are necessary to create such a vast body of knowledge, the quality of the entirety of Wikipedias has strong chances of being high, since the core community seems to have a deep grasp of what Wikipedia stands for and how knowledge should be presented and disseminated on the website.

The secondary research question of this thesis is concerned with the relationship between global importance and neutrality and quality of the article:

"To what extent does neutrality and quality increase with more contributions and when the subject of the article is a person of global importance?"

The results of this thesis seem to confirm the hypothesis that neutrality and quality of article increases with greater importance of the famous person analyzed, and with greater number of contributors. This finding is primarily demonstrated in the case of English Wikipedia. As it has been mentioned above, the articles from the English Wikipedia, whose subject is a person from one of the analyzed countries with global recognition and influence, uphold a very strong standard of quality, as well as neutrality and objectivity.

To begin with, articles that concern historically very important persons are strongest in terms of verifiability and referencing. With some of those articles featuring more than a hundred references, the qualitative analysis revealed that the majority of claims presented in the articles is substantiated by a reliable reference. Furthermore, articles whose subjects are persons of great importance tend to be longest, thus conforming to the hypothesis that with greater word count quality improves. For instance, the English article about Slobodan Milosevic, whose war crimes and nationalist politics have been receiving international attention since the 1990s, has a length of approximately 102.000 bytes. Similarly, the Serbian article about Jelena Jankovic, a world-famous tennis player from Serbia, is 283.197 bytes long.

The qualitative analysis of the highest-performing articles also confirmed the use of neutral, objective language that is free of emotion-rousing words, glorification or opinion-presenting. The qualitative analysis revealed that the majority of the highest-performing articles is concerned only with presenting substantiated facts, rather than opinionated or disputed claims. The intention to inform an audience, as opposed to an intention to convince or influence, has also been clearly identified in all of the highest-performing articles concerning persons of global importance, across the various language versions.

However, this finding was not supported across all language versions. Further confirming the *local hero hypothesis*, the increase in quality with greater importance and global relevance is only evident in the local Wikipedia version (that of the country from which

the famous person is coming) and the English Wikipedia version. When it comes to other languages, the quality and neutrality of content remains limited by the overall weak availability of content on those famous persons. Hence, the article on Robert Kranjec, a famous Slovenian ski jumper whose Slovenian and English language articles perform very well in terms of objectivity and neutrality, the Serbian version of the article has zero references, while the Serbo-Croatian Wikipedia does not even have its own version of the article.

CHAPTER 7

Limitations

7.1 Data limitations

The primary limitation of this thesis is the relatively small data sample. As an analysis of 18 Wikipedia articles, 6 from each of the language versions, this study is only a first step towards much more detailed research in the subject of the cross-lingual character of Wikipedia. Due to the vast amount of data available on the subject, and the contrastingly specific research goals of this thesis, along with technical constraints of acquiring and processing the data, this research is limited by the scope of its data.

The specific research goal of this thesis was to examine the dataset for quality from a variety of angles, therefore combining the cumulative knowledge that has been identified from the large body of literature tackling the subject of the English Wikipedia.

Because there is no previous research on the specific language versions that were examined in this thesis, and therefore no secondary sources to use as a replacement, the researcher decided to perform the variety of analysis that would ensure *depth* of data, as opposed to analyzing a larger data set, which would ensure a broader analysis but less depth.

Ultimately, this limitation is most important in terms of drawing *general* conclusions about the nature of the Wikipedia versions of the former Yugoslav republics. The researcher warns against the hazards of using this thesis to draw general conclusions, and advises that broader research is performed for these purposes. However, this thesis can be used a starting point and estimation of the state of the art in the ex-Yugoslav Wikipedia's.

7.2 Methodology limitations

When it comes to methodology, the primary methodological limitation of this study lies within the third, qualitative stage of the research. The researcher acknowledges

that there are much more comprehensive and detailed methods of information quality evaluation. However, because these kinds of analyses fall outside the expertise of the researcher, and because the general conclusions that can be drawn from the adapted Cornell University questionnaire satisfy the research objectives of this thesis, the researcher concurs that, albeit limited, the qualitative analysis performed in this thesis provides sufficient information for the conclusions drawn.

Furthermore, the questionnaire does not employ a ranking system, but rather a 1-0 system, which does not leave enough space for the intricacies of the examination of quality information, and can be improved through a better design of the questionnaire.

7.3 Causality limitations

Examining the causes of the various phenomena that are only identified in this thesis also falls outside its scope. Although the researcher identifies the possible and/or probable causes of certain occurrences, these hypotheses are by no means supported by empirical evidence.

In order to examine the causality and various factors that influence the occurrence of the identified phenomena, further research and a completely different research design is required.

7.4 Technical and time constraints

The technical, i.e. word count and Master's thesis requirements, and time, i.e. deadline, constraints of this research dictated the choice of the subject and goals of the research, which needed to be specific enough to accommodate for these conditions. Hence, there is a wealth of insight derived from the analyzed data which was not further explored in this research, but can nevertheless be explored in the future work of this researcher, or others.

7.5 Technical and time constraints

Because there has been no previous research done on the subject, the findings and results of this research can only be substantiated with theories, hypotheses and findings that have been arrived at through research of the English Wikipedia, along with a minority that tackled other-language versions.

The evidence base for these findings, therefore, is unique. For the purposes of confirming the results of the research, additional examination of the Balkan Wikipedia's and their quality would need to be performed to evaluate, confirm and/or dispute the findings presented in this study.

CHAPTER 8

Areas for future research

The possible areas for expansion of this research are two-fold. On the one hand, there is relatively little research done (and only recently) on the subject of multi- and cross-lingual Wikipedia analysis. This thesis, which looks at availability, quality and neutrality of Wikipedia content across six Wikipedia language versions and five languages (Serbo-Croatian not being an official language), may be further expanded to perform a cross-lingual analysis of content quality on a much larger scale. With practices of information quality evaluation improving greatly through automation, this field of research is expected to become much easier to facilitate and much more viable in the future.

On the other hand, this research can be expanded to include a larger sample of content from the Wikipedia's of the Balkan region, and therefore examine in much greater detail and with greater precision the relationships between these Wikipedia versions, analyzing the causal relationship between political, socio-cultural and identity relations that take place offline and the impact these factors have on Wikipedia content availability and quality. Insights such as the revisionist and nationalist tendency of the Croatian article regarding Franjo Tudjman indicate that there is plenty of space for research in this field. Furthermore, from a socio-historical perspective, the content of Wikipedia and the interplay of the various Wikipedia's can be studied in order to examine in greater detail the social dynamics and consensus building in war-torn regions with a mutual history.

Overall, the unique position Wikipedia as a concept holds in the world today, and its greater role in the Web 2.0 movement provides a multitude of research opportunities that emerged often throughout the performing of this research.

CHAPTER 9

Conclusion

In conclusion, albeit limited, this thesis holds great importance as a pioneering comparative study of cross-lingual availability of content in the Balkan region, whose principles are not limited by the region and can be applied regardless of the language of the Wikipedia.

The primary conclusion of this thesis is linked to the *critical mass* hypothesis, which stipulates that a greater number of users and visitors of Wikipedia results in better, higher-quality articles. The truthfulness of this hypothesis has been most prominently demonstrated in the results of the analysis of English articles. Seeing as the English Wikipedia is the most popular one, and by extension the one most likely to have reached critical mass, the vast differences in word count, number of references, and number of edits between English versions and Balkan-language versions of articles for Balkan persons of international prominence, this thesis proves that the quality of articles undoubtedly improves with greater number of contributors and visitors.

Furthermore, an important conclusion that can be drawn from this research is that the *local hero hypothesis*, which stipulates that persons from the same country ("locals") will receive greater attention and therefore better articles is also true. Throughout the analysis of articles in one language about other-country persons, the results have clearly demonstrated this type of favoritism described above. In turn, this raises important questions about Wikipedia's role in knowledge building and dissemination across cultures.

Finally, the results of this thesis point towards an important trend. By looking at the research aimed at the Wikipedia phenomenon chronologically and historically, and comparing the results of the abovementioned researches with the results of this thesis, it becomes evident that the trajectory of any language version of Wikipedia follows a certain pattern. Although it falls outside the scope of this research to analyze and examine the details of this pattern, it is important to note that these findings then underline the relevance of all the global researches performed on the English Wikipedia for other-language Wikipedia versions. Overall, the various findings of authors that studied the

English Wikipedia in terms of topical coverage, quality of information, availability of information, and neutrality and objectivity of Wikipedia have been confirmed in this study, thus highlighting the global relevance of the studies performed on the English Wikipedia.

List of Figures

List of Tables

Glossary

AC Average number of categories. 44, 46, 48, 50, 52–55

ADA Average daily average number of visitors. 44, 46, 48, 50, 52–55

AE Average number of edits. 44, 46, 48, 50, 52–55

AEL Average number of external links. 44, 46, 48, 50, 52–55

ALB Average length in bytes. 44, 46, 48, 50, 52–55

AR Average number of references. 44, 46, 48, 50, 52–55

AUE Average number of unique editors. 44, 46, 48, 50, 52–55

AV Average number of views (01.11-31.12.2016). 44, 46, 48, 50, 52–55

C Categories. 43–52

DA Daily average. 43–52

E Edits. 43–52

EL External links. 43–52

LA Languages available. 43–52

LB Length in bytes. 43–52

R References. 43–52

UE Unique editors. 43–52

V Views (01.11-31.12). 43–52

Bibliography

[Ama] Amazon. Alexa. http://www.alexa.com/siteinfo/wikipedia.org. Visited: January 15, 2017.

[AS12] Maik Anderka and Benno Stein. A breakdown of quality flaws in wikipedia. In *Proceedings of the 2Nd Joint WICOW/AIRWeb Workshop on Web Quality*, pages 11–18. ACM, 2012.

[ASW09] Denise Anthony, Sean W. Smith, and Timothy Williamson. The case of the online encyclopedia wikipedia. *Reputation and Reliability in Collective Goods*, 21(3), 2009.

[atW] Encyclopedias around the World. Encyclopedia). http://www.edinformatics.com/inventions_inventors/encyclopedia.htm. Visited: February 14, 2017.

[BB14] Pasko Bilic and Luka Bulian. Lost in translation: Contexts, computing, disputing on wikipedia. *iConference 2014 Proceedings*, Jan 2014.

[Blu08] Joshua E. Blumenstock. Size matters: Word count as a measure of quality on wikipedia. In *Proceedings of the 17th International Conference on World Wide Web*, pages 1095–1096. ACM, 2008.

[CC79] Robert Chambers and Robert Carruthers. *Chambers's Cyclopaedia of English literature; a history, critical and biographical, of British and American authors, with specimens of their writings, Tomes 3 and 4*. American Book exchange, 1879.

[CH11] Ewa S. Callahan and Susan C. Herring. Cultural bias in wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology*, 62(10):1899–1915, Jul 2011.

[DHPW05] Peter Denning, Jim Horning, David Parnas, and Lauren Weinstein. Wikipedia risks. *Communications of the ACM*, 48(12):152, Dec 2005.

[Eij10] Dr. Henk Simon Eijkman. Academics and wikipedia: Reframing web 2.0 as a disruptor of traditional academic power-knowledge arrangements. *Campus-Wide Information Systems*, 27(3), 2010.

[Gal07] Francis Galton. Vox populi. *Nature*, 75(1949):450–451, Jul 1907.

[Gil05] Jim Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, 2005.

[HG10] Brent Hecht and Darren Gergle. The tower of babel meets web 2.0: User-generated content and its applications in a multilingual context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 291–300. ACM, 2010.

[HL08] Alexander Halavais and Derek Lackaff. An analysis of topical coverage of wikipedia. *Journal of Computer-Mediated Communication*, 13(2):429—-440, 2008.

[Jem14a] Dariusz Jemielniak. *Common Knowledge?: An Ethnography of Wikipedia.* Stanford University Press, 2014.

[Jem14b] Dariusz Jemielniak. *Common Knowledge?: An Ethnography of Wikipedia*, chapter Introduction, page 18. Stanford University Press, 2014.

[Jem14c] Dariusz Jemielniak. *Common Knowledge?: An Ethnography of Wikipedia*, pages 83–84. Stanford University Press, 2014.

[KK08] Aniket Kittur and Robert E. Kraut. Harnessing the wisdom of crowds in wikipedia: Quality through coordination. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, pages 37–46. ACM, 2008.

[KM06] Josef Kolbitsch and Hermann Maurer. The transformation of the web: How emerging communities shape the information we consume. *Journal of Universal Computer Science*, 12(2):187–213, Feb 2006.

[Kuz06] Stacey Kuznetsov. Motivations of contributors to wikipedia. *SIGCAS Comput. Soc.*, 36(2), Jun 2006.

[KW08] Ayelt Komus and Franziska Wauch. *Wikimanagement: Was Unternehmen von Social Software und Web 2.0 lernen können.* Oldenbourg Wissenschaftsverlag, 2008.

[Lei14a] Thomas M. Leitch. *Wikipedia U: knowledge, authority, and liberal education in the digital age.* Johns Hopkins University Press, 2014.

[Lei14b] Thomas M. Leitch. *Wikipedia U: knowledge, authority, and liberal education in the digital age*, chapter Paradoxes of Verifiability, page 40. Johns Hopkins University Press, 2014.

[Lih10] Andrew Lih. *The Wikipedia revolution: how a bunch of nobodies created the world's greatest encyclopedia.* Aurum, 2010.

[Lin92] David C. Lindberg. *The beginnings of Western science: the European scientific tradition in philosophical, religious, and institutional context.* Univ. of Chicago Press, 1992.

[Liv] Livius. Pliny the elder, natural history. http://www.livius.org/articles/person/pliny-the-elder/pliny-the-elder-natural-history/. Visited: February 12, 2017.

[LST10] Wei-Ying Lim, Hyo-Jeong So, and Seng-Chee Tan. elearning 2.0 and new literacies: Are social practices lagging behind? *Interactive Learning Environments*, pages 203–218, 2010.

[Ma06] Cathy Ma. The social, cultural and economic implications of the wikipedia. *Intercultural Communication Studies*, XV(2):195–203, 2006.

[McF] Paul McFedries. It's a wiki wiki world. http://spectrum.ieee.org/computing/software/its-a-wiki-wiki-world. Visited: December 22, 2016.

[Med] MediaWiki. Api. https://www.mediawiki.org/wiki/API:Main_page. Visited: February 10, 2017.

[NC] Jeremy Norman and Co. John harris issues the first english encyclopedia arranged in alphabetical order (1704 – 1710). http://www.historyofinformation.com/expanded.php?id=3374. Visited: February 12, 2017.

[ND] New Mexico State University Nikos Drakos. Ontologies (and encyclopedic knowledge). http://www.cs.nmsu.edu/~tomohara/comps_review/node7.html. Visited: February 18, 2017.

[Oko09] Chitu Okoli. A brief review of studies of wikipedia in peer-reviewed journals. *Third International Conference on Digital Society*, page 1, 2009.

[OPSL15] J. Evans Ochola, Dorothy M Persson, Lisa A Schumacher, and Mitchell D Lingo. Wikipedia: the difference between information acquisition and learning knowledge. *First Monday*, 20(12), 2015.

[O'S09] Dan O'Sullivan. *Wikipedia: a new community of practice?*, page 79. Ashgate, 2009.

[Par07] David Parmenter. *Key Performance Indicators: Developing, Implementing, and Using Winning Kpis.* John Wiley & Sons, Inc., New York, NY, USA, 2007.

[Par08] David Parry. Wikipedia and the new curriculum: digital literacy is knowing how we store what we know. *Science Progress*, 2008.

[Pet09] Robert S. Petersen. Cult of the amateur: How today's internet is killing our culture, by andrew keen. *Design and Culture*, 1(3):385–387, 2009.

[Poe06] Marshall Poe. A closer look at the neutral point of view (npov). *Atlantic Media Company*, 2006.

[PZA06] Ulrike Pfeil, Panayiotis Zaphiris, and Chee Siang Ang. Cultural differences in collaborative authoring of wikipedia. *Journal of Computer-Mediated Communication*, 12(1):88–113, 2006.

[Ras08] Morten Rask. The reach and richness of wikipedia: Is wikinomics only for rich countries? *First Monday*, 13(6), May 2008.

[RL12a] Joseph M. Reagle and Lawrence Lessig. *Good Faith Collaboration The Culture of Wikipedia*. The MIT Press, 2012.

[RL12b] Joseph M. Reagle and Lawrence Lessig. *Good Faith Collaboration The Culture of Wikipedia*, page 4. The MIT Press, 2012.

[Ros06] R. Rosenzweig. Can history be open source? wikipedia and the future of the past. *Journal of American History*, 93(1):117—-146, Jan 2006.

[SAFJ09] Besiki Stvilia, Abdullah Al-Faraj, and Yong Yi Jeong. Issues of cross-contextual information quality evaluation-the case of arabic, english, and korean wikipedias. *Library & Information Science Research*, 31(4):232–239, 2009.

[San] Larry Sanger. Review of keen's cult of the amateur. `https://web.archive.org/web/20070825130320/http://blog.citizendium.org/2007/07/17/review-of-keens-cult-of-the-amateur-2/`. Visited: January 05, 2016.

[Shi08] Clay Shirky. *Here comes everybody: the power of organizing without organizations*. Penguin Press, 2008.

[SSL12] Garrett A. Sullivan, Alan Stewart, and Rebecca Lemon. *The encyclopedia of English renaissance literature. A-F*. Wiley-Blackwell, 2012.

[STGS05] Besiki Stvilia, Michael B Twidale, L Gasser, and Linda C Smith. Information quality discussions in wikipedia. In *Knowledge Management: Nurturing Culture, Innovation, and Technology - Proceedings of the 2005 International Conference on Knowledge Management*, pages 101–113. Graduate School of Library and Information science, University of Illinois at Urbana-Champaign, 2005.

[VWKvH07] Fernanda B. Viegas, Martin Wattenberg, Jesse Kriss, and Frank van Ham. Talk before you type: Coordination in wikipedia. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, pages 78–. IEEE Computer Society, 2007.

[Wei] Matt Weinberger. Microsoft had a secret, genius reason for making an encyclopedia in the nineties. `http://www.businessinsider.com/history-of-microsoft-encarta-2015-11`. Visited: December 15, 2016.

[WH07] Dennis M. Wilkinson and Bernardo A. Huberman. Cooperation and quality in wikipedia. In *Proceedings of the 2007 International Symposium on Wikis*, pages 157–164. ACM, 2007.

[Wika] Wikidata. Sparql query service. `https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/A_gentle_introduction_to_the_Wikidata_Query_Service`. Visited: February 11, 2017.

[Wikb] Wikimedia. List of wikipedias. `https://meta.wikimedia.org/wiki/List_of_Wikipedias`. Visited: December 15, 2016.

[Wikc] Wikimedia. Serbian wikipedia. `https://en.wikipedia.org/wiki/Serbian_Wikipedia`. Visited: December 17, 2016.

[Wikd] Wikimedia. Wikipedia: Five pillarsa. `https://en.wikipedia.org/wiki/Wikipedia:Five_pillars`. Visited: February 18, 2017.

[Wike] Wikipedia. About wikipedia. `https://en.wikipedia.org/wiki/Wikipedia`. Visited: October 19, 2016.

[Wikf] Wikipedia. English wikipedia. `https://en.wikipedia.org/wiki/English_Wikipedia`. Visited: February 14, 2017.

[Wikg] Wikipedia. History of encyclopedias. `https://en.wikipedia.org/wiki/History_of_encyclopedias`. Visited: February 15, 2017.

[Wikh] Wikipedia. History of wikipedia. `https://en.wikipedia.org/wiki/History_of_Wikipedia`. Visited: February 15, 2017.

[Wiki] Wikipedia. Page information for abortion. `https://en.wikipedia.org/w/index.php?title=Abortion&action=info`. Visited: December 20, 2016.

[Wikj] Wikipedia. Page information for "george w. bush". `https://en.wikipedia.org/w/index.php?title=George_W._Bush&action=info`. Visited: December 05, 2016.

[Wikk] Wikipedia. Wikidata. https://www.wikidata.org/wiki/Wikidata:Introduction. Visited: October 19, 2016.

[Wikl] Wikipedia. Wikidata spaql. https://en.wikipedia.org/wiki/File:Wikidata's_SPARQL_introduction_presentation.pdf. Visited: February 10, 2017.

[Wikm] Wikipedia. Wikipedia: Neutral point of view. https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_viewau. Visited: January 15, 2017.

[WL15] Adam M. Wilson and Gene E. Likens. Content volatility of scientific topics in wikipedia: A cautionary tale. *Plos One*, 10(8), 2015.

[ZCR15] Yiwei Zhou, Alexandra I. Cristea, and Zachary L. Roberts. Is wikipedia really neutral? a sentiment perspective study of war-related wikipedia articles since 1945. In *PACLIC*. ACL, 2015.