CULTURAL NEIGHBOURHOODS, OR APPROACHES TO QUANTIFYING CULTURAL CONTEXTUALISATION IN MULTILINGUAL KNOWLEDGE REPOSITORY WIKIPEDIA

by

Anna Samoilenko

Approved Dissertation thesis for the partial fulfillment of the requirements for a
Doctor of Natural Sciences (Dr. rer. nat.)
Fachbereich 4: Informatik
Universität Koblenz-landau

Chair of PhD Board: Prof. Dr. Ralf Lämmel
Chair of PhD Commission: Prof. Dr. Stefan Müller
Examiner and Supervisor: Prof. Dr. Steffen Staab
Further Examiners: Prof. Dr. Brent Hecht, Jun.-Prof. Dr. Tobias Krämer

Date of the doctoral viva: 16 June 2021

Cultural Neighbourhoods, or approaches to quantifying cultural contextualisation in multilingual knowledge repository Wikipedia

by Anna SAMOILENKO

Abstract

As a multilingual system, Wikipedia provides many challenges for academics and engineers alike. One such challenge is cultural contextualisation of Wikipedia content, and the lack of approaches to effectively quantify it. Additionally, what seems to lack is the intent of establishing sound computational practices and frameworks for measuring cultural variations in the data. Current approaches seem to mostly be dictated by the data availability, which makes it difficult to apply them in other contexts. Another common drawback is that they rarely scale due to a significant qualitative or translation effort.

To address these limitations, this thesis develops and tests two modular quantitative approaches. They are aimed at quantifying culture-related phenomena in systems which rely on multilingual user-generated content. In particular, they allow to: (1) operationalise a custom concept of culture in a system; (2) quantify and compare culture-specific content- or coverage biases in such a system; and (3) map a large scale landscape of shared cultural interests and focal points.

Empirical validation of these approaches is split into two parts. First, *an approach to mapping Wikipedia communities of shared co-editing interests* is validated on two large Wikipedia datasets comprising multilateral geopolitical and linguistic editor communities. Both datasets reveal measurable clusters of consistent co-editing interest, and computationally confirm that these clusters correspond to existing colonial, religious, socio-economic, and geographical ties.

Second, *an approach to quantifying content differences* is validated on a multilingual Wikipedia dataset, and a multi-platform (Wikipedia and Encyclopedia Britannica) dataset. Both are limited to a selected knowledge domain of national history. This analysis allows, for the first time on the large scale, to quantify and visualise the distribution of historical focal points in the articles on national histories. All results are cross-validated either by domain experts, or external datasets.

Main thesis contributions. This thesis: (1) presents an effort to formalise the process of measuring cultural variations in user-generated data; (2) introduces and tests two novel approaches to quantifying cultural contextualisation in multilingual data; (3) synthesises a valuable overview of literature on defining and quantifying culture; (4) provides important empirical insights on the effect of culture on Wikipedia content and coverage; demonstrates that Wikipedia is not context-free, and these differences should not be treated as noise, but rather, as an important feature of the data. (5) makes practical service contributions through sharing data and visualisations.

Zusammennfassung

Als mehrsprachiges System stellt Wikipedia viele Herausforderungen sowohl an Akademiker als auch an Ingenieure. Eine dieser Herausforderungen ist die kulturelle Kontextualisierung der Wikipedia-Inhalte und der Mangel an Ansätzen zu ihrer effektiven Quantifizierung. Außerdem scheint es an der Absicht zu fehlen, solide Berechnungspraktiken und Rahmenbedingungen für die Messung kultureller Variationen in dem Datenmaterial zu schaffen. Die derzeitigen Ansätze scheinen hauptsächlich von der Datenverfügbarkeit diktiert zu werden, was ihre Anwendung in anderen Kontexten erschwert. Ein weiterer häufiger Nachteil ist, dass sie aufgrund eines erheblichen qualitativen oder Übersetzungsaufwands selten skalieren.

Um diesen Einschränkungen zu begegnen, werden in dieser Arbeit zwei modulare quantitative Ansätze entwickelt und getestet. Sie zielen darauf ab, kulturbezogene Phänomene in Systemen zu quantifizieren, die auf mehrsprachigem, nutzergeneriertem Inhalt beruhen. Insbesondere ermöglichen sie es: (1) einen benutzerdefinierten Kulturbegriff in einem System zu operationalisieren; (2) kulturspezifische Inhalts- oder Abdeckungsverzerrungen in einem solchen System zu quantifizieren und zu vergleichen; und (3) eine großräumige Landschaft mit gemeinsamen kulturellen Interessen und Schwerpunkten abzubilden.

Die empirische Validierung dieser Ansätze ist in zwei Teile gegliedert. Erstens wird ein Ansatz zur Kartierung von Wikipedia-Gemeinschaften mit gemeinsamen redaktionellen Interessen auf zwei großen Wikipedia-Datensätzen validiert, die multilaterale geopolitische und sprachliche Redakteursgemeinschaften umfassen. Beide Datensätze zeigen messbare Cluster von konsistenten Mitredaktionsinteressen und bestätigen rechnerisch, dass diese Cluster mit bestehenden kolonialen, religiösen, sozioökonomischen und geographischen Bindungen übereinstimmen.

Zweitens wird ein Ansatz zur Quantifizierung von Inhaltsunterschieden anhand eines mehrsprachigen Wikipedia-Datensatzes und eines Multiplattform-Datensatzes (Wikipedia und Encyclopedia Britannica) validiert. Beide sind auf einen ausgewählten Wissensbereich der Nationalgeschichte beschränkt. Diese Analyse ermöglicht es erstmals im großen Maßstab, die Verteilung der historischen Schwerpunkte in den Artikeln zur Nationalgeschichte zu quantifizieren und zu visualisieren. Alle Ergebnisse werden entweder von Fachexperten oder von externen Datensätzen kreuzvalidiert.

Die wichtigsten Beiträge der Dissertation. Diese Dissertation: (1) stellt einen Versuch dar, den Prozess der Messung kultureller Variationen in nutzergeneriertem Datenmaterial zu formalisieren; (2) stellt zwei neue Ansätze zur Quantifizierung der kulturellen Kontextualisierung in mehrsprachigem Datenmaterial vor und testet sie; (3) schafft einen wertvollen Überblick über die Literatur zur Definition und Quantifizierung von Kultur; (4) liefert wichtige empirische Erkenntnisse über die Wirkung von Kultur auf den Inhalt und die Abdeckung von Wikipedia; zeigt, dass Wikipedia nicht kontextfrei ist, und dass diese Unterschiede nicht als Rauschen, sondern als ein wichtiges Merkmal des Datenmaterials behandelt werden sollten. (5) leistet einen praktischen Beitrag durch das Teilen von Datenmaterial und Visualisierungen.

Contents

Chapter 1

Introduction

Modern techno-social systems are intrinsically complex and interesting to study. With the proliferation of communication and mobile technologies, it became possible to collect large amounts of traces left by human activity on the Web. Either created intentionally by the user (user-generated content, UGC), or generated as a by-product of automatic collection of users' meta data and logs, these data are the 'digital footprints' of the modern society. They contain information about human movement, monetary and trade flows; friendship-, professional-, and collaborative ties; circadian patterns of activities, knowledge organisation and spread of innovative ideas, opinion dynamics, and many more aspects of collective behaviour.

Large amounts and high granularity of user-generated data open many exciting, unprecedented opportunities. Growing evidence demonstrates that these online traces are interconnected with the real life economic, social, and political outcomes. Mining them provides real-time estimations of various large-scale societal processes, from epidemic disease spreading to current touristic mobility interests, to public opinion during political events, such as elections and protests. Being able to extract the relevant signal from the avalanche of user-generated data has become a highly-valued skill on the modern employment market, both academic and commercial. In academic context, an entire new field of Computational Social Science or Social Physics is forming around the idea that the digital records of collective online activity could provide insights into collective behaviour of people offline, and even predict it. In commercial applications, firms and brokers who are able to timely obtain, analyse, and monetise online activity data are privileged with measurable economic and strategic advantages. Most successful of the modern businesses are those who harvest UGC to improve the quality of their services, such as search engines and recommendation systems. Finally, UGC has become indispensable at the intersection of (Computer) Science and Engineering. Many practical, cost- and time- efficient solutions to applied problems come from such areas as artificial intelligence and machine learning, natural language processing, and information retrieval, and critically depend on UGC.

Problem statement. Together with opportunities and innovative applications that UGC have opened for modern science and industry, several challenges arise. This work focuses on one of such challenges, which the author believes to be of particular importance: the relation between UGC and culture. As more users of diverse backgrounds go online, the content on the Web is becoming increasingly multilingual and multipolar, reflecting its contributors. In the literature, this phenomenon has been referred to as *cultural bias* or *cultural contextualisation* of the UGC. Although the amount of evidence illustrating the interconnection between UGC and cultural background of its contributors is mounting up, our understanding of the subject remains sparse and limited to case studies.

Based on the review of the current literature, presented in Chapter 2, this thesis identifies the following gaps in the literature on cultural effects in UGC (also discussed in Section 2.6).

- First of all, operationalising culture and quantifying culture-related variation in UGC are relatively new topics in computational domains. This presents methodological challenges at a fundamental level. In Computer Science community, there is a lack of established procedures for measuring cultural contextualisation in UGC. There is a need to move away from

the practice of letting the data dictate the approach, to developing novel frameworks that are applicable in multiple contexts.

- Secondly, current literature on cultural contextualisation is thin on scalable computational frameworks. It is rare to see studies which go beyond comparing variations across several selected communities. Additional research is needed in developing approaches that allow to zoom out, and compute the landscape of multilateral culture-related relationships in UGC.

- Finally, although the empirical evidence on cultural contextualisation is mounting, it is still a domain in need of expanding. Little is known about the large-scale effects that cultural context has on the scope of UGC, in particular, in multilingual environments like Wikipedia. Specifically, there is a lack of empirical evidence describing the global outlines of culture-related similarities and differences in UGC. Additionally, it is unclear whether the content generated by experts is substantially different from various cultural perspectives in UGC.

Objectives. This thesis aims at providing tools for facilitating future research concerned with the analysis of cultural aspects in UGC. It does so by addressing the above-mentioned gaps in the current literature on the subject. Of special interest is providing methodological frameworks, as well as operationalisation techniques for detecting and quantifying culture-related phenomena in UGC. Additionally, this thesis has an objective of investigating the Encyclopedia Wikipedia as a particularly impactful and popular example of UGC. In particular, this work aims at answering a number of empirical questions which help to push forward the current understanding of culture-related differences and similarities in multilingual UGC.

Scope of work. I start with a detailed overview of the literature presented in Chapter 2. Through this, I establish the key concepts relevant for this work, explain my motivations for studying Wikipedia, and identify the gaps in current research which drive this work. Additionally, I gather from different domains, an extensive overview on the attempts to define and quantify culture. Although admittedly not a complete reference, this overview is meant to be useful to other computational researchers who seek a literature-motivated way to operationalise culture-related phenomena in their research.

In this thesis, I present two modular approaches for quantifying cultural effects in UGC: they are introduced in Sections 3.1 and 4.1. Each approach is validated through two case studies of multilingual, large-scale datasets. Empirically, this thesis encompasses two major themes: identifying *similarities* (Chapter 3) and *differences* (Chapter 4) across the content generated by cultural communities of users in multilingual Wikipedia.

Chapter 3 presents an approach to quantifying and visualising the global ties of shared information interest across cultural communities. The approach is based on a combination of statistical filtering, and methods from Network Science. It is neutral to the knowledge domain, language-agnostic, and scalable to an arbitrary large number of communities. It is also modular and can be easily applied in other domains. The approach is validated on two large datasets and two definitions of cultural communities, reflecting linguistic and national communities of Wikipedia editors. Each dataset comprises a myriad of multilingual Wikipedia edits. In both case studies, the approach successfully depicts large-scale clusters of aggregated information interests, and identifies bridges between them. Moreover, it demonstrates through testing that these information preservation choices are culturally contextualised. In fact, despite globalisation, information interests of cultural communities on Wikipedia remain diverse. Precisely, I show that they are shaped by a number of social, geopolitical, linguistic, historical, and economic factors. Empirically, these results are the first comprehensive, large-scale analysis of such nature. One novelty of these analyses is that they embrace multiple Wikipedia language editions in their entirety, rather than focusing on a smaller subset. Thus, I map the global ties of shared information interests, for the first time at such scale.

Chapter 4 takes a closer look at how linguistic points of view are reflected in Wikipedia articles. Here, I present an approach (Section 4.1) to quantifying content differences in culturally contextualised, multilingual UGC. In the interest of capturing the nuances of differences in most detail, I validate the approach by narrowing the empirical scope down to one specific knowledge domain - articles on national histories. However, in principle the approach is extendable to other domains. The central idea of the approach is to reduce the notion of history to a single quantifiable *unit of comparison* – year mentions – which has equivalent meaning across all the studied narratives regardless of their language. This approach is successfully tested on a multilingual Wikipedia dataset, and, to illustrate that it has applications beyond the Wikipedia data, applied to a dataset of historiographical writing by the experts of Encyclopedia Britannica. To showcase the modularity of the approach, it is extended to include not only temporal analysis of national focal points, but also linguistic and semantic features. Finally, the validity and empirical value of this approach is confirmed by history experts. The empirical results uncover multiple biases across the scope of the analysed narratives. Some of these are to a large extent shared across all multilingual narratives, while others are only present among certain blocs or even between certain pairs of linguistic communities. Additionally, this chapter throws light on the previously unknown differences between public- and expert-written narratives.In particular, I empirically demonstrate that when it comes to historiography, the experts' writing leans towards spacial and territorial concepts, with emphasis on religious and cultural tensions. At the same time, the popular accounts disproportionally focus on wars and violent conflicts. While both perspectives are factually correct, the results remind the reader that there is no ground truth outside of cultural context.

1.1 Research questions, proposed approaches, results

One of the main purposes of this work is to present a series of approaches to quantifying culture-related phenomena in UGC. In validating these approaches, this thesis also addresses a series of empirical questions. Specifically, it focuses on collective knowledge production in Encyclopedia Wikipedia, as a particularly rich and relevant example of culturally contextualised UGC.

The empirical part of this work is split into two major research themes: *(1) mapping the emerging patterns of shared information interests across cultural communities of editors on Wikipedia* (Chapter 3); and *(2) comparing the content of encyclopedic articles across cultural communities* (Chapter 4). In this section, I give a detailed outline of research questions motivating each of the studies, outline the methodological details of proposed approaches, as well as summarise the main findings.

1.1.1 Mapping communities of shared information interest

Cultural globalisation and the world without borders have become popular notions in the modern mindset. With proliferation of communication technologies and faster, more affordable transport, it has become much easier to get the message across. Ideas, fashions, innovations, opinions, and commercial brands seem to diffuse around the planet with unprecedented speed, transcending national and linguistic barriers. Does it mean that the world of ideas is becoming increasingly homogeneous? Are cultural barriers to information exchange falling down? It is curious to imagine the world where all cultural communities became uniform with regards to their collective interest. It would mean that their shared interests converged to a specific set of universally known concepts. In Chapter 3, I investigate whether this is the case.

Multilingual encyclopedia Wikipedia presents a perfect opportunity to get an impression on how culturally proximate or remote communities are with regards to their information interests. In Chapter 3, I show how to map the global landscape of shared information interests across cultural communities. In particular, I focus on geopolitical and linguistic cultural communities. In the presented studies, these communities are approximated by the language and geographical location of Wikipedia editors. Additionally, I investigate whether cultural communities group, based on their strong shared interest in specific concepts, and what explains the presence of such common interests.

Chapter 3 opens by formalising the computational model, the essence of the approach. The model is then applied to quantify bilateral shared information interests. Section 3.2 investigates cultural communities approximated through the geographical location of Wikipedia editors. The specific research questions that I address include:

- How to quantitatively construct a network of shared information interests based on large-scale multilingual Wikipedia editing data?
- What factors best explain the strength of bilateral ties and formation of clusters?

Geopolitical belonging is not the only way to approximate cultural borders. In Section 3.3, I formalize cultural communities through language. This section addresses the following research questions:

- Is the set of languages covering a concept of Wikipedia random?
- Do certain editions show consistent interest in editing the same concepts?
- What socio-linguistic features explain common editing interests between language communities on Wikipedia?

Approach summary. The approach described in Chapter 3 defines *shared information interest* as a significant interest of Wikipedia editor communities (both geographical and linguistic) in

editing articles about the same topics. The borders of language communities are defined by the language editions of Wikipedia, and the geographical communities refer to the countries from which the editors are contributing, regardless of the language. The approach consists of several steps. I first use statistical filtering to identify language or country pairs which show consistent interest in articles on the same topics. Based on this dyadic information, I create a network of interest similarity where nodes are languages or countries, and links are weighted as the strength of shared interest. Then I cluster the network and inspect it visually to inform the generation of hypotheses about the mechanisms that contribute to interest similarity. Finally, these hypotheses are expressed as transition probability matrices, and I test their plausibility using two statistical inference techniques – HypTrails (Singer et al., 2015) and MRQAP (Krackardt, 1987) (Multiple Regression Quadratic Assignment Procedure). Using both Bayesian and frequentist approaches, I obtain similar results. This suggests that these findings are robust against the chosen statistical measure. Thus, the empirical results are validated by external datasets, as well as by their correspondence with the existing literature. See Section 3.1 for the details of this approach.

Main findings summary. The first empirical study, presented in Section 3.2 examines *shared interests of geopolitical communities*. It focuses on the contributions by unregistered editors whose IP addresses can be mapped to a specific country. The analysis provides a world map of shared information interests. Structural analysis of the underlying network shows that information interests are indeed not homogeneous, but split into 18 strongly interconnected country clusters. These clusters can be explained by factors related to language, religion, trade volumes, geographical proximity, and historical background, such as colonial past.

The analysis presented in Section 3.3 provides further answers to the empirical investigation of global shared information interests, this time, exploring in detail *interest profiles of linguistic communities*. I construct a large-scale network of interest similarities between 110 language communities, which are polarised into 24 linguistic clusters. This network structure is partially explained by several sociocultural factors, including shared religion, bilinguality, linguistic and geographical proximity of languages, and population attraction. Finally, this section shows that the set of language editions covering a concept on Wikipedia is not a random choice.

1.1.2 Quantifying content differences in a specific knowledge domain

While studying cross-cultural similarities of information interests is an intriguing question, it is sometimes even more interesting to investigate where the differences in content representation lay. Wikipedia articles exist in multiple languages. Since some topics are covered by multiple language editions, how much content do these editions borrow from each other? Do they present a similar view on the topic, or contradict each other? If so, which facts are omitted, and what could explain this? Educated guesses bring contradicting intuitions. On the one hand, encyclopedias consist of facts, and at its core Wikipedia's encyclopedic content should be to a large degree universal. On the other hand, Wikipedia is written by volunteers and in the absence of an editing authority. This suggests that its content is a constant working progress, not free of gaps and over-/under-emphasising. Furthermore, how does this content, produced by knowledgeable enthusiasts, compare to the work of professionals, who are paid to write encyclopedic articles? And finally, how can all these content differences be operationalised and quantified in a large-scale, multilingual setting?

These are the general questions with which Chapter 4 preoccupies itself. In order to answer them, I narrow my inquiry to a single knowledge domain – writing about history, or historiography. Historiography is an interesting case for a cross-cultural comparison. It is at the center of all social groups, from community clubs to entire nations, providing the feelings of roots, belonging, and identity. The chapter's narrative is split into several sections. In Section 4.2, I focus on quantifying Wikipedians' narratives on national histories, and compare them across multiple language editions. In particular, I answer the following research questions:

- What are the most documented periods of history of the last 1,000 years in Wikipedia?
- What are the temporal focal points in the descriptions of national histories in Wikipedia?
- Are country timelines consistent across language editions?

In Section 4.3, I extend the inquiry beyond Wikipedia, and compare historiographical writing by Wikipedia's volunteers, with the articles in Encyclopedia Britannica written by professional historians. My research questions include:

- How do the descriptions of national histories in English Wikipedia compare to the corresponding articles in Encyclopedia Britannica?
- What are the differences in the temporal and topical aspects of coverage, and in linguistic presentation of the material?

Approach summary. In both research projects I apply a similar computational approach to the analysis of textual historiographical data. In doing so, I demonstrate that this approach is suited for large-scale comparative studies. In particular, I focus on Wikipedia articles on all UN member states in 30 language editions. The approach concentrates on *year dates* as accessible representations of more complex historical structures. To be able to compare the descriptions across languages, I retrieve from article texts all date mentions (in the form of 4-digit numbers between 1000-2016), and use them as a language-independent unit of comparison (Rüsen, 1996). I propose a simple randomisation technique to extract *significant focal points of national histories* – time periods of significantly high mentions, compared to a random expectation model. I combine visual interpolation and expertise of invited history experts in order to evaluate how the results of this approach compare with the existing historical knowledge. I use hierarchical clustering to group countries whose histories are represented similarly on Wikipedia. Finally, I compute inter-language agreement on history of each country, using the Jensen-Shannon divergence measure. To compare linguistic features, I compute text statistics, apply a range of well-established readability tests, and run a Part of Speech analysis. The empirical results are validated by history experts. Extended details of this approach can be found in Section 4.1.

Main findings summary. The analysis in Section 4.2 provides insights on the national historiographic narratives in 30 language editions of Wikipedia. It demonstrates that Wikipedia narratives about national histories are distributed unevenly across the continents, with significant focus on the history of European countries (*Eurocentric bias*). Moreover, the clusters of countries with similarly distributed focal points map well to geopolitical blocs. Finally, mapping countries according to their Jensen-Shannon divergence scores shows that the national historical timelines vary across language editions, although average interlingual consensus is rather high.

The second case study, presented in Section 4.3, compares Wikipedia's crowd-sourced historiographical narratives with those written by professional historians for Encyclopedia Britannica. My research finds out, that Wikipedia leans to presenting history as a sequence of political events, putting a disproportional emphasis on periods of war and violent conflicts, with a specific preference to the events well-known to the general public. At the same time, Britannica is concerned with a more spatial and territorial concept of the history of states, emphasising the conflicts with underlying religious or cultural tensions. These differences are also reflected in the semantic analysis, which shows that Wikipedia relies on political and military words, while Britannica is heavy on vocabulary with religious connotations and on geographical terms.

1.2 Thesis contributions

This work's contributions are discussed in several places throughout the thesis. Current section outlines the contributions in detail. Abstract gives a short summary of the main contributions. Individual contributions of each empirical study are discussed in:

- Section 3.2 (p.42),
- Section 3.3 (p.50),
- Section 4.2 (p.73), and
- Section 4.3, (p.84).

This thesis makes several important contributions to the existing research on cultural contextualisation of UGC and quantifying culture in computational fields.

- First and foremost, this thesis contributes **two validated approaches** to quantifying cultural effects in UGC. Moreover, this work aims to be didactic to the extent that its approaches can be easily reproduced and adapted by researchers in other domains. Additionally, several features make the approaches unique and valuable to the research community:
  - they are tailored for comparison of multilingual, culturally contextualised digital data.
  - They are flexible and modular. This allows to easily extend and re-adapt them for applications in various contexts outside the experiments presented in this thesis.
  - They encompass a broad spectrum of different methods. This allows researchers to gain diverse, holistic insights into potential culture-related effects in multilingual UGC.
- To continue, this thesis contains a series of important **empirical findings**. It, thus, contributes to hitherto sparse computational literature on quantifying cultural contextualisation in UGC, and in multilingual Wikipedia in particular. This work:
  - provides important empirical insights on how culture shapes Wikipedia content and coverage. It demonstrates that Wikipedia is not context-free. It contains gaps and biases, and thus, does not represent 'universal ground truth'.
  - shows that the culture-related local differences in UGC are omnipresent and not random, and should be treated rather as a feature of the content.
  - is one of the first which visualises the landscape of multipolar cultural relationships on a truly large-scale. This provides a unique insight into the global effects of cultural context on UGC.
  - demonstrates that global cultural interconnections are not dominated by one powerful player, but instead form locally established blocs. It also validates that Wikipedia data can be successfully used to get insight into global, intercultural relationships.
- Additionally, this thesis presents **one of the first efforts to formalise the process of measuring cultural variations in user-generated data**. It fills a specific, and so far rather thin niche, in this literature. Particularly, this work uniquely combines computational methods typical elsewhere in the computational literature, with qualitative evaluation and thorough theoretical grounding, more common in Anthropology and Social Sciences. This practice is novel and is not yet a standard in computational fields, but it is critical when quantifying culture-related effects. This is validated by domain knowledge experts and external socio-demographic datasets.

- Finally, this work has resulted in several **practical service contributions to the research community**:
  - The datasets collected for the empirical studies and produced during the analysis are made publicly available.
  - Visualisations of national timelines with historical focal points are made publicly available. This might be useful, for example, to historiography researchers, who wish to conduct additional investigations, but do not necessarily want to work with the raw data.
  - This thesis also provides a useful summary of existing research on defining and quantifying culture, with lessons across fields. This overview might be useful for researchers working on similar topics.

1.3 Publications related to this thesis and author contributions

This section outlines the list of publications related to this cumulative thesis, as well as elaborates in detail on my own contributions to each study.

The following project was born during my research visit to the Ice Lab in Umea University, Sweden, and the idea stems from discussions between Martin Rosvall, Andrea Lancichinetti, Fariba Karimi, Ludwig Bohlin, and myself. The main idea of extracting communities of shared interest based on Wikipedia editing activity is preserved in both articles. The first article defines the communities of interest as separated by country borders, and this part of research was primarily lead by Fariba Karimi. I was responsible for leading a parallel part of the investigation, which focused on the multilingual side of Wikipedia and defined inter-community borders via languages.

- **Article 1:** [(Karimi et al., 2015)] Karimi F., Bohlin L., Samoilenko A., Rosvall M., Lancichinetti A. (2015) Mapping bilateral information interests using the activity of Wikipedia editors. *Palgrave Communications* 1, 15041. doi:10.1057/palcomms.2015.41

  First and foremost, I participated in developing the empirical framework of the study and grounding it in the existing literature. I took direct participation in the development of the statistical filtering method introduced in this paper, as well as interpreted the resulting network, including the clustering outcomes. In terms of the multiple regression analysis, I was responsible for collecting and formalising the data for some of the analysed hypotheses.

  The project was primarily driven by Fariba Karimi, but all authors actively contributed to the elaboration and testing of the the filtering method, interpreted the results, and wrote parts of the manuscript.

- **Article 2:** [(Samoilenko et al., 2016)] Samoilenko A., Karimi F., Edler D., Kunegis J., Strohmaier M. (2016) Linguistic neighbourhoods: Explaining cultural borders on Wikipedia through multilingual co-editing activity. *EPJ Data Science* 5 (9). doi:10.1140/epjds/s13688-016-0070-8

  I was responsible for driving this project, from establishing the rhetoric and empirical foundation of the research questions, through data acquisition and analysis, to visualisation, interpretation, and communication of the results. First of all, I developed the framework for sampling the language editions and selecting the time frame of the analysed data, wrote the necessary scripting for retrieving the data from the live servers of Wikimedia Foundation, cleaned and merged the data. The statistical filtering method applied for this project is based on the work introduced in the Article 1, and for this project I wrote my own Python implementation of it. Network visualisation and clustering were finetuned by myself in coordination with Daniel Edler, who consulted me on applying the Infomap software that he developed. The idea to apply the HypTrails approach for testing hypotheses about the network edges stems from discussions with Markus Strohmaier. Philipp Singer is acknowledged for developing the openly available Python implementation of HypTrails which I used. He also consulted me on the inner workarounds and data normalisation for HypTrails. Finally, the idea to apply the MRQAP as a parallel framework to explaining the network edges was inspired by a talk given by Michael Macy at the first International Conference on Computational Social Science in Helsinki, Finland.

  All the authors were constantly involved in discussions regarding the analysis and the interpretation of the results, as well as the project framing. All authors contributed to writing the manuscript.

The remaining publications are a part of a project on computing historiography which was born at the GESIS Off-Campus Meeting in La-Roche-en-Ardenne, Belgium. The initial idea stems from discussions between Katrin Weller, Florian Lemmerich, and myself.

- **Article 3:** [(Samoilenko et al., 2017)] Samoilenko A., Lemmerich F., Weller K., Zens M., Strohmaier M. (2017) Analysing Timelines of National Histories across Wikipedia Editions: A Comparative Computational Approach. *Proceedings of the Eleventh International AAAI Conference on Web an Social Media (ICWSM 2017)*. 15-19 May 2017. Montreal, Canada. Pages (210-219)

  In this project, Florian Lemmerich and I worked closely on developing a statistical approach to quantifying digital narratives about history based on the multilingual Wikipedia data, and in particular, defining the model for statistical extraction of the national focal points. I was responsible for collecting the data, extracting the date mentions, performing all statistical analyses, and visalising the results. Additionally, I set up the evaluation procedure and performed the related computations. I also made all the collected data publically available online (see the Dataset below) in pre-processed and cleaned format.

  All authors met regularly to discuss the progress and refine the methodological framework. The project benefited a lot from History-related expertise of Maria Zens, Katrin Weller, and Florian Lemmerich who worked closely on interpreting the extracted national focal points and identifying historical events corresponding to these time periods. I would like to also acknowledge the contribution of Sebastian Stier, Mathieu Génois and Mohsen Jadidi, who participated in some of the meetings and shared their ideas and exerience. All authors contributed to the writing of the manuscript.

- **Article 4:** [(Samoilenko et al., 2018)] Samoilenko A., Lemmerich F., Zens M., Jadidi M., Génois M., Strohmaier M. (Don't) Mention the War: A Comparison of Wikipedia and Britannica Articles on National Histories.

  The idea of this research belongs to myself, and is inspired by the work on the Article 3. The kind permission of the Editing Board of Encylopedia Britannica, which I procured, made this project real. I was responsible for obtaining the data, adapting the statistical model to extract significant national focal points, implementing statistical comparisons of temporal distributions, and performing the linguistic part of the analysis. I was also responsible for designing the evaluation procedure and assembling its results. Finally, I produced all visualisations.

  Florian Lemmerich, Mathieu Génois, Mohsen Jadidi, and myself had regular discussions regarding the statistical framework. Florian Lemmerich also conducted the top word usage part of the analysis. Maria Zens, Florian Lemmerich, Mathieu Génois and myself were directly involved into interpreting the extracted statistical differences to arrive at a comprehensive account of historiographical perspectives offered by both encyclopedias. All authors contributed to the writing of the manuscript.

- **Dataset:** [(Samoilenko, 2017)] Samoilenko, A. (2017): Multilingual historical narratives on Wikipedia. Version: 1. **GESIS Data Archive**. Dataset. http://doi.org/10.7802/1411

  This dataset is based on the data I collected for the Article 3 study (Samoilenko et al., 2017). The raw data encompasses the text of Wikipedia articles related to the national histories of all UN member states from 30 language editions. For this dataset, I extracted all date mentions from the raw data. The dataset is formatted into .csv files, and is available to other researchers for free at the GESIS Online Data Archive.

The papers presented in this section have been published at prestigious multidisciplinary journals such as EPJ Data Science and Palgrave Communications, and some of the top tier conferences in the field of Computational Social Science, such as ICWSM and WWW.

I have presented this work at various international conferences, such as:

- Web Science'15 in Oxford, UK;

- NetSciX'16 in Wroclaw, Poland;
- ICWSM'17 in Montreal, Canada;
- GESIS Winter Symposia in 2015 and 2016,
- the IC2S2'17 in Cologne, Germany, and
- the First European Symposium on Societal Challenges in Computational Social Science in London, UK.

I have also been invited to give talks on this research at the seminars at:

- the University of Haifa, Israel (2017),
- the University of Torino, Italy (2017), and
- the University of Oxford, UK (2017).

Finally, I have designed a website [http://annsamoilenko.wixsite.com/homepage/projects] that gives a detailed overview of each project presented in this cumulative thesis. It also links to the available data and resources for free downloads. Additional visualisations which are not presented in the papers are also available there.

Overall, the research presented in this thesis has been positively received by the academic community, and has been complimented by anonymous reviewers on multiple occasions as an excellent example of Computational Social Science work. Article 2 has also received an award at the NetSciX conference in 2016.

1.4 Thesis structure

The rest of this thesis is structured as follows. I continue by outlining the necessary conceptual background in Chapter 2. This chapter gives definitions to the key concepts relevant to this work, such as UGC, culture, and cultural contextualisation in UGC. Additionally, it lists my motivations for studying Wikipedia as a particular example of UGC. It also gives a detailed overview of the methodologies that have been applied to measure cultural similarities and quantify culture-related effects in UGC.

Further chapters focus on presenting the approaches, and the empirical part of the work. In particular, Chapter 3 starts with the introduction of a computational approach for quantifying the similarity of shared interest in the multilingual collaborative context of Wikipedia (Section 3.1). Subsequent Sections 3.2 and 3.3 test the approach on two empirical studies. This chapter concludes with a discussion of empirical results, limitations, and a summary of implications.

Following it, Chapter 4 introduces another computational approach, this time focusing on quantifying similarities and differences in multilingual textual data. This chapter starts with a discussion on selecting an appropriate unit of comparison, around which the approach centers (Section 4.1). The approach is applied to quantifying historical narratives. The empirical part is presented by two studies. In Section 4.2, I test the approach by comparing historiographies of nations across 30 Wikipedia editions. Next, in Section 4.3, I extend the approach and compare crowd- and experts perspectives on historiographies. Similar to the previous chapter, these empirical studies are followed by a discussion of limitations, implications, and concluding remarks.

Finally, the last part of this manuscript, Chapter 5, presents the final thoughts of the author together with contributions and implications of the current research, as well as directions for future work. At last, this theses closes with the list of bibliographical references.

Chapter 2

Related work

"There is not much point in trying to say what culture is... What can be done, however, is to say what culture does."

Thornton (1987)

2.1 User-generated content (UGC)

User generated content (UGC) has become an indispensable part of the modern online experience. In 2006 TIME magazine named 'YOU' the person of the year. This manifestation acknowledged the fact that could no longer be ignored – both the computing community as well as the public at large were undergoing a silent revolution which Tim O'Reilly described as a switch from Web 1.0 towards Web 2.0 (O'Reilly, Tim, 2005). The core ideas around Web 2.0 focus on building architectures that allow richer user experience, direct engagement with the content of the Web, and participation in making the Web more valuable. In particular, such architectures assume a shift beyond the level of a single device towards building distributed platforms that aim at harnessing collective intelligence and promoting collaborative value creation. Blogs, forums, social networking websites, and various wikis are the most prominent examples of UGC which arose thanks to this technological shift. They represent the philosophy and technology of collective participation, co-creation, co-consumption, and sharing of various expertise online.

Motivation for studying UGC. UGC has affected in many profound ways modern economy, business, engineering and Computer Science, and even the way academic research itself is conducted. UGC has become heart and brains of many modern computing systems, and an indispensable part of daily reality for anyone with online connection. The shift towards new participatory technologies has revolutionised computing by putting 'YOU', the user, and more precisely, user-generated content (UGC) in the center of the digital universe. Twitter, YouTube, Reddit, Wikipedia, Facebook, Amazon, StackOverflow, OpenStreetMap, Yandex.Traffic, AirBnB, TripAdviser, Yelp, Foursquare are just a few examples of businesses which became possible thanks to collaborative participation and "produsage" (Bruns, 2008) of massive amounts of individuals. While the commercial value of these businesses is substantial, one may argue that it almost entirely depends of the contributions of their users, just like the users co-depend on these technologies and services to exist. Nowadays UGC is increasingly used to improve the value of the services offered by various businesses, and the more users turn to and generate content for the platform, the better value they are getting out of using it. Nevertheless, it is not only the end users and digital businesses who benefit from the UGC boom. UGC has become the life blood of the algorithms behind these platforms, and an indispensable source of data in Computer Science and Software Development communities (Baeza-Yates, 2009). Although the quality of UGC is often questioned, it is compensated by its quantity. In fact, already in 2007 Ramakrishnan and Tomkins estimated that UGC generated daily from 8 to 10GB while the professional Web only generates 2GB in the same time. Recent advances in such areas as artificial intelligence, machine learning and recommendation systems, information retrieval, topic modeling, and natural language processing, as well as

their applications owe a lot to this sudden avalanche of UGC and social bookmarking and folksonomies in particular. Additionally, UGC has impacted the way academic research is conducted. Such Web portals like Zooniverse harnesses human intelligence on a large scale by encouraging volunteers to participate in scientific research by solving small tasks such as ranking, classification, and annotation of digital data. Such initiatives allow to optimise and accelerate current research practices, and have already resulted in more than 100 academic publications Zooniverse.org (2018) in the fields of Astronomy, Humanities, Ecology, and Biology.

Given the impact and importance of UGC in many areas of modern life, business, and science, from the academic standpoint, there are many unanswered questions around UGC. This thesis focuses on one group of such questions which are concerned with the *content itself*. Generally speaking, this stream of research typically investigates the quality, topical and geographical coverage, focal points, and imbalances in UGC. This particular work investigates the relationship between UGC and the cultural background of the contributors. More precisely, this thesis develops the frameworks and approaches which aim at quantifying the imbalances, similarities, and differences in UGC which are an effect of the culture-related characteristics of the communities of users who "produse" this content.

Conceptual definition of UGC. The term UGC has become widely adopted (Van Dijck, 2009). Nevertheless, several researchers have conceptualsed a very similar idea under different names: Web 2.0 (Chadwick and Howard, 2009), produsage (Bruns, 2008), citizen journalism (Bruns, 2005), participatory news (Deuze et al., 2007), user-generated media (Shao, 2009), user-created content (Vickery and Wunsch-Vincent, 2007), and at least further 43 names, according to a recent analysis (Dylko and McCluskey, 2012). It is difficult to agree on a common definition of UGC, multiple authors have emphasised its different characteristics. First of all, UGC is characterised by the presence of active users who voluntary nominate themselves as contributors. For example, Bruns describes "produsage", a very similar concept to UGC, as "collaborative and continuous building and extending of existing content in pursuit of further development" (Bruns, 2008, p.21). Secondly, user contributions may vary in degree and format, and be very narrow; however a large number of people can produce meaningful outcomes through constantly interacting with the content. In his *Infotopia*, Sunstein deliberates on an idea that Chadwick and Howard consider fundamental for understanding UGC: collective intelligence. This idea suggests that amateurs working together and voluntarily often produce content of better quality than paid experts working alone. Thirdly, UGC is always unfinished and characterised by constant refinement and experimentation. Moreover, unfinished content does not mean bad content (Chadwick and Howard, 2009; Jarvis, 2009). Rather, publishing incomplete material invites future participation. Finally, Vickery and Wunsch-Vincent writes that UGC should reflect a creative effort of the user and be created outside of the user's professional work. Also, UGC is not owned, published online, and openly accessible to everyone (Bruns, 2008; Vickery and Wunsch-Vincent, 2007). Krumm et al. writes about UGC as the content which "comes from regular people who voluntarily contribute data, information, or media that then appears before others in a useful or entertaining way, usually on the Web".

Aggregating several trends and characteristics which emerge from the literature, this thesis adopts the following definition of UGC: *UGC is a) an information product that is b) published online and openly available, c) created through large numbers of contributions by multiple users d) working collaboratively, voluntarily, and outside of their professional routines*.

2.2 Motivation for studying Wikipedia

Wikipedia is probably one of the most interesting examples of techno-social systems built entirely on UGC that a researcher could study. This online encyclopedia is being written by volunteers and in the absence of any editing authority, in multiple languages, and in real time. Technologically, Wikipedia is a platform created around the idea of participatory knowledge accumulation. It is permanently open to change and improvement, and represents a constant working progress. While there are many prominent examples of UGC, it is hard to find a platform with more impact on the real world than Wikipedia. For years Wikipedia has been one of the most accessed platforms online (Silverwood-Cope, 2012; Zickuhr and Rainie, 2011). As of 2018, it is among the top five visited websites globally[1]. However its online popularity is not the only reason that attracts researchers to studying Wikipedia.

This section continues introducing the reader to the encyclopedia, and elaborates on why, despite being an online phenomenon, Wikipedia impacts daily offline life in many profound ways. In particular, the narrative focuses on Wikipedia's importance in four domains: its impact on a) education and academic research, b) Computer Science and modern algorithms, c) Business and Economy, and finally, d) Public Policy.

2.2.1 Wikipedia in education and academic research

Wikipedia has produced a profound impact on the academic community. Since its inception in 2001, Wikipedia has attracted researchers attention as a complex socio-technological system; a unique and successful example of ongoing massive human collaboration; a rich, open multilingual dataset; and a relevant source and storage of knowledge for both academics and lay readers.

The literature that Wikipedia has inspired is vast. A 2009 review (Okoli and Schabram, 2009) identified over 400 research studies which focus on the encyclopedia either as the major topic of research or as a source of data. The interest in Wikipedia has been growing ever since (Okoli et al., 2012, 2014), attracting researchers with various backgrounds. An entire field of Wikipedia studies has emerged as a result, investigating Wikipedia's content (Brown, 2011; Callahan and Herring, 2011; Halavais and Lackaff, 2008), credibility (Blumenstock, 2008; Giles, 2005), size (Lam and Riedl, 2011), technological infrastructure (Slattery, 2009), collaboration patterns (Brandes et al., 2009; Keegan et al., 2012; Pfeil et al., 2006), editing community (Ciffolilli, 2003; Gallus, 2016; Zhu et al., 2013b), contributor motivations (Kuznetsov, 2006; Xu and Li, 2015; Yang and Lai, 2010; Zhu et al., 2013a), readership (Antin and Cheshire, 2010; Heilman and West, 2015), and conflict dynamics (Kittur et al., 2007; Sumi et al., 2011; Yasseri et al., 2012b), to name just a few research directions. Apart from being the object of research, Wikipedia has gradually made its way into education (Head and Eisenberg, 2010; Judd and Kennedy, 2010; Weller et al., 2010) making a positive impact in classrooms. Among other effects, it has been found that Wikipedia editing can improve student learning and retention (Kennedy et al., 2015). More generally, Wikipedia is increasingly seen by academics as an important channel for public communication of science, and "possibly, the main source of knowledge for generations to come" (Jemielniak and Aibar, 2016), in- and outside of the classroom. It has also been found to amplify the diffusion of the scientific findings published in open access venues Teplitskiy et al. (2017). The impact of Wikipedia on academic community goes even beyond that. New causal evidence suggests that Wikipedia is not just a platform that provides access to knowledge, including scientific knowledge; it also shapes scientific agenda and the language that is used in academic publications (Thompson and Hanley, 2018). Finally, the way the researchers themselves are depicted by Wikipedia is discussed as a possible alternative metric for academic success (Samoilenko and Yasseri, 2014). This means Wikipedia might have tangible impact on academic careers and funding.

[1] http://www.alexa.com/siteinfo/wikipedia.org (accessed 16, October 2018)

2.2.2 Wikipedia in Computer Science and modern algorithms

Wikipedia is an impactful knowledge database in the broad context of information technology ecosystem, and it has probably become one of the most important datasets in contemporary computing. It is mainly utilised by other information systems in two broad ways: for content re-use, and as a training set for contemporary machine learning algorithms.

Many modern information systems automatically source their content from Wikipedia. It is frequently integrated with voice-activated systems such as Apple's Siri (Lardinois, 2016), Amazon's Echo and Alexa (Kensinger, 2015), and other intelligent personal assistants (Dale, 2015). Wikipedia content has extensive applications in a wide range of Computer Science sub-fields. Chat bot systems use Wikipedia content to improve the conversational abilities of artificial agents, for example, to handle rapid changes in topical threads Breuing et al. (2011) or to provide extra contextual knowledge (Breuing, 2010). Wikipedia content has also been used to improve the performance of sophisticated topic models (Coursey and Mihalcea, 2009; Yao et al., 2016), recommendation systems (Zhang et al., 2012), search queries (Devlin, 2015; McMahon et al., 2017; Singhal, 2012), and in (cross-lingual) information retrieval (Chen et al., 2017; Müller and Gurevych, 2008; Nguyen et al., 2008; Sorg and Cimiano, 2012).

Moreover, Wikipedia is widely used for constructing entire knowledge bases, notably, in seeding and automatically refining Web knowledge graphs (Paulheim, 2017). Such systems use not only the semi-structured article content of the encyclopedia, but mine the structured key-value pairs in Wikipedia's infoboxes, its category system, inter-entity linkage structure, and the links matching entities across languages. Some prominent systems relying on Wikipedia data include open large-scale knowledge graphs like Wikidata (Vrandečić and Krötzsch, 2014), DBpedia (Lehmann et al., 2015), YAGO (Kasneci et al., 2006), as well as their commercial analogues, such as Google's Knowledge Graph and Knowledge Vault (Dong et al., 2014; Singhal, 2012), Yahoo!'s Knowledge Graph (Paulheim, 2017), and Facebook's Entities Graph (Sun and Iyer, 2013). Wikipedia is knit tightly into the broad ecosystem of online platforms and services, having a tangible impact on other websites through content re-use. One of the most prominent examples of such relationship is the interconnection between Wikipedia and modern search engines, and Google in particular. Wikipedia's content makes Google a better search engine (McMahon et al., 2017), and this effect is likely to exceed the algorithmic improvements to the search. Wikipedia also provides substantial value to other online platforms, improving the quality, visitation, engagement, and even revenues of the websites such as Reddit and StackOverflow (Vincent et al., 2018).

Apart from useful content applications, Wikipedia also serves as a corpus for training state-of-the-art AI algorithms (Collier et al., 2018; Eckles and Bakshy, 2017; Ponza et al., 2018; West et al., 2012) and is also partially responsible for driving the methodological progress in the area (Nickel et al., 2016). Extraction of structured data from Wikipedia and similar knowledge repositories is an automatic process which is prone to inconsistencies. Detecting and solving these has become a separate branch of tasks for algorithm designers and engineers. In this, Wikipedia and its structured data offsprings like DBpedia serve both as source of training data, as well as inspiration for novel algorithmic approaches in, e.g. ontology enrichment and association rule mining (Jang et al., 2015; Lehmann and Bühmann, 2010; Töpper et al., 2012), outlier/anomaly detection in numerical data and knowledge graph interlinks (Fleischhacker et al., 2014; Paulheim, 2014; Péron et al., 2011; Wienand and Paulheim, 2014), and finding erroneous entity relations (Lehmann et al., 2012; Paulheim and Bizer, 2014).

2.2.3 Wikipedia in Business and Economy

Wikipedia has a real impact on Business and Economy. First of all, mere presence on Wikipedia has been proven to lead to positive outcomes for businesses, firms, and entire markets. It also has an effect on individuals and their private economic decisions, such as consumption of services and

even touristic mobility. For example, (Hinnosaar et al., 2017) has found that locations with better coverage on Wikipedia attract more tourists, and even small improvements of article content lead to up to a 9% increase in documented hotel stays. The authors report a potential 160,000 euro of additional annual tourist revenue for an average city in Spain, thanks to improved Wikipedia coverage only. (Aitken et al., 2014) found a correlation between drug sales and Wikipedia traffic to a subset of health-related articles. Similarly, Wikipedia usage logs have been found useful for forecasting box office revenues(de Silva and Compton, 2014; Mestyán et al., 2013) and stock performance (Cergol and Omladič, 2015; Liu et al., 2014; Moat et al., 2013, 2014; Wei and Wang, 2016) Another study (Boulton et al., 2018) has investigated Wikipedia's impact on the information environment around the firms going through the initial public offerings (IPO). The authors find that IPO firms with a Wikipedia article enjoy greater attention from investors and benefit from positive long-term effects compared to those without an article. In addition, (Xu and Zhang, 2013) examines the general value of Wikipedia for providing firm information. The study reports that Wikipedia articles on firms are important in the financial market, acting as a long-term, reliable source of information on companies' important milestones. The authors remark Wikipedia's role in bridging information asymmetries between managers, analysts, and investors, and thus improving the overall information climate in the financial market.

Secondly, Wikipedia usage logs are frequently re-used by other businesses and online platforms. Several studies utilise Wikipedia traffic patterns for building sophisticated financial trading algorithms and to get unique insights into the state of markets, and collective interest of decision-makers, such as traders and speculators. For instance, (Dickerson, 2018) focuses on the case of Bitcoin and reports a high correlation between the Wikipedia search volumes for 'Bitcoin' and the cryptocurrency's market price. The findings suggest that Wikipedia viewership statistics can be used to construct highly profitable trading algorithms, as well as to monitor for early warning signs, and even anticipate price bubbles and crashes. Additionally, (Kristoufek, 2013) reports on Wikipedia's role in enhancing public interest in Bitcoin, and through that, driving up the cryptocurrency's prices. This relationship is reciprocal, i.e. not only do the search levels drive the prices but also the prices influence the search. In other words, Wikipedia participates in a cycle where rising information interest in Bitcoin causes more people, including the general public, to gain interest, and thus driving real-world price bubbles and subsequent market crashes.

Finally, many online businesses and platforms re-use Wikipedia content to improve their services and enhance user experience. Wikipedia is known to improve the quality of search recommendations and is especially valuable in information retrieval (Devlin, 2015; Singhal, 2012). Wikipedia content re-use makes Google a better search engine, and this likely results in more positive outcomes than algorithmic improvements (McMahon et al., 2017). Several studies (Xu et al., 2015; Zhang et al., 2012) investigate contextual advertising strategies for online businesses, and report that leveraging Wikipedia content provides substantial improvement compared to existing online recommendation approaches. Moreover, such large-scale online communities as StackOverflow and Reddit are able to attract more visitations, keep the users engaged, and increase potential advertisement revenues through Wikipedia content re-use. Back-of-the-napkin revenue estimations in (Vincent et al., 2018) suggest that Wikipedia is responsible for generating up to $124,200 annual advertising revenues for Reddit and $180,900 - for StackOverflow.

These studies demonstrate that the online repository Wikipedia contributes to economic decision-making and has a tangible impact on real-life macro- and microeconomic outcomes. Moreover, information asymmetries in Wikipedia articles can hinder or enhance real life individual economic activity and business revenues. At the low cost of improving Wikipedia presence, the return on investments is high for businesses: expanding potential customer base, receiving more attention from investors, supporting a stable information climate around their activities, and eventually, increasing their revenues.

2.2.4 Wikipedia in Public policy

Wikipedia has tangible implications for public policy, first of all, as an effective tool for information dissemination, secondly, as a data source for approximating large-scale social dynamics, and thirdly, as a platform for social mobilisation and democratic debate which can influence real political outcomes.

First of all, Wikipedia is relevant for public policy as a tool for information dissemination. It is one of the most visible and attended websites, and is arguably the primary source of information online on any topic. This makes it valuable for various stakeholders who desire to communicate important information to a wider public, in a cost-effective way. For example, when it comes to the medical information and health care, Wikipedia is one of the most viewed resources globally (Heilman and West, 2015). Its information dissemination potential is larger than that of the World Health Organisation or the UK National Health Service platforms (Masukume et al., 2016; Murray, 2018). Wikipedia's reach, regardless of the quality of the content, makes it a major player in influencing the public thought on health information and related health choices (Rasberry, 2014). Thus, much research in health and public policy making calls for monitoring and curating the health-related content on Wikipedia(Azzam, 2017). Extending this idea beyond health-related topics, (Shafee et al., 2017) argues that Wikipedia should be used to improve public understanding of science and scientific literacy in general. Apart from informing the lay readers, Wikipedia is also relevant for the work of policy analysts, as it provides the necessary information in a more digestible form compared to the less comprehensible language of scientific publications(Gluckman, 2016). Finally, various cultural institutions and public libraries incorporate Wikipedia into their digital outreach strategies as an effective marketing tool to raise awareness of the existence of their resources (Perrin et al., 2017). Wikipedia has proven effective in boosting discoverability of digital collections, thus expanding the user base of the archival repositories, and generating a large number of new digital patrons(Galloway and DellaCorte, 2014; Lally and Dunford, 2007). Moreover, traffic driven to digital collections by Wikipedia outperforms by multiple times the traffic driven by other marketing strategies (Elder et al., 2012; Szajewski, 2013). Finally, linking repositories' resources with Wikipedia requires little or no maintenance, and provides a greater chance of reaching appropriate audiences(Sliger Krause et al., 2017).

Secondly, Wikipedia can be an interesting complementary data source for policy-making when it comes to monitoring and forecasting large-scale population trends, from seasonal diseases to public opinion and touristic mobility. In particular, Wikipedia viewership and editing volumes are used to extract information on trending news and events (Ahn et al., 2011; Althoff et al., 2013; Ciglan and Nørvåg, 2010; Kämpf et al., 2015), as well as to estimate the popularity of politicians and political parties (Yasseri and Bright, 2014, 2016). Wikipedia can begin to offer useful information for disease monitoring and mitigation. A large body of literature focuses on using Wikipedia access logs for the monitoring and forecasting of infectious disease spreading(Generous et al., 2014; Priedhorsky et al., 2017), especially when traditional population surveillance systems are not available in real time. Recent examples (Bardak and Tan, 2015; Hickmann et al., 2015; McIver and Brownstein, 2014) use Wikipedia, alone or in combination with other social media (Sharpe et al., 2016) to forecast influenza season, and allow for accurate prediction several weeks before the epidemic starts. Additionally, Wikipedia usage trends can be effectively used by public and private sectors in tourism planning and forecasting the tourism demands (Khadivi and Ramakrishnan, 2016)

Thirdly, Wikipedia serves not only information purposes but also as a platform for action, mobilisation, and democratic public debate. In 2011-12 Wikipedia together with other online communities and corporate websites united in a massive act of collaborative online protest against the online copyright infringement legislation — the Stop Online Piracy Act (SOPA) and the Protect IP act (PIPA). Initially playing a purely informational role, Wikipedia eventually became a platform for debate, attracting over 2,000 activists to the talk pages, and finally resulting in the decision to

black out all of its content for a day (Lee, 2012; Potter, 2012). As a result of the protests, the US Congress backed off and announced that the legislation would be postponed. While Wikipedia was one of many players in the protests, it was a distinct and successful one(Oz, 2012). It demonstrated in action a new model of direct democracy, and the potential for a non-profit platform like Wikipedia to host self-governed public deliberation, which lead to tangible political outcomes in the real world(Benkler et al., 2015; Powell, 2012).

2.3 Cultural contextualisation of UGC

The concept of cultural contextualisation in UGC has been coined by Hecht, p.47, who describes it as "the cause of some of the content diversity" in UGC. His original thesis mostly focuses on the examples of cultural contextualisation in multilingual Wikipedia, although several studies examine the same phenomenon in other projects. For example, demographic differences such as age and gender, as well as personality result in strikingly different language use on Facebook (Schwartz et al., 2013). Moreover, anger and fear emotions are more present in Twitter posts of users with higher income, while sadness, surprise and disgust emotions are more associated with lower income (Preoţiuc-Pietro et al., 2015). As shown through semantic tagging of culture-specific concepts, the patterns of interpersonal communication in YouTube comments vary across countries (Thakker et al., 2017). Additionally, a comparison of TripAdvisor hotel reviews between Chinese and English speakers reveals substantial differences both in structured (ratings) and unstructured (text) content features which the authors attribute to the collectivism/individualism cultural divide (Zhang et al., 2016). Over 45 percent of Flickr photos are local to the photographer (Hecht and Gergle, 2010b). Finally, in OpenStreetMap peer-produced content about rural areas is of systematically lower quality, likely because the participation of rural populations in peer production seems to be lower than in urban populations (Hecht, 2013; Johnson et al., 2016).

To sum, there is a growing literature on culture-related factors in UGC which has examined multiple online media and several approaches to culture, from socio-demographic factors like age, income, gender, and social status, to geopolitical factors like nationality and location, to language. Such interest towards the topic is explained by the relative novelty of the phenomena in the context of UGC. Cultural contextualisation naturally emerges in online projects which offer their users a certain degree of freedom in their contributions (Hecht and Gergle, 2009). Naturally, there is a growing need for new methods and approaches for extracting meaning from this unstructured content (Gandomi and Haider, 2015).

Cultural contextualisation in multilingual Wikipedia. Wikipedia is one of the most interesting examples of cultural contextualisation of UGC. The online encyclopedia is a prominent example of collective knowledge accumulation, and it is becoming one of the most interesting and convenient sources for academics to study cultural processes (Schich et al., 2014). Contributing to Wikipedia means more than writing encyclopedic content: it allows communities to store cultural memories of events (Keegan et al., 2011; Keegan, 2013; Pentzold, 2009), document their point of view (Massa and Scrinzi, 2011, 2012), and give prominence to people (Samoilenko and Yasseri, 2014). This collective sifting of culturally-relevant knowledge is such an important social process that conflicts and edit wars frequently emerge before reaching consensus (Yasseri et al., 2014). Wikipedia is one of the most linguistically diverse projects online, with a constant base of editors contributing in almost 300 languages (Wikipedia, 2016), ranging from almost 5M in the largest edition (English) to just 89 in Cree, the smallest one (Wikipedia, 2016). The language communities not yet represented on Wikipedia seek the inclusion as an opportunity to establish and promote their language and culture in the digital realm (Kornai, 2013). There are currently 160 open requests for new Wikipedia language editions in the Wikimedia Incubator (The Wikimedia Incubator, 2015).

There is no central authority that dictates which topics must be covered, and every editor is free to select their own, as long as they are consistent with the notability guidelines (Wikipedia, 2015). All language editions have their own notability guidelines and are edited independently from each other, although an editor can also co-edit several editions in parallel. Apart from this rather low percentage of multilingual editors (Hale, 2014b), these editing communities exist in relative isolation from each other, and are not forced by the platform to discuss the content choices or points of view introduced in the articles of each edition. As a result, even when articles on the same concept exist in different language editions, they are not translated replicas of each other,

but instead reveal consistent differences. In the literature these imbalances across language editions are attributed to the phenomenon of cultural contextualisation (Bao et al., 2012; Hecht and Gergle, 2009), however the terms "linguistic points of view"(Massa and Scrinzi, 2011, 2012), "national points of view" (Rogers et al., 2012)', "cultural biases" (Callahan and Herring, 2011; Laufer et al., 2014), or "culture gap" (Miquel-Ribé and Laniado, 2018) have also been used. Cultural contextualisation is based on the idea that each Wikipedia edition constitutes a community of editors united by (at least to some extent) shared cultural background. This cultural background then inevitably shapes each language edition in a unique way, both in terms of what is covered, and how. Moreover, a recent study has reported that about a quarter of each Wikipedia edition consists of the articles on culturally specific content (Miquel-Ribé and Laniado, 2018), which the authors describe as "a natural expression of cultural diversity", and the base of the imbalances between Wikipedia language editions.

An extensive body of literature has recently emerged, which studies how cultural contextualisation has shaped multilingual Wikipedia. In particular, these studies can be grouped into several streams of research:

- **Cultural contextualisation in article discourse.** Discourse-oriented studies compare the representation of Wikipedia articles on the same topic across language editions, and analyse the ways in which the cultural background of the editors is reflected in their contributions. Hecht and Gergle argues that Wikipedia editions are subjected to self-focus bias, in the sense that their editors contribute "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". As a result, the articles on the local content tend to be more developed and precise. Notably, Rosenzweig writes that Wikipedia is designed to be biased and reflect the interests of its editors. At least compared to the sources written by paid professionals, which are supposed to be neutral. Continuing the work of Hecht, Ribé and Rodríguez quantifies the amount of content isolation and editing interest to local content articles, and develops the index of autoreferentiality, a measure for quantifying self-focus bias in Wikipedia's language editions. Several comparative studies report that articles on the same topic are presented differently in Wikipedia language editions. This has been found to apply to a wide range of articles (Bao et al., 2012) from cultural heritage (Pentzold et al., 2017), to biographies of famous people (Callahan and Herring, 2011; Filatova, 2009), to national cuisines (Laufer et al., 2014). However, the differences become especially pronounced in the articles on sensitive issues like geopolitics (Apic et al., 2011; Massa and Scrinzi, 2011), history (Rosenzweig, 2006), and particularly traumatising recent events (Rogers et al., 2012). Studying online representations of geographical places, Graham and Zook have demonstrated that fundamentally different narratives can be created about places and topics in different languages. Finally, to illustrate the cross-lingual differences in article composition, several visualisation tools have been proposed, including Manypedia (Massa and Scrinzi, 2011, 2012), Omnipedia (Bao et al., 2012), and Contropedia (Borra et al., 2014).

- **Cultural contextualisation in Wikipedia coverage structure.** The studies of topical coverage seek to examine the distribution of Wikipedia articles in particular knowledge domains. These studies typically demonstrate that large language editions like English are not supersets of the smaller ones, and each edition contains unique concepts which are not covered by others. For example, several studie examine the hypothesis of the Global Consensus of World Knowledge, and argue that the assumption that encyclopedic knowledge is universal across cultures and languages is false (Hecht, 2013). Specifically, Hecht and Gergle reports that the concept overlap between the two largest editions, English and German, is only 51%. Additionally, (Warncke-Wang et al., 2012) finds out that the topics covered universally by Wikipedias include countries, cities, and lists of events, while narrower topics usually exist only in a limited number of editions. Several authors argue that the surface of the Earth itself

is represented unevenly across Wikipedia editions: Hardy et al. and Hardy find that there is generally a decreasing likelihood of edits to geotagged articles with increasing distance between editor and article. Indeed, Graham et al. examined geotagged articles in 44 language editions and found that the Global North is well represented in local language Wikipedias. The authors also found that there is not much written in Wikipedia on the Global South, and when so, it is likely to be only in English. Likewise, the similarity of content coverage between language editions decreases as distance increases (Warncke-Wang et al., 2012). An early study by Halavais and Lackaff compares the topical coverage of Wikipedia to that of books on academic subjects and concludes that Wikipedia is driven by the interests of its users, and hence lacks heavily in some areas while being elaborated in others. For example, (Kittur et al., 2009) reports "Natural and physical sciences" as the fastest growing topical area in all of Wikipedias. Additionally, a study by Bellomi and Bonato reports that English Wikipedia has a strong bias towards covering the Western culture and history. Finally, Aragon et al. analyzed the betweenness centrality of biographical articles in the largest 15 language editions, and found that the most central figures in most language edition reflect country-specific preferences.

- **Cultural contextualisation in Wikipedia editing process.** Several studies have compared the editing process and collaboration patterns across the language editions. For example, (Pfeil et al., 2006) compared the editing practices in French, German, Japanese, and Dutch Wikipedia, and reported cultural differences in the style and pace of contributions. (Nemoto and Gloor, 2011) found differences between English, German, Japanese, Korean, and Finish language Wikipedias with regards to their talk practices and conflict resolution. Moreover, Hara et al. reports differences in communication styles across the editors of English, Hebrew, Japanese, and Malay Wikipedias. additionally, Stvilia et al. studied the "Featured Article" phenomenon, and found that the Arabic, English, and Korean Wikipedians have a different understanding of information quality. Furthermore, Yasseri et al. points out differences in the circadian patterns of editorial activity across 34 Wikipedia language editions. To add, Chinese and other Wikipedias define their own editorial policies and guidelines (Liao, 2009). Finally, Zlatić et al. acknowledges the differences in growth patterns across editions, however the study also argues that when controlling for the maturity of editions, all Wikipedias exhibit a similar growth process.

These differences in number, selection, and content of articles across languages are not accidental, but relate to the cultural differences between the underlying editor communities. A brief review of the literature demonstrates that Wikipedia is rich in cultural material. On top of that, all Wikipedia data are recorded and openly available for academics. This makes the encyclopedia an attractive object for research on how culturally-mediated behaviour results in differently contextualised UGC. In particular, this thesis contributes to the study of cultural contextualisation of multulingual Wikipedia in the domains of coverage structure (Chapter 3) and article discourse (Chapter 4).

2.4 Defining and operationalising culture

Definitions. The concept of culture is vast, complex, and is surrounded by too many competing theories and schools of thought, to permit one generally accepted paradigm of looking at it. Definition of culture and its borders is a long-debated and still unresolved issue; a 1951 review of the works on the subject already contained 164 definitions of culture (Kroeber and Kluckhohn, 1952). It starts with the first mention of culture by Tylor in 1871 (Tylor, 1871, p.1); and his view still reflects well the current anthropological meaning of the notion:

> Culture, or civilization,... is that complex whole which includes knowledge, belief, art, law, morals, custom, and any other capabilities and habits acquired by a man as a member of society.

Defining culture has been a problem of fundamental importance to many fields. For example, Kroeber and Kluckhohn saw culture as a central concept in intellectual thought and science, and compared its importance to that of the notion gravity in Physics, evolution in Biology, and disease in Medicine. His 1951 review tracks the development of the concept 'culture' across national and disciplinary borders. It classifies academic attempts to define culture into several distinct groups:

- *Descriptive* definitions follow Tylor's original approach, i.e. enumerate the culture's content. In other words, culture is described as the sum total of the characteristics of a society.
- *Historical* definitions emphasise tradition and social heritage as central to culture. In this sense, the preservation and curation of historical knowledge is central to culture, both in terms of describing the past, and documenting the relational links between historical facts. This thesis will explore this view on culture in more detail in Chapter 4.
- *Normative* definitions describe culture as a rule or a way of living. Culture is viewed both as blueprint for action, and a register of sanctions for failure to follow the shared patterns.
- *Psychological* definitions see culture as a problem-solving device as it consists of learnt techniques of social adjustment. This paradigm puts emphasis on social transmission of behaviour and learning, which eventually results in socially desired habitual behaviour.
- *Structural* definitions emphasise that culture is an organised system of knowledge, norms, and paradigms. It is thus inevitably an abstraction, a conceptual model, a system of designs for living, rather than the living itself.
- Finally, *genetic* definitions highlight that culture is a product of human living, both tangible and intangible. It is the the ideas and symbols behind human creations. At last, culture is what distinguishes men from animals.

Half a century after Kroeber's review, the number of definitions of culture is still increasing (Hofstede, 2001). Such diversity can be explained by the complex nature of culture as a phenomenon, and importantly, by the fact that it is studied in parallel by multiple fields. In Anthropology, Philosophy, Psychology, Social Sciences, and Management, 'culture' is known under the names 'worldviews' (Freud, 1933; Koltko-Rivera, 2000), 'cultural orientations' (Kluckhohn, 1949), 'schemata' (Bartlett, 1932), 'value orientations' (Kluckhohn and Strodtbeck, 1961), 'unconscious systems of meaning', 'canons of choice', 'culture themes', 'configurations' (Kluckhohn, 1949; Kluckhohn and Strodtbeck, 1961), 'world hypotheses' (Pepper, 1942), 'world outlooks' (Maslow et al., 1970), 'philosophy of life' (Jung, 1951), 'construct systems' (Kottler and Hazler, 2001), 'visions of reality' (Messer, 1992), 'personal constructs' (Kelly, 1955), 'basic assumptions' (Coan, 1979), and many others. A recent review (Koltko-Rivera, 2004) concluded that although the vocabulary differs across different streams of research, in essence, they all explore the same

phenomena, and are often redundant. Following multiple definitions of culture given across academic disciplines, Koltko-Rivera identified four criteria on which virtually all of them agree: *culture a) is a complex, multi-level construct, which is b) shared among individuals belonging to a group or society; c) it is formed over a long period of time; and d) relatively stable.*

This thesis adopts this definition of culture, although admitting that it is too general to be applied directly in concrete research projects. Instead, in Chapters 3 and 4 it restricts the concept of culture even further, in order to be able to answer particular research questions. By narrowing the context, this thesis has no claims to be comprehensive or exclusive. It rather serves as a mere illustration, that the computational approaches to operationalising culture which it presents, can be adapted easily to fit various definitions of culture.

Relevant aspects from theoretical research on culture. Definition of culture for a particular research project often depends on the specific empirical question one wants to answer. Different aspects of culture are appropriate to emphasise in different kinds of inquiry. Given the complexity of the 'culture' construct, it is out of scope of this work to give a complete review of the theories built around it. Without claims of being comprehensive, this section focuses on those elements and ideas surrounding the study of culture which are especially germane when thinking about the intersection of culture and human interaction with technology, cultural manifestations in UGC, and in particular, in collective preservation of knowledge. Several theories and views of culture are central to this work, which can be summarised into three statements:

- **Cultures are systems of knowledge.** First of all, this thesis is built on the assumption that culture has some effect on the way people think, act, and interact with the environment. One of the major trends in anthropological and Social Science literature has been viewing cultures as mental infrastructures which an individual would acquire without any conscious effort through absorption of the environment (Dumont, 1979; Geertz, 2008; Goodenough, 1981; Lévi-Strauss, 1990; Schneider and Schneider, 1980). These unconscious infrastructures consist of systems of ideas, symbols, meanings, values, and beliefs, which guide people's minds and actions like vessels. At the roots of this paradigm lies the famous ideational definition of culture by Goodenough, p.167: "A society's culture consists of whatever it is one has to now or believe in order to operate in a manner acceptable to its members". Hutchins, p.374 even goes as far as to claim that "cognition is a fundamentally cultural process". For this thesis, the notion that culture dictates the ways of human life is perhaps overly deterministic and somehow outdated. Nevertheless, like the entire field of cultural studies, this work is deeply rooted in the belief that human behaviour and ways of thinking are contextualised by culture. This work thus sides with the researchers who adopt a less deterministic position on culture. In particular, (Swidler, 1986) compares culture to a toolkit, from which individuals draw various beliefs, motifs, and practices which are appropriate for the situation at hand. The items in this toolkit are vast in number. While parts of them might undoubtedly be universal, others are specific to a particular culture. Some see them as organised into logically consistent, durable, and well-formulated "cultural worldviews", or a massive "junkyard"(Martin, 2010) or "clump" (McLean, 2016) of cultural elements scattered in one's mind. Martin sees them as a "network of concepts and ideas" which are connected by the ties of mental models, gestalts, and associations. Regardless of the organisation, the view that these mental infrastructures exist is important for this thesis. Precisely, the mere presence of these infrastructures implies that they have a potential to create physical, measurable experiences when particular people take a particular action. For example, writing a book, sharing a memory, designing an artwork, or, in the context of this thesis, generating content online can all be viewed as physical manifestations of one's cultural mental infrastructure. This leads us to the next point.

- **Culture is what culture does.** Considering the number of attempts to define and classify culture, it might seem that culture is viewed by academia as a thing. Nevertheless, according

to Street, "culture is a verb, it does rather than is". Many researches share this perspective, suggesting to focus on the measurable outcomes of culture, such as experiences, actions, expressions, and other forms of human interaction with the world. Theoretical perspectives on culture as a verb leave researcher much interpretative freedom. To illustrate, Sapir, p.233, for example, suggests that "culture may be defined as what a society does and thinks". In this context, simple mundane acts like choosing a newspaper to read, watching a sports game, or editing a Wikipedia page on own city's history could all be interpreted as cultural acts which reflect the person's worldview, to some extent. In fact, Koltko-Rivera suggests that culture is expressed through the preference in one's behaviour and language. Bryson and Jones, p.74 emphasises this point further, by suggesting that "...culture is human energy organised in patterns of repetitive behaviour". These positions are crucial to this thesis, because they *warrant it possible to operationalise and quantify culture as an aggregarion of a multitude of small mundane human behaviours which are in essence cultural expressions indicative of larger underlying cultural mental models*. In the context of UGC it means that by collecting and analysing the traces and products of human *activity* on the Web, it is possible to extract meaningful cultural patterns. This thesis demonstrates that this is true in practice.

- **Real world is culturally contextualised.** As a consequence of the previous two postulates, it follows that most of human activity should be at least to some extent contextualised by culture. In his well-known theory of language as joint action (Clark, 1996), Herbert Clark describes culture as a "communal common ground" which certain sets of people share and other people lack. This common ground, or "shared expertise", consists of a set of "facts, beliefs, procedures, norms, and assumptions". According to Clark's theory, cultural communities could be thus identified through the shared common ground, or expertise. On the other hand, the absence of this common ground results in the invisible, but uniquely real boundaries between cultural communities. It is through detecting the presence and the shape of these boundaries that it becomes possible to measure culture and its manifistations in the real world. Thorton eloquently describes this as following:

  > One thing that culture does is create boundaries of class, ethnicity (identification with a larger historical group), race, gender, neighbourhood, generation, and territory within which we all live. Boundaries are created and maintained when people [...] internalise modes of thought to the extent that they become entirely automatic and unconscious. These boundaries come to seem uniquely real and permanent. Their creation is only obvious when we step outside our normal day-to-day interactions. (Thornton, 1987, p.27)

Both Clark and Thorton speak of two sides of the same phenomenon of cultural contextualisation. On the one hand, *cultural contextualisation implies similarities* in the behaviours and opinions which are shared by the members of a community. Applied in the UGC context this suggests, that online content such as Tweets, Wikipedia articles, or Foursquare reviews would inadvertently reflect the "beliefs, procedures, norms, and assumptions" which are shared by the members of the community who produced this content. On the other hand, each of such communities would produce the content which reflects its own unique cultural point of view, and thus, is different from everyone else. In other words, online content generated by users who are members of various cultural communities, *is expected to reflect a great diversity of culturally slanted opinions and biases*. Finally, Clark's and Thorton's theories imply that it is virtually impossible to come across an example of UGC which would not be culturally contextualised to some extent. This thesis looks at the examples of such similarities and differences in the content created by various cultural communities of Wikipedia editors. It demonstrates that by examining and comparing myriads of small actions like editing

Wikipedia articles, one can arrive at comprehensive conclusions about the real world shape of cultural borders and clusters of "communal common ground".

Operationalising culture and cultural borders. Defining the borders of culture has proven to be a challenging task. A given community might be characterised by one dominant cultural tradition, as well as multiple, subdominant subcultures (Keesing, 1990). To account for this diversity, models of culture are distinguished by their level (e.g. national, organisational, individual), and the nature of the group on which they focus. For example, cultural groups might be defined by ethnicity, gender, age, religion, social status, or otherwise. In order to simplify the complexity of the culture construct, scholars often operationalise cultures through simple criteria. One of the most popular ways to define culture has been through the country of origin or the current citizenship of the studied respondents. In fact, according to (Schaffer and Riordan, 2003), approximately 79% of cross-cultural studies published between 1995 and 2001 used nationality or citizenship as a proxy for culture. Other approaches argue that drawing cultural borders should not be done on the basis of national borders alone, especially when comparing multiethnical countries. A review by (Peterson and Smith, 1997) provides a comprehensive list of other possible determinants which might help researchers operationalise cultural borders in their studies. They include language, proximity and topography, religion, economic development, technological development, political boundaries, industry type, and climate. Other suggestions include ethnic origin (Allik and McCrae, 2004; Okazaki and Sue, 1995) and similar historical background (Koltko-Rivera, 2004).

The following theoretical perspectives are particularly relevant to this thesis:

- **Culture and language.** The relationship between language and culture is probably one of the most discussed in the literature (Bloomfield, 1945; Hoijer, 1948; Silvia-Fuenzalida, 1949; Voegelin and Harris, 1945). While it is universally agreed that culture cannot be reduced to the language alone, researchers in multiple fields identify language as one of the main components of culture, and moreover, one of the factors that distinguishes cultural communities from each other. From the cognitive perspective, language is a useful indicator of culture, since it influences one's values. This view is based on the belief that knowledge transfer and the human cognition itself are to a large extent carried out through language:

  > The notion of culture is inseparably linked to language on the grounds that culture is thought and transmitted as a text through language. [...] ...acquired knowledge is being continuously stored in a manner that makes it relatively accessible when necessary (Bloch, 1991, p.184).

  Early theoretisations on the relationship between culture, cognition, and language became known among the popular culture as the famous Sapir-Whorf hypothesis(Sapir, 1921; Whorf, 1940). It exists in two forms: its stronger form, known as linguistic determinism, suggests that language determines thinking; the weaker formulation postulates that language only influences thinking. Modern scientific view sides with this weaker, linguistic relativism interpretation (Crystal, 2003; Pinker, 1994). Agar, p.71 writes on the topic that "Language carries with it [...] patterns that mark the easier trails for thought and perception and action." Hofstede, p.21 adds: "Our thinking is affected by the categories and words available in our language." Accumulating evidence suggests that this extends not only to one's native language: new languages learned also influence thinking, non-linguistic mental concepts, and result in acquiring hidden cultural knowledge (Dubin, 1989; Nisbett, 2004). Quite similarly, in ethnolinguistics and anthropology, language is an important bearer of culture (Silvia-Fuenzalida, 1949; Voegelin and Harris, 1945) – its meanings have to be learnt socially and represent the way of life as seen by a particular community. Finally, from the socio-historical and psychological perspectives, language is also central to culture for several reasons. It reflects the collective agreement of a language community to view the world in a certain way, and helps

a community to perpetuate its culture, develop its identity, and archive accumulated knowledge (Kramsch, 1998). Language-speaking communities form distinct and unique cultures around themselves (Bucholtz and Hall, 2008; Geertz, 1973). Moreover, cultural communities bond around a shared language. It has been argued that designation of a national language facilitates the development of national group identity and is a pre-requisite to the formation of a stable nation (Fasold, 1984).

- **Culture and nation.** In the literature cultures have often been treated as if they reside within the national borders. Equating cultures with countries has become a common academic practice, to the point that culture and country are used in publications as synonyms (Taras et al., 2016). Theoretically, these operationalisations hold on the ideas that: a) cultures tend to be somehow geographically localised, although these borders can be pliable; and b) cultures have a strong political and ideological edge and the power to unite people into imaginary comradeship of fellow members, often called a nation.

  As discussed above, values are at the root of culture; and geography and geopolitics have profoundly affected the distribution of culture-related values and profiles. For a long time geographic distance and national borders have, to various degrees, limited interpersonal exchanges, flows of information, labor, and products. As a result, cultural values commonly cluster within country borders. For example, (Allik and McCrae, 2004) looked at the personality traits across 36 countries and found that geographically proximate countries often have similar profiles. According to a large body of research, these values typically include the preference towards political and social organisation. In articular, "the link between nation and culture tends to occur because people prefer to interact with other people and be guided and politically governed by institutions consistent with their values"(Peterson and Smith, 1997, p.934). Or, in the words of Gupta and Ferguson: "places [...] have a logic of their own", which shows through political and economic determinations of nation states. Apart from politics and economy, Duncan and Jackson list the discourses of ideology, social organisation, and subordination as important components of national culture. Another relevant theorising of culture comes from the field of cultural geography, where national cultures are often discussed in the context of political contest and domination. Paul, p.49-50 describes nation as " a unified cultural community" which "constructs and defends the image of national-culture, homogeneous in its whiteness yet precarious and perpetually vulnerable to attack from enemies within and without." Thus, culture as a system of political power shows itself through colonisation, ethnic wars, and conflict once societies clash with each other (Gregory and Ley, 1988; Latour, 1987; Mitchell, 1995). As a consequence, cultures split the world into discrete national communities which are in the state of perpetual mutual resistance and struggle for domination (Baker and Biger, 2006; Gregory and Ley, 1988). Or, in the words of Mitchell, p.108, "culture differentiates the earth" into "us" and "them". In the context of UGC and Wikipedia in particular, the contest between the national cultures, for example, may show themselves in the way they describe their own and others' national histories, as Chapter 4 of this thesis shows. Finally, deliberating on the nature of cultural borders, Jackson, p.4 concludes that they "change shape according to changing historical and geographical circumstances". Some authors, in fact, reject the notion of tangible borders all together. For example, Hall describes nations as mental models, systems of representation, and cultural identities. Hall's national cultures can go beyond political entities as they unite ethnic nationalities into one people of shared cultural identity and a sense of belonging. Furthermore, the theory of imagined communities introduced by Anderson talks about nations as invented political communities which are imagined in a sense that most of the members of even a small nation will never meet their fellow-members. At the same time, such communities are united by a strong sense of comradeship and elastic but finite boundaries beyond which lie other nations. To that end, the concept of locality itself

becomes blurred as people travel and leave homelands, are displaced, immigrate, etc. As a result, (Gupta and Ferguson, 2007) talks about "the partial erosion of spatially bounded social worlds and the growing role of the imagination of places from a distance". As such, nations can be described as imagined communities attached to imagined places.

While it is possible to use multiple approaches to operationalising culture, this thesis focuses on two which are most popular, well-defined, and applicable in the context of UGC. In particular, in Chapter 3 culture is approximated by the language. Chapter 4 adopts the geopolitical definition and uses national borders as delimiters between cultures. Both of these operationalisations are admittedly severe simplifications of the concept of culture, and are not intended to be interpreted as comprehensive solutions to how cultural borders should be delineated. However, the relationships between culture, language, and nation are some of the most described in the theoretical literature on the subject. Additionally, many cross-cultural studies have already been run using country of origin or language as a cultural delimiter. By using similar criteria, this thesis ensures the possibility to cross-validate its findings with the results proposed by a large body of preceding cultural research. Finally, country and language of a contribution, be it a Wikipedia edit, a tweet, or an anonymous review, are among the most easily extracted features in UGC, and selecting them for the current analysis has a large practical value. To conclude, the operationalisations of cultural borders proposed in this thesis serve merely as a starting point in the investigation of cultural contextualisation in UGC. As such, they welcome subsequent research to investigate other delimiters of cultural communities in online spaces.

2.5 Approaches to quantifying culture

Lessons from across fields. First examples of quantitative studies of culture come from the fields of Management, Business, Organisational studies, and Psychology. They represent a shift from the traditional anthropological studies which offered descriptive documentations of the observable external layers of culture. Instead, they aim to formalise and compare cultures through a number of faucets, such as values and attitudes that guide human behaviour. These cultural faucets have been typically measured through self-reported survey responses, which were then aggregated, converted into scores along several dimensions, and finally used to map cultures as points in this multi-dimensional space. The selection of the cultural faucets usually depended on the disciplinary background of the researcher and availability of survey respondents. One of the cornerstone quantitative works which inspired a lot of subsequent research in the area of culture studies was the 1984 paper by Hofstede (Hofstede, 1984). He compared the surveys on work-related values of the HERMES Corporation employees in 40 countries. The study found substantial cultural differences in they employees' "collective programming of the mind". Hofstede summarised these cultural differences into four dimensions: power distance; uncertainty avoidance, individualism, and masculinity. By giving each country a score in each of the four categories, Hofstede finally mapped the world into 8 distinct clusters of stereotypical human behaviour. Coming from a background in Management science, Hofstede focused on work-related values, however other cross-cultural scholars emphasised other types of values and attitudes depending on their disciplines. For example, Social Sciences focused on people's attitudes to social and political issues, such as economics, religion, sexual behaviour, gender roles, family values, and ecological concerns (Inglehart et al., 1998). Psychologists, on the other hand, were interested in comparing self-perception (Singelis, 1994) and basic social axioms which guide human behaviour (Bond et al., 2004). Following the groundbreaking work of Hofstede, other researchers have proposed their own versions of cultural dimensions (see House et al. (2004); Inglehart and Baker (2000); Schwartz (1994); Smith et al. (1996) to mention just a few prominent works). A recent cross-field review (Taras et al., 2009) has counted 121 distinct constructs, or facets, that have been proposed to measure cultural dimensions. A close inspection, however, confirmed that both conceptually and empirically most of them closely correspond to the original dimensions proposed by Hofstede. Moreover, his original four dimensions are often used for validation purposes in multicultural projects.

While this type of works has laid a sound foundation for a number of theories in culture research, they all have important limitations when it comes to the underlying data. Virtually all of these studies are based on non-representative, self-report questionnaires; the vast majority uses convenience samples which include very narrow target groups, typically, students (Taras et al., 2009). Additionally, Taras et al. reports issues with representation and coverage: some of the largest studies include between 40 and 60 cultural groups, while the rest cover only 2 to 10, typically skewed towards the largest and most conveniently available datasets. Other criticism (Taras and Steel, 2005) includes a certain degree of negligence when it comes to discussing statistical matters such as the variance of cultural scores between the groups. Conceptually, the progress of this line of research on culture has been made by adding and exploring new cultural dimensions, improving the properties of the questionnaires, and sample qualities. This research has largely ignored the emergence of the new types of culturally rich data which became available with the advent of UGC, as well as the methodological tools for measuring and comparing cultures beyond the traditional questionnaires.

Quantifying culture in computational fields. New wave of research on culture comes from such computational fields such as Physics, Computer and Network Sciences. They explore new types of data, shifting from traditional self-reported questionnaires towards observational digital data such as activity logs, digitised datasets, and user-generated content. Together with the new types of data, new methodological approaches emerge, with the focus on large-scale analysis. This

section gives detailed examples of some of these methodologies and presents an overview of the diversity of explored data sources.

Curiously, most of the recent large-scale computational studies of culture utilise networks to quantify, visualise, and compare cultural communities. For instance, Herring et al. examines culture-related differences across the blogosphere. The study compares inter-user connectivity and linguistic preferences across a random selection of Russian, Portuguese, Finnish, and Japanese LiveJournal.com user profiles, and discusses the differences by visually interpolating user-to-user networks. Likewise, several other studies have chosen to approximate cultural communities through language. One of them (Ronen et al., 2014) looks at how cultures gain global visibility through linguistic dominance. The authors infer the networks of international communication between the educated elites from the patterns of multilingual tweeting and Wikipedia editing, and the data on book translations collected by the UNESCO's Index Translatorium project. For each data type, the authors build a global language network. The links between the language nodes represent significant connections expressed with t-score values. The position of a language in the network, measured as the Eigenvector Centrality, determines the influence of a language, and thus the culture that its speakers carry, on the global arena.

Several studies operationalise culture through national borders. For example, State et al. map the global patterns of cross-country communication in Twitter network. The study measures the density of contact between users in different countries as a fraction of observed and expected relationships of mutual following, given the number of Twitter users in each country. This statistic is used as a link weight in the network of international cross-country communication ties. Another study (Barnett and Benefield, 2015) examines the thesis that cultural homophily between countries predicts international Facebook communication ties. The authors visualise a Facebook communication network based on the linking data taken from (Newman, 2012) and report its statistics. Additionally, the study examines several hypotheses that might explain the Facebook network. To assess their significance, the authors apply correlation/logistic regression (likelihood ratio χ^2 test and Nagelkerke's R^2 test which compares the significance of the models as a function of deviance, as opposed to the variance in linear models). Another example comes from García-Gavilanes et al. who examines how culture shapes international cultural boundaries by studying the Twitter mention and retweet network across 100 countries. Locations data are taken from the user profile info. The study uses the gravity model to explain the network ties, and does so by applying multiple linear regression. Another project (Platt et al., 2015) examines the international cultural impact of online YouTube videos in 57 countries, testing the hypothesis of culture exporting, trend-setting nations. The study represents the data as a video-nation matrix of the number of times a video trends in a nation (similar to document-term matrix in natural language processing). Nation-to-nation similarity is computed probabilistically as a symmentric conditional co-affiliation equal to the probability that video trends in both countries. Contextual factors that predict the co-affiliation ties are tested with ordinary least squares linear regression.

Quite a different approach is introduced by Schich et al. who examine the processes driving cultural history through the birth and death locations of more than 150,000 notable individuals. The study analyses the locations which played significant role in the life of the notable elites, and constructs a worldwide historical migration network connecting their birth-to-death locations. Assuming every death in a location to count as a vote for its attractiveness, the authors determine the most attractive migration sports by their PageRank and the Eigenvector Centrality in the network. Cultural narratives are examined on a case-by-case basis. This study is distinctly different from others because it combines quantitative methods of massive data retrieval, descriptive statistics, a variety of visualisation techniques (cartograms, timelines, movies, demographic tables, networks, etc.), and a qualitative interpretation by the field experts. The data comes from Freebase.com, the General Artist Lexicon, and the Getty Union List of Artist Names.

A practical perspective on culture is offered in another study (Mocanu et al., 2013) which argues that language use in online media is indicative of some culture-related dynamics in the real

world. In particular, the authors use online Twitter activity to develop linguistic indicators for mapping offline social communities, as well as the patterns of language polarisation/homogeneity, and seasonal tourist flows. They examine probability density functions of user activity for specific countries and languages, and map language provenance of Twitter users, normalised by their location and activity level in that language. Map granularity is on the country, region, and city levels.

Close to computational linguistics, another study (Michel et al., 2011) focuses on how culture shapes language, grammar, and lexicography. The authors analyse the corpus of 5,195,769 digitized books from Google's collection. They track the growth of lexicon, the development of culture-related concepts such as "slavery", and the change around the language used to talk about such concepts. The analyses include the development of n-gram frequency over time, grammatical complexity change, and the timelines of attention to and forgetting of famous individuals.

Finally, coming from the domain of semantic tagging, Thakker et al. examine cultural variations in online interpersonal communication. The study builds an ontology of culture-specific concepts which enables automatic tagging of culture-related mentions in textual content. The core ontology is built based on the theoretical models of culture and extended with DBpedia linked data.

Quantifying culture in Wikipedia. When it comes to Wikipedia, a large proportion of culture-related research has chosen the encyclopedia as the source of data. This section focuses on the details of methodologies which have been used in the literature. In particular, the studies can be divided into several groups, according to their methodological framework.

- **Network science.** Many studies use networks in their analysis, and several of them focus on biographies. For example, one (Aragon et al., 2012) compares Wikipedia biographical networks across 15 language editions. The authors build language-specific article linkage networks, where link weights represent the number of outlinks from the text of one article to the other. The study compares descriptive statistics for each of the networks, as well as builds a network of cross-language similarities.There, the edges are computed as Jaccard coefficient, i.e. the ratio between the number of links present in both language networks (their intersection) and the number of links in their union. A similar approach has been used by another team (Gloor et al., 2015) who studied cultural chauvinism in the English, Chinese, German, and Japanese Wikipedias, and compared biographical networks of influential historical leaders who lived at the same time. The most influential people are ranked by a combination of PageRank and in-degree. The resulting lists are compared across Wikipedias by reporting counts and percentages. Another study of culture through biographies (Eom et al., 2015) applies a number of ranking algorithms to the hyperlink networks of 24 Wikipedias. The top 100 most important historical figures in each Wikipedia are filtered from the total ranked list by removing all non-biography articles. The study further introduces a network of 24 language-approximated cultures, where edge weights are proportionate to the number of foreign figures quoted in top 100 of a given culture (each person is characterised by the main language of the country where they were born). Other than biographies, another study (Gloor et al., 2015) has compared the distribution of the most central topics in German, Portuguese, English, and Spanish Wikinews. In this work, the topics are represented by Wikipedia concepts referenced in the Wikinews posts, ranked by their betweenness centrality in the concept-concept network. Finally, another study (Hale, 2014b) has compared the top 46 Wikipedia editions with regards to the proportion of multilingually active editors. Additionally, Hale analyses the directed weighted network of inter-edition relationships where the edges are computed as the log of the number of editors who primarily edit one language edition but also edit the other one.

- **Statistical models.** Several statistical frameworks have been suggested in quantifying culture-related phenomena in Wikipedia. For example, a study by Laufer et al. examines cross-cultural interest by analysing the descriptions of national cuisines in 27 Wikipedias. The authors propose to use Jaccard similarity as a measure of cultural understanding. It is computed as the overlap between the set of concepts in the Wikipedia article on a national cuisine in its "native" language, and its sister-article in another language. Another study (Yasseri et al., 2012a) compares editing patterns among 34 Wikipedias, deriving an average circadian daily and weekly activity patterns for each of the editing communities. The curves are calculated based on the timestamps assigned to edits, for time of the day and day of the week. A series of studies have examined cultural variations among Wikipedias through conflict patterns among the editors. As such, one (Yasseri et al., 2012b) has proposed a controversy measure to asses the severity of edit wars over Wikipedia articles. The measure is language independent and is determined by the the number of edits in the pairs of editors reverting each other's contributions, and the total number of editors involved in the article. A subsequent analysis (Yasseri et al., 2014) has successfully applied this measure to compare the most controversial topics among 10 different versions of Wikipedia. Another example comes from Kim et al.. The authors examine cultural variations between Wikipedia editors who edit multiple editions versus those who mainly edit just one. The study compares the groups in terms of descriptive editor engagement statistics and the editors' language proficiency (entropy of n-grams, entropy of parts-of-speech frequency, and the difference in the article usage by primarily and non-primarily editors). Additionally, it applies Bayesian topic modeling to map Wikipedia articles to 100 distinct topics, and computes the proportion of editor interest in each topic. The between-group differences are tested using a standard two-tailed independent samples t-test. Finally, a quite recent study (Miquel-Ribé and Laniado, 2018) has proposed an algorithmic way to quantify the amount of cultural context content across 40 Wikipedias. Cultural content of a language edition is defined as the articles which are geotagged with coordinates of a territory where the language is spoken, or contain keywords related to the language or the territory. Further articles are added to this set if they are classified into a Wikipedia category whose title contains one of these keywords, and their text contains a large enough proportion of outlinks which point to the original set of articles. The differences between the editions are summarised as descriptive statistics, plots of proportion of cultural content over time, and the heatmap of cultural content shared between each of the 40 editions.

- **Mixed methods.** Several studies have mixed quantitative and qualitative methods in order to arrive at more comprehensive accounts and achieve higher validity. For example, Hara et al. compares several types of Talk pages across the English, Hebrew, Japanese, and Malay Wikipedias. The study applies interpretative content analysis, the authors manually code over 2,700 Wikipedia posts to develop a coding scheme of topic categories. Subsequently, the frequency tables of each edition are aggregated as related to Eastern or Western culture, and cross-tabulation analysis is run for statistical comparison. Another study (Pfeil et al., 2006) has analysed all changes to the article "Game" in French, German, Japanese, and Dutch Wikipedias, manually classifying each type of change. The classification is developed using ground theory. Approximating cultures with the countries where the studied languages are spoken, the authors report Pearson correlation between the relative percentage of changes under each category and the score of the countries along several Hofstede's dimensions of culture (Hofstede, 1980).

- **Web tools.** Several interactive tools have been proposed to the research community in order to explore the differences between Wikipedias. As such, Massa and Scrinzi has proposed a Web tool called Manypedia. Manypedia allows to compare the current version of a Wikipedia article across language editions, with automatic translation of the article into up

to 56 languages. Apart from the text, it allows to compare the images, the views statistics, the number of received edits, and some other descriptive article statistics. Another system, Omnipedia (Bao et al., 2012) combines entries from different language versions, exposing the users to a range of different cultural perspectives. It visualises the topics discussed in a given Wikipedia article, side-by-side with the topics in its sister-articles in other editions. The algorithm is based on a number of natural language processing techniques, which make possible, for example, cross-lingual concept alignment in the presence of missing inter-language links and ambiguities caused by conceptual drift.

2.6 Conclusion

Several general trends are evident from this overview. First of all, quantifying culture and its effects, despite being old, is still a new domain. The recent computational turn in quantifying culture is distinctly different from the original studies in Psychology and Management. First of all, there is no established approach to quantifying culture-related trends. The field is yet to see cornerstone studies like those by Hofstede which would bring structure and set the tone to measuring cultural phenomena. More commonly, the current literature focuses on making evident the presence of culture-related variations in UGC, and demonstrating that these variations are statistically significant and empirically substantial. Collecting empirical evidence and testing out the tools which are helpful for extracting cultural patterns seems to be the necessary stage in this young emerging field of studies. However, what seems to lack is the intent of establishing sound computational practices and frameworks for quantifying cultural variations in the data. More than that, when it comes to UGC, there seems to be no widely adopted definition of what culture is or how to operationalise it. In fact, despite implicitly studying cultural differences, several studies avoid using the term all together, potentially, in order not to face conceptual debates. Finally, quantitative literature on culture is still rather small and sporadic, mostly dictated by the availability of the data. This often makes it difficult to apply the same approaches in other contexts or adapt the proposed measures to new types of data.

At last, when it comes to practical lessons from the literature, there are also several best practices to be learned. For example, some of the most comprehensive from the examined studies tend to mix and match multiple methods and data sources. This allows them to achieve a holistic perspective on the studied phenomena, and to be able to tell a story substantiated with the analysis. Moreover, several studies cross-validate their findings by using real word population statistics to statistically explain the variations found online. Finally, recent studies of culture are done by increasingly interdisciplinary teams, and it is evident that combining expertise from various fields and involving domain experts is beneficial for the analysis.

Chapter 3

Mapping communities of shared information interest

Cultural globalisation has become a popular concept, idealising free and borderless exchange of ideas, values, and shared knowledge across the world. It is true that certain intellectual notions, fashions, opinions, and commercial brands have diffused and gained cultural meaning around the globe. However, it is difficult to asses, whether these anecdotal examples indicate the presence of a general trend towards homogenisation of information interests across cultural communities.

Wikipedia data can be very useful in order to get an impression on how culturally proximate, or diverse communities are with regards to their information interests. There are several reasons to this.

First of all, Wikipedia is an open and all-inclusive system by design. It unites the collaborative effort of editors with various linguistic, geographical, and professional backgrounds by allowing anyone to contribute. Secondly, all contributions to Wikipedia are voluntary. This justifies that the editors are driven purely by their enthusiasm. Thus, the topics they select to work on, reflect the true information interests of the editors. Thirdly, all activity on Wikipedia is recorded and attributed to specific users. This means it could be aggregated by communal features. For example, aggregating editing activity across locations of the editors could give an impression of an interest profile of a specific region, city, of an entire country. Similarly, aggregating the edits of an entire language edition can be indicative of large-scale information preferences in this linguistic group.

Evidently, Wikipedia editors are not representative of the general population (and rather, are known to be mostly male, white, and educated). Nevertheless, analysing their aggregated editing activity and knowledge curating process is a plausible first step towards a quantitative, rigorous analysis of whether cultural globalisation of interests is happening.

In order to quantify information interest of editor communities (defined by, for example, language, location, age, gender, or other characteristics), I propose to use the amount of edits as a proxy. This way, one can gain an impression of the attention given to each concept, by counting the number of edits to the corresponding Wikipedia article. By inference, a larger number of edits will indicate larger relative importance of this concept in the community.

Using this formalisation across various communities, it then becomes possible to assess dyadic relationships between communities. For example, it can be used to measure shared information interest and shared attention to concepts. A number of questions become relevant in this context.

- Are all communities uniform with regards to their collective interest? (In other words, have their shared interests converged to a specific set of universally known concepts?)
- Do certain communities group based on their shared interest in specific concepts? (this would also mean that other communities are less interested in, or unaware of some concepts)
- How to map the general landscape of shared information interests across cultural communities, and what does it look like?

In this chapter, I examine these questions quantitatively. First of all, I present a suitable scalable approach to operationalising and extracting significant shared information interests. Secondly, I validate it on large empirical datasets of multilingual editing activity on Wikipedia.

At the core of the approach, there is a statistical filtering method for extracting meaningful dyadic relationships from multinomial data. I combine it with other statistical inference methods. This approach allows to formulate, quantify, and order by plausibility, hypotheses about the nature of the extracted relationships.

To test and validate this approach empirically, I focus on Wikipedia. It is a powerful example of culturally contextualised UGC, specifically due to its wide linguistic and geographical coverage, and international popularity. Additionally, it offers a formalised design shared across its linguistic versions.

Still, the approach can be useful in the context of any multidimentional collaboative system. It is applicable whenever it is meaningful to group contributors by certain shared features in their profile - be it location, field of interest, language, socio-demographic characteristics, etc. Since the approach is based on measuring the activity levels (to approximate interest), it overcomes linguistic barriers. Thus, it is very useful in multilingual environments, as long as a contribution (e.g. edit) have the same meaning and cost in each of the studied communities.

Overall, this approach helps to formalise the process of quantifying large-scale multi-community relationships with regards to shared interests. It can also be extended to an arbitrary possible number of communities and length of the timeframe, both in reasonable computational time.

The remainder of this chapter is structured the following way. Section 3.1 outlines the approach details. It starts with empirical background that is relevant to quantifying inter-cultural borders and points of shared interest. It then introduces the statistical filtering method for extracting significant bilateral information interests. The rest of the chapter tests this approach by applying it to longitudinal Wikipedia editing data. To start, Section 3.2 focuses on quantifying bilateral information interests across geopolitical communities of Wikipedia editors. In this setup, editors are grouped based on their physical location at the moment of contribution. To continue, Section 3.3 explores the linguistic definition of a cultural community. Here, I quantify shared information interests across 110 Wikipedia language editions.

3.1 Approach: Extracting and understanding ties of shared interests

This section presents a short technical summary of the statistical filtering that colleagues and I developed. The filtering is applied to extract a network of shared interests from Wikipedia co-editing data. Later, I demonstrate how this network could be clustered into communities. Finally, I illustrate how these can be used in order to motivate and test empirical hypotheses about the factors that explain some of the network's structure.

Editor communities. To study information interest similarities, I focus on interest profiles of editor communities. *Editor communities* are groups of independent Wikipedia editors united by a common feature. This could be geographical location, gender, age, professional interest, or language in which they contribute, to name a few possible grouping criteria. I approximate an interest profile of an editor community with the selection of Wikipedia concepts which are edited by the editors who belong to this group.

In this thesis, I empirically explore two possible definitions of editor communities — grouped by (a) location of the editor, and (b) the language edition to which they contribute. The approach would also work with other grouping criteria. In the context of Wikipedia, for example it is also possible to examine editor communities based on belonging to a city or geographical region, or on other characteristics such as age or gender, when it is possible to retrieve them.

Shared interest ties. Intuitively, shared interests between editor communities can be conceptualised as a network, where each node represents an editor community. Two nodes are connected by an edge if these communities share significant interest in the same concepts, such that the stronger the shared interest, the higher the weight of the link. A naive approach to drawing the edges between these nodes is to use the raw co-edit counts. While the idea is simple and intuitive, it has a serious limitation.

Since there are millions of articles on Wikipedia, most editor community pairs are likely to have co-edited at least one article. In fact, when the co-editing activity is aggregated across all articles, it will result in a hairball network, where all nodes are connected with each other. The problem is, it is difficult to know, which of these co-editing ties represent true shared interests, and which exist due to chance,or merely due to a large and active community of Wikipedia editors.

To take a particular example, if we delimit Wikipedia communities by language, the resulting network will show strong interconnections between English, German, and Swedish language communities because these are the largest and the most active editions of the encyclopedia. Moreover, smaller language editions have a much tighter pool of editors and subsequently, much lower aggregated activity levels. Thus, their edges will have negligibly small weights compared to the most active core. This is why, a more sophisticated approach is needed that accounts for these effects.

3.1.1 Step I: Statistical formalisation of the shared interest model

The approach that I present in this section filters out the connections in a co-editing network that exist only due to community size, activity level, or noise. The idea is to compare the empirical co-edit counts with the co-edit levels predicted by a Null model, preserving only those links that are statistically significant.

The Null model assumes that edits are randomly assigned to communities, proportional to the cumulative editing activity of these communities. This random assignment can be drawn from a multinomial distribution of community activity levels, which reflects the proportional edit count of each community in the entire dataset. Mathematically, each edit of community i occurs with

probability $p_i = \frac{1}{M}\sum_c k_i^c$, where M reflects the total number of edits from all communities and k_i^c is the total number of edits by the community i on the concept[1] c.

This Null model expresses the expectation that each edit happens independently from all other edits, while preserving the community activity levels observed in the empirical data. Thus, a co-edit between every pair of communities is possible. Its probability depends only on the proportional activity of these communities, and is independent from any other factors. Using the Null expectation of edit activity defined above, for each pair of editor communities p_i and p_j, the probability to co-edit a concept by chance is expressed as:

$$\mathrm{E}[w_{ij}^c] = n_c(n_c - 1)p_i p_j, \tag{3.1}$$

where n_c is the total number of edits to the concept c. To determine which edges cannot be explained by chance, I compare the observed co-edit levels w_{ij}^c on the concept c with the expected edit count given the Null model $E[w_{ij}^c]$:

$$z_{ij}^c = \frac{w_{ij}^c - \mathrm{E}[w_{ij}^c]}{\sigma_{ij}^c}, \tag{3.2}$$

where the standard deviation σ_{ij}^c, according to the multinomial theorem, is defined as

$$\sigma_{ij}^c = \sqrt{n_c(n_c-1)p_i p_j((6-4n_c)p_i p_j + (n_c-2)(p_i+p_j)+1)}, \tag{3.3}$$

The resulting value in the Eq.3.2 reflects the difference between empirical and expected edit counts expressed in standard units called z-scores. This also reflect the number of standard deviations that separate the expected value from the observed one. Note that this is a z-score for just one concept c and communities i and j. In order to find the z-score for a pair of communities i, j over the entire set of concepts, I sum their z-scores over all c:

$$z_{ij} = \sum_c z_{ij}^c. \tag{3.4}$$

Using the Bonferroni correction (Dunn, 1961), a link is considered to be significant if the probability of observing the total z_{ij}-score is less than α/N, where N is the number of communities, and α is the selected significance level (for example, $\alpha = 0.05$). Since the total z_{ij}-score is a sum over many independent variables, one can approximate the expected total z_{ij}-score distribution with a normal distribution. The normal distribution has average value 0 and standard deviation $\sqrt{L}$, where L is the number of Wikipedia articles. Thus, the threshold for the significant link weight is $t = a\sqrt{L}$, where a is derived from the condition that $P(z_{ij} > a) = \alpha/N$, where N is the number of communities and P is the standard Gaussian distribution (with zero average and unit variance). If the total z_{ij}-score is larger than the threshold, we create a weighted edge between the nodes i and j with weight $\widetilde{w_{ij}}$ according to

$$\widetilde{w_{ij}} = \begin{cases} z_{ij} - t & \text{if } z_{ij} > t \\ 0 & \text{if } z_{ij} \leq t. \end{cases} \tag{3.5}$$

The resulting weighted edges are used to build a network of shared topical interests. In it, the nodes represent the communities, and the edges are statistically significant and weighted by the strength of shared interests, quantified via z_{ij}-scores. This model can also be described as inferring

[1] I use the term *concept* to refer to the subject of a Wikipedia article, and *article* to refer to each particular instance of writing on a concept in a certain language edition. Thus, a concept may be represented in multilingual language editions by several unique article instances, each connected together by inter-language links.

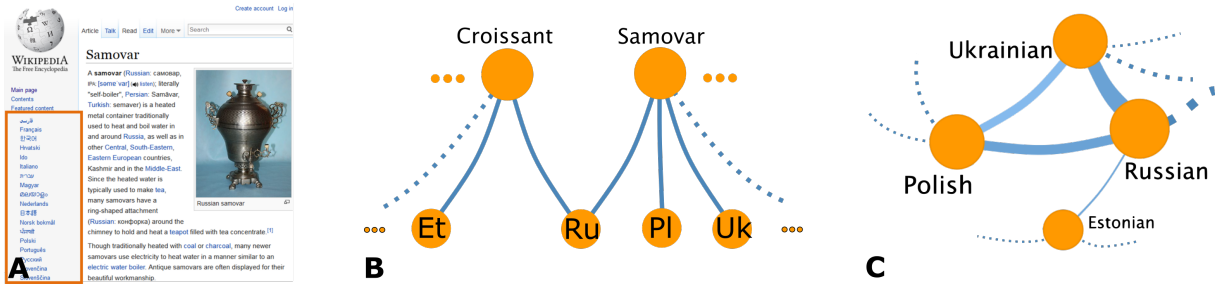


Figure 3.1: Illustration of the z-score-based filtering method. The method requires three steps: (a) to retrieve all edits to each concept in all linked language editions; (b) to compare the empirical and expected probabilities of each language pair to co-edit a concept; and (c) to create a filtered network of languages with significant shared interests. In the final network, 'heavier' links signify stronger co-editing similarity between the nodes.

significant links in a bipartite system in which communities are in one set of nodes, and concepts are in the other set. I illustrate the logic of the shared interest model in Fig. 3.1, where for the purpose of the example, I define an editor community as the set of all editors who contribute to the same language edition of Wikipedia.

Other methods exist to evaluate the significant correlation between entities in bipartite systems. For example, Zweig and Kaufmann (2011) proposed a systematic approach to one-mode projections of bipartite graphs for different motifs. In another work, Tumminello et al. (2011) used the hypergeometric distribution and measured the p-value for each subset of the bipartite network. Moreover, Lancichinetti et al. (2015) proposed a community detection method to classify article topics more efficiently. To add, Serrano et al. (2009) used a disparity filtering method to infer significant weights in networks. Finally, Ronen et al. (2014) adopted a statistical approach to determine significant links between languages in various written documents. However, the presented model for inferring significant shared interests has an important advantage. It preserves the average level of activity of each community, but randomizes the temporal order and the concepts that these communities edit. Thus, it brings out the significant connections between communities regardless of their size or activity level.

3.1.2 Step II: Clustering editor communities of similar interests

Constructing a network of shared information interests is only the first step on the way to understanding the mechanisms that bring communities together. Networks are useful when it comes to understanding the interplay of relational ties across the nodes, and clustering is one of the possible visualisation tools to study them.

Clustering allows to group network nodes into communities in which the nodes are more densely interconnected with each other than with the rest of the network. In case of a network of shared co-editing interests, clustering brings to the foreground entire groups of nodes: communities which (1) demonstrate significant interests in co-editing the same concepts, and which (2) are unique in this interest, with respect to the rest of the network.

If the network results in one strongly interconnected component (which is often referred to as a 'hairball network'), it means that all nodes have converged with regards to their information interests. However if clusters are found, then the co-editing profiles are not similar in all regions of the network. This indicates that factors exist that inhibit universal spread of information interests, and create borders between communities. Studying and testing intuitions about these borders helps in understanding the global mechanisms behind shared information interests.

To investigate the presence of effective barriers to global information exchange, I first examine the large-scale structure of the obtained network of shared interests (see Section 3.1). The aim is to highlight the clusters of editor communities that share interest in the same information.

To reveal such clusters among the pairwise connections, I use a network community detection method based on random walks. The clustering algorithm can be envisioned as a random walker game, in which different editor communities are active in sequence. In this relay race, two nodes (editor communities) share interest proportional to the weight of the edge between them. The random walker travels from node to node, picking the next hop proportional to the weight of the outgoing links of its current node. Accordingly, the sequence of nodes forms a random walk, and certain sets of nodes with strong internal connections will be visited for a relatively long time. This process describes a community-detection method known as the map equation (Rosvall et al., 2010; Rosvall and Bergstrom, 2008). I use the map equation's associated search algorithm Infomap (Edler and Rosvall, 2013) to identify the clusters of nodes. By revealing the network's large-scale community structure, I am able to map out the global landscape of information interests based on co-editing behaviour of Wikipedia editor communities.

3.1.3 Step III: Understanding shared interests

Network clusters provide valuable intuitions about the mechanisms that might unite the nodes. Although interpreting clusters is a subjective and somehow qualitative process, I turn these interpretations into quantifiable hypotheses. These hypotheses represent the mechanisms that are hypothetically responsible for some of the observed network structure.

Practically, hypotheses are represented as adjacency matrices where each cell contains the expected edge weight between two nodes given the hypothesis. This weight is essentially the probability that the link exists between two nodes, and can vary between 0 when the link is not expected and 1 if the link is certainly present, given the hypothesis is true.

Visualisation of network clusters provides useful intuitions which help inform hypothesis generation. Depending on configuration of the observed network, access to additional data sources, and researcher's creativity, an arbitrary number of such hypotheses can be constructed and tested against each other. In the following sections I demonstrate that hypothesis matrices can be constructed using the official statistical data on socio-demographics, geographical distances, population densities, language proliferation, colonial history, etc. Hypothesis testing is performed through statistical inference which compares the hypothesis matrices with the adjacency matrix of the co-editing network, and ranking the significant hypotheses according to their explanatory power.

Overall, in this chapter I combine multiple techniques from network theory and statistical inference in a novel three-step approach. This approach allows not only to extract a network of shared interest, but also quantifies intuitions about the processes that produce this network's structure. This approach generalises to other context than Wikipedia, and is flexible with respect to the definition of a community. It is based on quantifying activity levels of community members, and thus is independent from the language of a dataset. It is designed for the context where multiple communities exist and participate in collective creation of a product (in the context of this work – Wikipedia articles). The approach is robust against biases related to different sizes and activity levels within communities. Finally, it allows for testing an arbitrary number of hypotheses related to the structural composition of the network. The rest of this chapter validates this approach by applying it to study empirical Wikipedia co-editing data.

3.2 Validation I: Mapping bilateral information interests

This section presents an analysis of a large dataset of Wikipedia co-editing data. It aims to empirically validate the approach described above, and in doing so, gain insights on the landscape of global bilateral information interests.

In this study, colleagues and I define *communities* by country borders, letting each country represent a separate community of interest. In particular, this study investigates how today's world map of information interests look like. It also asks, what factors may create barriers to information exchange between countries.

This section is based on the results published in the Palgrave Communications journal (Karimi et al., 2015), and presented at the First Conference on Computational Social Science in Helsinki, Finland in 2016. My contributions to this work are outlines in Section 1.3. In order to reflect the fact that this is a collaborative work, where justified, the narrative switches to plural academic 'we'.

"We live in a global world" has become a cliché (Kose and Ozturk, 2014). Historically, the exchange of goods, money, and information was naturally limited to nearby locations, since globalization was effectively blocked by spatial, territorial, and cultural barriers (Cairncross, 2001). Today, new technology is overcoming these barriers and exchange can take place in an increasingly international arena (Friedman, 2000). Nevertheless, geographical proximity still seems to be important for the trade of goods (Fagiolo et al., 2010; Kaluza et al., 2010; Overman et al., 2003; Serrano et al., 2007) as well as for mobile phone communication (Lambiotte et al., 2008) and scientific collaboration (Pan et al., 2012). However, since the Internet allows information to travel more easily and rapidly than goods, it remains unclear what are the effective barriers of global information exchange. As information exchange requires shared interests, we therefore need to better understand global connections in interest, and the factors that form these connections.

Although globalization of information has been discussed extensively in the research literature (Fischer, 2003; Friedman, 2000; Nye Jr, 2004), currently there is no method to quantitatively map bilateral information interests from large-scale data. Without such a method, it becomes difficult to justify qualitative statements about, for example, the complex interplay between shared values and conflict on a global scale. I use data mining and statistical analysis to device a measure of bilateral information interests, and apply this measure to construct a world map of information interests.

To study interests on a global scale, I use the free online encyclopedia Wikipedia, which has evolved into one of the largest collaborative repositories of information in the history of mankind (Mesgari et al., 2014). The free online encyclopedia consists of almost 300 language editions, with English being the largest one (Wikipedia, 2014). This multilingual encyclopedia captures a wide spectrum of information in millions of articles. These articles undergo a peer-reviewed editing process without a central editing authority. Instead, articles are written, reviewed, and edited by the public. Each article edit is recorded, along with a time-stamp, and, if the editor is unregistered, the computer's IP address. The IP address makes it possible to connect each edit to a specific location. Therefore one can use Wikipedia editors as sensors for mapping information interest to specific countries.

Approach. In this study, colleagues and I use Wikipedia editors as sensors for mapping information interest to specific countries. In particular, we use co-editing of the same Wikipedia article as a proxy for shared information interests. To find global connections, we look at how often editors located in different countries co-edit the same concepts on Wikipedia. To infer connections of shared interest between countries, we develop a statistical model and represent significant correlations between countries as links in a global network. In order to explain the global structure of the network, we use regression to test hypotheses about the factors that may impact the formation of shared information interests.

Empirical questions. The particular focus of this study is on the following research questions:

- **RQ1:** What does today's world map of information interests look like?
- **RQ2:** What factors create the barriers of information exchange between countries?

Empirical findings. We quantitatively construct a global network of bilateral information interests based on the Wikipedia co-editing activity. Through structural analysis of the network, we find that countries can be mapped into 18 clusters with similar information interests. Statistical analysis of the network ties suggests that interests are polarized by factors related to geographical proximity, language, religion and historical background. We quantify the effects of these factors using regression analysis and find that information exchange indeed is constrained by the impact of social and economic factors connected to shared interests.

Contributions. We devise a scalable statistical model that identifies countries with similar information interests and measures the countries' bilateral similarities. By including over 10 years of Wikipedia editing in almost 300 language editions, we are able to quantitatively construct a truly global world map of bilateral information interests, and for the first time gain a bird's eye view perspective on the structure of information highways interconnecting various countries. This research pushes forward the literature on globalization and communication, and highlights the efficient barriers on the highways of cross-national and cross-cultural information exchange.

3.2.1 Data collection

As one of the largest and most linguistically diverse repositories of human knowledge, Wikipedia has become the world's main platform for archiving factual information (Mesgari et al., 2014). One important feature of Wikipedia is that every edit made to an article is recorded. Thanks to this detailed data, Wikipedia provides a unique platform for studying different aspects of information processes, for example, semantic relatedness of topics (Auer and Lehmann, 2007; Radinsky et al., 2011), collaboration (Keegan et al., 2012; Kimmons, 2011; Török et al., 2013), social roles of editors (Welser et al., 2011), and the geographical locations of Wikipedia editors (Lieberman and Lin, 2009).

In this work, we used data from Wikipedia dumps[2] to select a random sample from the English Wikipedia edition, which is the largest and most widespread language edition. In total, the English edition has around 10 million articles, including redirects and duplicates. Since retrieving the editing histories of all articles is computationally demanding, we randomly sampled more than six million articles from this set. For each English article, we retrieved the complete editing history of the same article in all language editions that the English Wikipedia page links to. Finally we merged all language editions together to create a global editing history for each article. For each edit, the editing history includes the text of the edit, its time-stamp, and, for unregistered editors, the IP address of the editor's computer. From the IP address associated with the edit, we retrieved the geolocation of the corresponding editor using an IP database [4]. For the purpose of spatial analysis, we limited the analysis to edits from unregistered editors, because data on the location for most of the registered Wikipedia editors are unavailable. The resulting dataset contains more than six million (6,285,753) Wikipedia articles and about 140 million edits in total. We use these edits to create interest profiles for countries.

Relating information interests to geographical location. We identify the interest profile of a country by aggregating the edits of all Wikipedia editors whose IPs are recorded in the country. If an article is co-edited by editors located in different countries, we say that the countries share a common interest in the information of the article. In other words, we connect countries if their editors co-edit the same articles. Indirectly, we let individuals who edit Wikipedia represent the

[2] Available on http://dumps.wikimedia.org/enwiki/[3]
[4] We used http://www.ip2location.com/[5]

population of their country. While Wikipedia editors in a country certainly do not represent a statistically unbiased sample, there is a higher tendency that they edit contents that are related to the country in which they live (Hecht and Gergle, 2010b). Therefore, we approximate the interest profile of a country with collective editing behavior of editors in that country.

Inferring the location of all editors on the country level is non-trivial. Although we have data on all edits, we do not know the location of registered editors because their IPs are not recorded. One proposed approach to tackle this problem makes use of circadian rhythms of editing activity to infer the location of the editors (Yasseri et al., 2012a). This method approximates the longitude of a location but provides little information about its latitude. Therefore, we must limit the analysis to the activity of unregistered editors with recorded IP addresses. This will arguably affect the results. Not only do registered editors contribute to 70% of all 140 million edits, they also have somewhat different behaviour. For example, many of the most active registered users take on administrative functions, develop career paths, or specialize in covering selected topics (Arazy et al., 2015). On the other hand, some unregistered editors are involved in vandalism, but often their activity nevertheless indicates their interest.[6] While we can only speculate about how including registered editors would affect the results, unregistered editors can nevertheless provide useful information about shared interests between countries.

3.2.2 Approach and results

I discuss the results at four levels of detail, from the big picture to the detailed dynamics, and highlight different potential mechanisms for barriers of information exchange. First, I show a global map of countries with shared information interests, and continue with the interconnections between the clusters. Then I consider each cluster separately and examine the interconnections between countries within the clusters. Finally, I apply multiple regression analysis to examine explanatory variables that may stimulate or hinder global information exchange.

The world map of information interests

The network of shared co-editing interests is inferred based on the interest model introduced in Section 3.1. The Bonferroni-corrected threshold for the significant link weight is $t = 3.52\sqrt{L}$, where $L = 6,285,753$ is the number of Wikipedia articles. 3.52 is derived from the condition that $P(z > 3.52) = 0.05/N$, where $N = 234$ is the number of countries, and P is the standard Gaussian distribution (with zero average and unit variance).

Results. The resulting network is illustrated as a map in Fig. 3.2, where countries of the same cluster share the same color. Between the 234 countries, we identified 2,847 significant links that together form a network of article co-edits. By clustering the network, we identified 18 clusters of strongly connected countries (see Table 3.1 for a detailed list of countries in each cluster). The world map of information interests suggests that cultural and geopolitical features can explain the division of countries. For example, the United States and Canada share a long geographical border and extensive mutual trade, and are clustered together despite the fact that other English-speaking countries are not. Moreover, religion is a plausible driver for the formation of the cluster of countries in the Middle East and North Africa, as well as the cluster of Russia and the Orthodox Eastern-European countries (Gupta et al., 2002). Another factor in the formation of shared information interests is language. For example, countries in Central and South America are divided into two clusters with Portuguese and Spanish as common languages in each cluster, respectively. Colonial history can also shape similarity in interests, as in the cluster of Portugal, Angola and Brazil, as well as the cluster of former Soviet Union countries (Hensel, 2009). Overall, there is

[6] See Wikipedia's policy and fight against vandalism here: https://en.wikipedia.org/wiki/Vandalism_on_Wikipedia

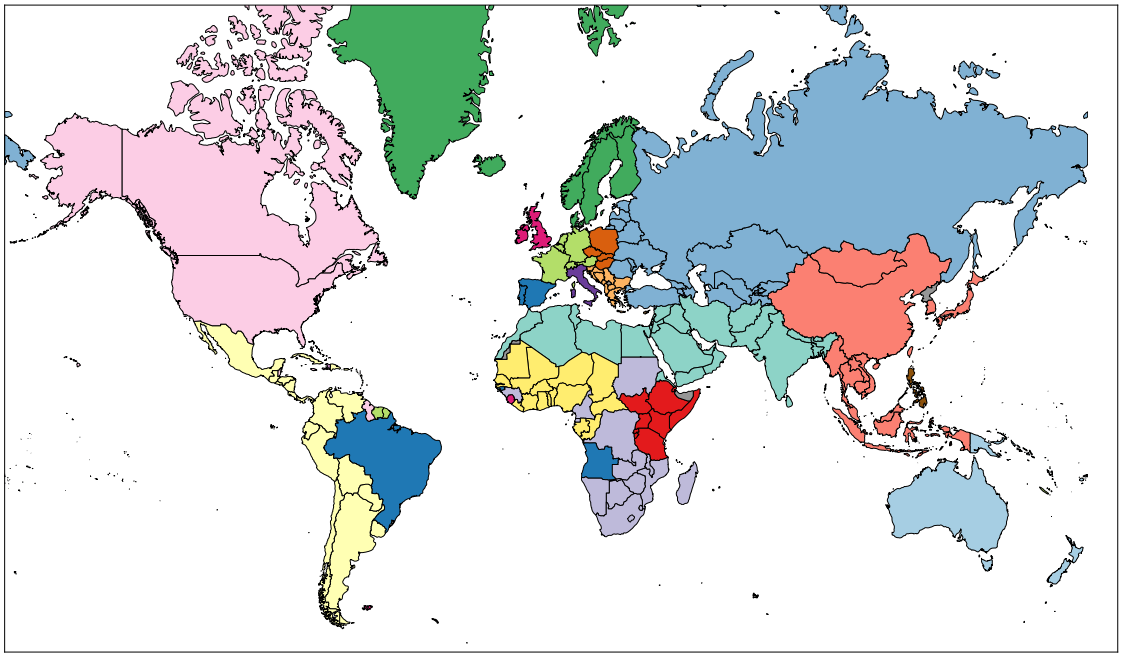


Figure 3.2: World map of information interests based on the national Wikipedia co-editing profiles. The network of significant co-editing interests is clustered using the Infomap algorithm (Section 3.1). Countries that belong to the same cluster have the same color. Countries colored in gray do not belong to any cluster. We find 18 country clusters of shared information interests. The map suggests that this division of countries can be related to a number of cultural and geopolitical features.

strong empirical evidence that geographical proximity, common religion, shared language, and colonial history can explain the division of countries.

Interconnections between the clusters

To examine the connections between clusters, I look at the network structure at the cluster level. The network in Fig.3.3 shows the connections between the clusters of countries illustrated in Fig. 3.2 with the same color coding. Connections tend to be stronger between clusters of geographically proximate countries also at this level. Interestingly, the Middle East cluster in turquoise has the strongest outlinks to other clusters, forming a hub that connects East and West, North and South. Interpreting the strong connections as potential highways for information exchange, the Middle East is not only a melting pot of ideas, but also seems to play an important role in the exchange of information.

Bilateral ties within the clusters

To get better insights into how the clusters are shaped, I zoom into the inter-country networks within clusters. In the upper left corner of Fig. 3.3, I show the strongest connections within the Central European cluster. It suggests that the links between some countries can be related to the overlap in their official languages. For example, Belgium has three official languages, Dutch, French and German. Indeed, Belgium is connected closely with the Netherlands, France, and Luxembourg. We observed the same pattern in other clusters, and the triad of Switzerland, Germany,

[7]Countries that are not included in the upper left panel with the Central European cluster are Suriname, French Polynesia, New Caledonia, Mayotte, RE, Saint Pierre and Miquelon, and North Korea.

and Austria is another example of strongly linked countries with a shared language. In the context of Wikipedia editing, the effect of multilingualism on article editing might show itself in the presence of editors who contribute simultaneously to several language editions (Hale, 2014b).

In order to illustrate what shared interests can shape the bilateral connections, I look at a number of concrete examples. First, I rank the concepts c according to their significant z_{ij}^{c}-scores for each pair of countries i and j, focusing on the top-ranked concepts. In Table 3.2 I report on the results for two European country pairs: Germany–Austria in the European cluster, and Sweden–Norway in the Scandinavian cluster. In both cases, the concepts with the most significant co-edits relate to local and regional interests, including sports, media, music, and places. For example, the top-ranked concepts in the Germany–Austria list include an Austrian singer who is also popular in Germany, and an Austrian football player who is playing in the German league. The top-ranked concepts in the Sweden–Norway list shows a similar pattern of locally related topics, for example, a host of a popular TV show simultaneously aired in Sweden and Norway, a Swedish football manager who has been successful both in Sweden and Norway, and a music genre that is nearly exclusive to Scandinavian countries. Altogether, the top concepts suggest that an important factor for co-editing is related interests, which in turn may be an effect of shared language, religion, or

Cluster	Countries
1	Saudi Arabia, United Arab Emirates, Egypt, India, Kuwait, Jordan, Qatar, Pakistan, Bahrain, Palestine, Oman, Algeria, Morocco, Lebanon, Syria, Iraq, Tunisia, Yemen, Bangladesh, Libya, Sri Lanka, Iran, Nepal, Maldives, Israel, Mauritius, Afghanistan, Bhutan, Eritrea
2	Argentina, Colombia, Venezuela, Guatemala, Peru, Mexico, Chile, Ecuador, Uruguay, Honduras, Costa Rica, Dominican Republic, Panama, Paraguay, El Salvador, Bolivia, Puerto Rico, Nicaragua, Cuba
3	Hong Kong, Taiwan, China, Malaysia, Singapore, South Korea, Vietnam, Indonesia, Thailand, Macau, Japan, Cambodia, Burma (Myanmar), Brunei, Mongolia, Laos, Timor-Leste
4	Russian Federation, Ukraine, Belarus, Azerbaijan, Kazakhstan, Latvia, Armenia, Estonia, Georgia, Lithuania, Moldova, Romania, Turkey, Uzbekistan, Kyrgyzstan, Turkmenistan, Tajikistan
5	Serbia, Bosnia and Herzegovina, Montenegro, Croatia, Macedonia, Greece, Bulgaria, Slovenia, Cyprus, Albania
6	South Africa, Sudan, Zimbabwe, Cameroon, Democratic Republic of the Congo, Botswana, Zambia, Namibia, Mozambique, Swaziland, Equatorial Guinea, Guinea, Madagascar, Malawi, Lesotho, Sao Tome and Principe
7	France, Switzerland, Austria, Germany, Belgium, Netherlands, Luxembourg, Monaco, Suriname, Liechtenstein, French Polynesia, New Caledonia, Mayotte, Réunion, Saint Pierre and Miquelon
8	United States, Canada, Bermuda, Palau, Bahamas, Caribbean Islands*
9	Nigeria, Senegal, Ghana, Ivory Coast, Burkina Faso, Benin, Mauritania, Mali, Liberia, Niger, Gambia, Gabon, Togo, Republic of the Congo, Chad, Central African Republic
10	Kenya, Uganda, Djibouti, Somalia, Tanzania, Rwanda, Ethiopia, South Sudan, Burundi, Comoros
11	Sweden, Denmark, Norway, Finland, Greenland, Faroe Islands, Iceland, Åland Islands, Malta
12	Curaçao, Saint Martin, Guadeloupe, Sant Maarten, French Guiana, Aruba, Martinique, Haiti, Wallis and Futuna
13	Slovakia, Czech Republic, Hungary, Poland, Niue
14	Fiji, New Zealand, Australia, Samoa, Vanuatu, Kiribati, Cook Islands, Tonga, Solomon Islands, Papua New Guinea, Nauru, Marshall Islands, American Samoa, Norfolk Island
15	Spain, Portugal, Angola, Brazil, Cape Verde, Andorra, Guinea-Bissau
16	United Kingdom, Ireland, Guernsey, Jersey, Isle of Man, Sierra Leone, Gibraltar, Falkland Islands, Tuvalu, British Indian Ocean Territory
17	Philippines, Guam, Northern Mariana Islands, Micronesia
18	Italy, San Marino, Holy See (Vatican City)

* Caribbean Islands in the list are: Jamaica, Trinidad and Tobago, Saint Lucia, Barbados, Antigua and Barbuda, Guyana, Saint Kitts and Nevis, Grenada, Saint Vincent and the Grenadines, Belize, US Virgin Islands, Dominica, Cayman Islands, British Virgin Islands, Anguilla, Turks and Caicos Islands.

Table 3.1: Clustering results. In total 234 countries are assigned to 18 clusters

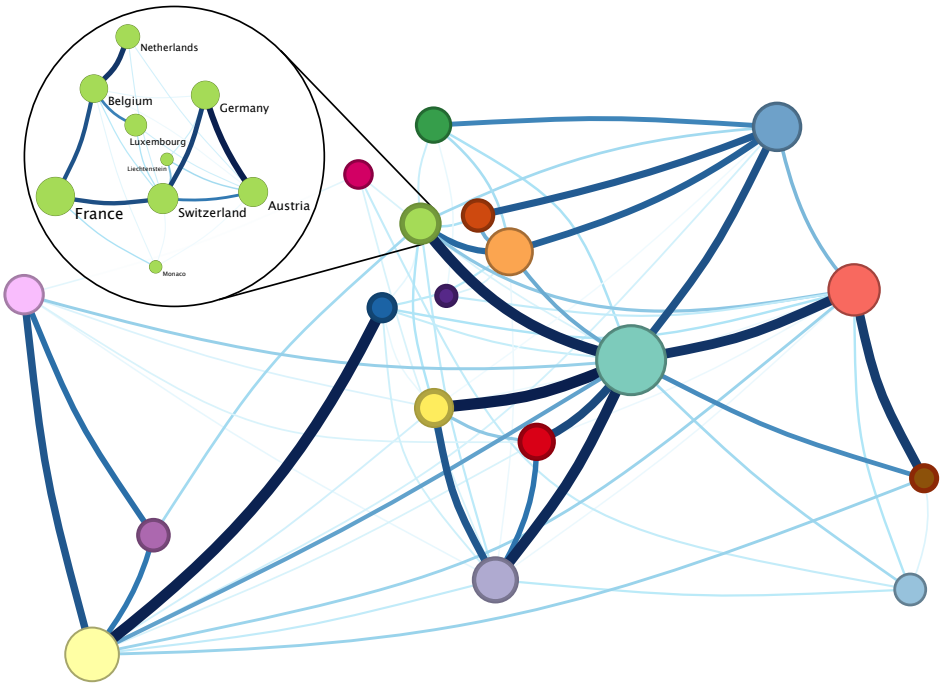


Figure 3.3: World network of information interests. The size of the nodes represents the sum of all link weights within the clusters. The links represent connections between clusters obtained from the cluster analysis with Infomap (see Section 3.1); the thicker the line, the stronger the connection. Clusters are coloured in the same way as in Fig. 3.2. The upper left corner shows the most significant connections between countries in the Central European cluster[7].

colonial history, as well as geographical proximity or large volume of trade between countries.

Regression analysis of hypotheses related to the strengths of shared interests

During the previous sections of analysis, several hypotheses became evident that might explain some of the structural variation in the extracted network of bilateral interests. These hypotheses point that geographical proximity, trade relationships, past colonial ties, linguistic and religious factors might all play role in the countries' shared information interests.

Rank	Germany and Austria	Sweden and Norway
1	Christina Stürmer	Tipuloidea
2	Erste Allgemeine Verunsicherung	Dansband
3	Steffen Hofmann	Boyoz
4	Nazar (rapper)	Fredrik Skavlan
5	Piefke	Erik Hamrén
6	ATV (Austria)	Sweden
7	Klagenfurt	Anders
8	Karl-Heinz Grasser	Peter Jöback
9	Wolf Haas	Causerie
10	Zillertal	List of the busiest airports in the Nordic countries

Table 3.2: Top 10 Wikipedia concepts co-edited by country pairs Germany-Austria and Sweden-Norway, according to the filtering analysis based on the interest model. These examples suggest that locality plays an important role in the patterns of shared interests, which include regionally popular figures such as sport stars and musicians, TV-shows, and locally relevant geographical objects.

In order to test these hypotheses in a formalised way, I apply Multiple Regression Quadratic Assignment Procedure (MRQAP) analysis. This method is specifically suited when there are collinearity and autocorrelation in the data (Dekker et al., 2007; Krackhardt, 1988). The MRQAP analysis is performed using the `netlm` function in the `sna` R package (Butts, 2008). The dependent variables in the regression model are the significant z_{ij}-scores that have been obtained from the data, i.e. weighted links of the co-editing network. The independent variables represent the intuitions about the socio-economic factors that might contribute to the observed network structure, and include the following hypotheses:

- **H1: Geographical proximity**
  The geographical proximity is calculated as the Euclidean distance between each pair of countries using Haversine formula. The point for each country is based on the rounded latitude and longitude of the centroid or the center point of the country (CIA, 2011).

- **H2: Trade data**
  Data on free trade areas and customs union are collected for the year 2000 (WTO, 2015). World Trade Organisation members are obliged to notify the regional trade agreements in which they participate. 157 countries reported their trade flow in 2000. The trade volume is estimated by averaging the import and export volume of each pairs of countries (Subramanian and Wei, 2007).

- **H3: Colonial ties**
  This data come from the Colonial history dataset available for download at (Hensel, 2009). There is a tie of weight one between two countries if they had a colonial relationship since 1900 until now.

- **H4: Language similarities**
  The data on the languages that are spoken in each country are collected from the Ethnologue database (Ethnologue, 2015). The data contain information on more than 7000 known alive languages in the world. It is regarded to be one of the most comprehensive sources of information on language usage. Based on the data, the weight of the tie between each pair of countries is calculated based on the number of languages that are co-spoken. For example the weight between Sweden and Finland is 5 because there are co-spoken languages: 'fin' : Finnish, 'fit' : Finnish, Tornedalen, 'rmf' : Romani, Kalo Finnish, 'sme' : Northern Saami, 'swe' : Swedish.

	R_0	R_1	R_2	R_3	R_4
Intercept	0.41	0.3	2.33	2.33	2.28
Shared language	0.91* (69)	0.82* (64)	0.77* (60)	0.75* (58)	0.74* (57)
Shared religion		2.76* (46)	2.6* (44)	2.6* (43)	2.44* (40)
Log distance			-0.23* (-23)	-0.23* (-23)	-0.23* (-23)
Colonial tie				4.5* (22)	4.35* (21)
Log trade					0.03* (10)
Adjusted R-squared	0.13	0.19	0.20	0.21	0.22
F-statistic	7,774	3,590	2,610	2,110	1,716
dF	30,874	30,873	30,872	30,871	30,870

Table 3.3: The results of the MRQAP analysis. Significant edit co-occurrences (z_{ij}-scores) form the dependent variable matrix, which we regress on the independent matrices (representing the hypotheses which explain the observed network structure) in different models. Values in parentheses are t-statistics. The features are ordered by importance, from shared language (most effect) to trade. The number of examined country pairs is 62,001. Values marked with an asterisk have a p-value less than 0.01. The data suggest that shared language and religion are among the strongest facilitators of shared information interests between country communities of editors on Wikipedia.

- **H5: Shared religion**
  The data on the religion composition of countries was taken from World Religion database (WRD, 2015). The results is a binary matrix where countries that share the same religion have a tie of weight one, and zero otherwise.

Results. All independent variables show significant correlation with the data (see Table 3.3). To observe the variation between different independent matrices, I combine them in different models. In model R_0, the influence of shared language explains 13% of the observed network. In model R_1, by adding the shared religion hypothesis, the power of the model increases to 19%. After adding the geographical proximity of countries, the model R_2 shows a slight increase the R-squared. The observed relationship between inter-country distances and the z_{ij}-scores is negative, since short distances correspond to high proximity. Models R_3 and R_4, respectively, add colonial ties and trade hypotheses. Including all these explanatory variables into the regression model improves the explanatory power of the model to 22%. The correlation of each variable with the observed z_{ij}-scores can be inferred from the t-statistic. Shared language shows the strongest association, followed by shared religion, geographical proximity, colonial ties, and volume of trade (see Table 3.3).

3.3 Validation II: Linguistic neighbourhoods

This section provides further validation of the approach. It continues the empirical investigation of global shared information interests, this time, exploring in detail *interest profiles of linguistic communities*. It in particular focuses on how different cultural communities select and document their cumulative knowledge in different language editions of Wikipedia.

This section is based on the results previously published at the European Physics Journal Data Science (Samoilenko et al., 2016), and presented at the WebScience conference in Oxford, UK in 2015, at the NetSciX conference in Wroclaw, Poland in 2016 where it won an award for the best poster, and at the IC2S2 conference in Cologne, Germany in 2017. My contributions to this work are outlined in Section 1.3. In order to reflect the fact that this is a collaborative work, where justified, the narrative switches to plural academic 'we'.

Wikipedia is the largest crowd-sourced encyclopedia today. Is is also a platform that allows editors from multiple backgrounds to document knowledge in different language editions. The collective traces left by editors of Wikipedia can be utilized as proxies for cultural communities. Thus, by examining the overlap in the knowledge that these communities preserve through editing, it is possible to gain an impression of how culturally proximate or different the corresponding communities are. Certainly, co-editing similarities among language communities of Wikipedia editors are just a particular dimension of culture and are not representative of cultural similarities among the communities in general. Yet, Wikipedia plays a critical role in today's information gathering and diffusion processes and Wikipedians constitute an important cultural subset of educated and technology-savvy elites who often drive the cultural, political, and economic processes (Ronen et al., 2014).

Empirical questions. In this analysis, I tap into the traces left by editors of Wikipedia to gain new insights into how language communities Wikipedia editors relate to each other via the shared information interests. This analysis focuses of the following research questions:

- **RQ1:** What are the common editing interests between language communities on Wikipedia?
- **RQ2:** What factors can explain the landscape of these shared interest ties between the language communities?

Approach. This analysis assumes that collective interest of a language-speaking community is reflected through the aggregation of articles documented in the corresponding language edition of Wikipedia. These articles are thus used as an approximation of the topics which are culturally relevant to that language community. It is important to note, though, that by no means they are representative of the entire underlying cultural community. I define *cultural similarity* as a significant interest of language communities in editing articles about the same topics. In other words, the interests of language communities are similar when both communities significantly agree regarding the choice of topics they edit.

The approach consists of several steps. I first use statistical filtering (see Section 3.1) to identify language pairs which show consistent interest in articles on the same topics. Based on this dyadic information, I create a network of interest similarity where nodes are languages and links are weighted as the strength of shared interest between them. Then I cluster the network and inspect it visually to inform the generation of hypotheses about the mechanisms that contribute to cultural similarity. Finally, I express these hypotheses as transition probability matrices, and test their plausibility using two statistical inference techniques – HypTrails (Singer et al., 2015) and MRQAP (Krackardt, 1987). Using both Bayesian and frequentist approaches, I obtain similar results, which suggests that the findings are robust against the chosen statistical measure.

Empirical findings. This study finds that the topics that each language edition documents are not selected randomly, however small the underlying community of editors. I test several

hypotheses about the underlying processes that might explain the observed non-randomness, and find that bilingualism, linguistic similarity of languages, and shared religion provide the best explanations for the similarity of interests between cultural communities. Population attraction and geographical proximity are also significant, but much weaker factors bringing communities together.

Contributions. The main contribution is empirical. Colleagues and I expand the literature on culture-related research by (a) presenting a large-scale network of interest similarities between 110 language communities, (b) showing that the set of languages covering a concept of Wikipedia is not a random choice, and (c) by statistically demonstrating that similarity in concept sets between Wikipedia editions is influenced by multiple factors, including bilingualism, proximity of these languages, shared religion, and population attraction. We also combine multiple techniques from network theory, Bayesian and frequentist statistics in a novel way, and present a generalisable approach to quantify and explain culture-related similarity based on editing activity of Wikipedia editors.

The remainder of this chapter is structured as follows. I first describe in detail the process of data sampling and collection (Section 3.3.1). Section 3.3.2 focuses on identifying and explaining co-editing interests, gives a technical overview of the quantitative methods, and finally, reports the results.

3.3.1 Data collection

There are almost 300 language editions of the encyclopedia, which vary greatly in size. This makes selecting the sample a nontrivial decision: on the one hand, many editions are rather small. For example, at the time of writing (2017) there are 8 Wikipedia language editions which each contain less than 10 articles. Thus, even the entire data from them would not be sufficient for statistical analysis. On the other hand, the purpose of this analysis is to preserve as many linguistic dimensions as possible while trying to extract the global network of shared information interests. As a compromise, this analysis focuses on a sample of 126 largest editions which contained more than 10,000 article pages, as of July 2014 (Wikipedia, 2016).

Sampling procedure. To account for variations in editions' age, number of active contributors, and growth rates, I selected the time frame such that (1) to ensure a sufficient amount of editions existed in the beginning of the observation; and (2) to allow enough time for each edition to accumulate concepts. I traced back each edition to its first registered article page, and found out that 110 out of 126 largest editions had been created before 01.01.2005. I excluded 11 editions which appeared later (min, vo, be, new, pms, pnb, bpy, arz, mzn, sah, vec) and those whose language codes could not be mapped to the ISO 639-1 standard (be-x-old, zh-yue,bat-smg, map-bms, zh-min-nan). The remaining 110 editions became the focus of my subsequent analysis which covers the period of 9 years between 01.01.2005 and 31.12.2013.

I sampled from each edition separately, collecting IDs of all article pages created between 2005 and 2013 (excluding other types of pages, redirects, and pages created by bots). For each ID I also collected the entire editing history in all linked language editions. Thus, each ID corresponds to a concept[1] (the topic of the article regardless of the language), and all interlinked language editions represent various linguistic points of view on the concept. After removing duplicates, the dataset includes 3,066,736 unique concepts and a total of 1,360,647,795 article pages in different languages. The data were collected between 20.12.2015 and 25.01.2016 from Wikimedia servers directly, using the access provided by Wikimedia Tool Labs (Wikimedia, 2015).

As a note, one algorithmic limitation of such approach could be hidden in relying on Wikipedia's inter-language link graph to identify articles on the same concepts in different language editions. This has some known issues with the lack of triadic closure and dyadic reciprocity (Bao et al.,

2012). To ensure that the maximal set of interlanguage links related to a concept is retrieved, I collect all articles with their interlanguage links from each edition separately, removing duplicates afterwards. Thus, all existing inter-language links are extracted.

3.3.2 Approach and results

In this section, I describe the procedure of extracting cultural similarities from co-editing activity in Wikipedia, and present the network of significant shared interests between 110 language communities. The section begins with summarising the pre-analysis check of whether the language-concept overlap in Wikipedia is random.

Testing for non-randomness of co-editing patterns

Theoretically, each concept covered in Wikipedia could exist in all 288 language editions of the encyclopedia. This is possible because Wikipedia does not censor topic inclusion depending on the language of edition, and anyone is free to contribute an article on any topic of significance. However in practice, such complete coverage is very rare, and concepts are covered in a limited set of language editions. Is this set of languages random? To answer this question, I analyse matrices of language co-occurrences based on a random sample of the data (200,748 concepts).

This analysis is based on a random sample of $N = 200{,}748$ concepts. At first, I construct the matrix of empirical co-occurrences C_{ij}, based on the probability of languages i, j to have an article on the same concept. I also construct a synthetic dataset where I preserve the distribution of languages and the number of concepts, $N = 200{,}748$, but allow languages to co-occur at random. I use the resulting data to produce the matrix of random co-occurrences C_{ij}^{rand}, and compare it to the matrix of co-occurrences C_{ij} by calculating confidence intervals. This null model corresponds to the belief that in Wikipedia each concept has equal chances to be covered by any language, with larger editions sharing concepts more frequently purely because of their size. Comparing two matrices reveals some preliminary intuitions regarding the extent to which co-editing patterns are non-random.

I establish that language dyads do not edit articles about the same concept (co-occur) by chance. Large editions share concepts more frequently than expected: although in the data *EN-DE* and *EN-FR* overlap in 45% of cases, only 15% is expected by the null model. To little surprise, the amount of overlap between editions in the data decreases with the size of the editions. One notable exception is the Japanese edition which, despite being among the ten largest Wikipedias, co-occurs with other top editions noticeably less frequently. Similarly, the Uzbek edition, being among the ten smallest in the dataset, shows high concept overlap with large editions. By simply plotting frequencies of co-occurrences, I do not observe any local blocks or clusters, neither among large nor small editions (see Fig. 3.4). These overlap differences are statistically significant (95% confidence level), and the null model explains only 1,386 out of 11,990 language pairs (11% of observed data, white cells in the matrix). Such low explained variation suggests that concept overlap is not random and cannot be explained only by edition sizes. Instead, there are non-random, possibly cultural processes, that influence which languages cover which concepts on Wikipedia. Having evidence that the data contain a signal, I continue the investigation by performing network analysis.

The network of shared interests among the language communities

The procedure for extracting the network structure is described in detail in Section 3.1. The Bonferroni-corrected threshold for link significance in the right tail is $t = 3.32\sqrt{L}$, where $L = 3{,}066{,}736$ is the number of concepts. 3.32 is derived from the condition that $P(z > 3.32) = 0.05/N$, where $N = 110$ is the number of languages, and P is the standard Gaussian distribution (with zero

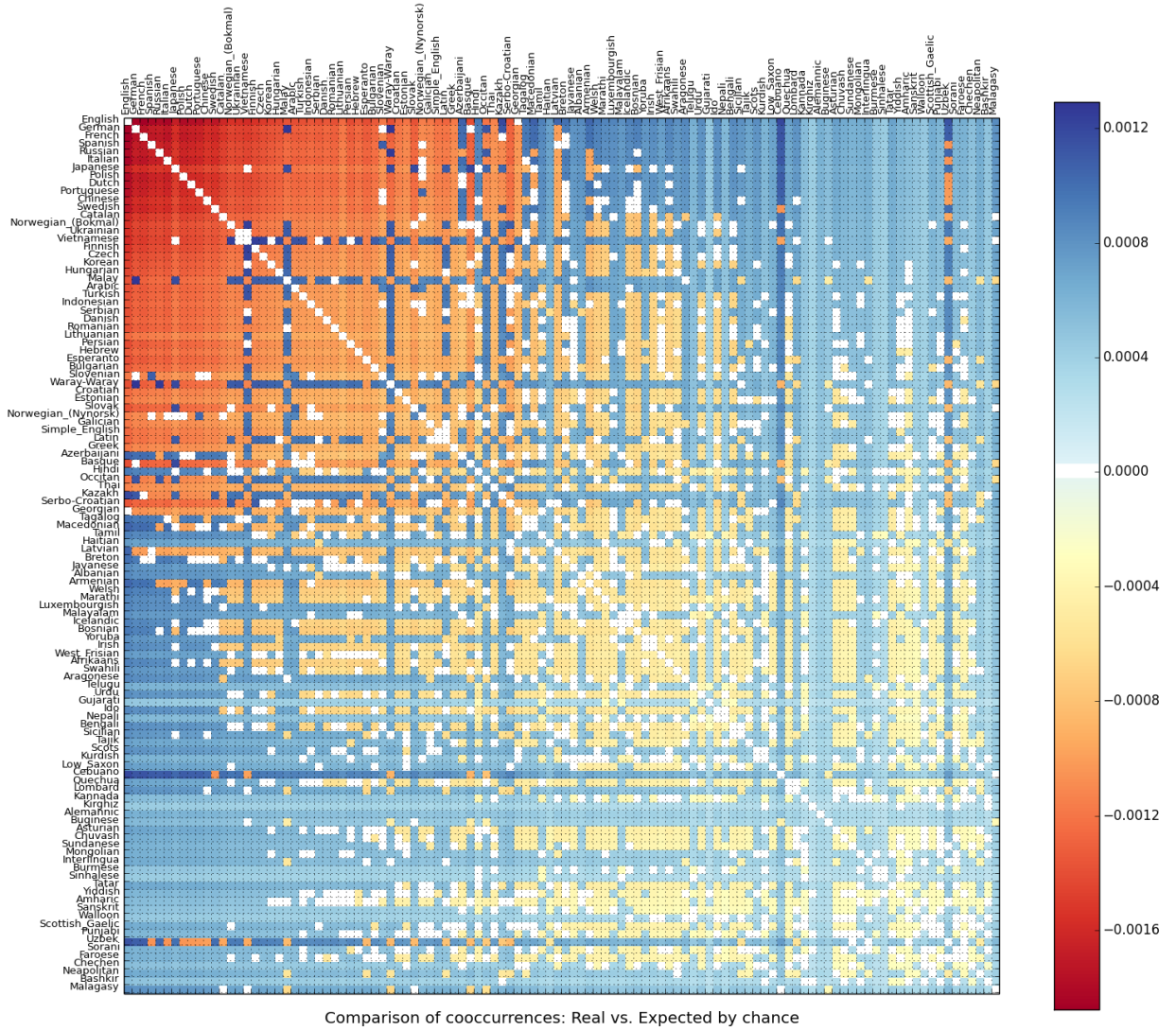


Figure 3.4: Comparison of empirical and experimental data on editing co-occurrences. Based on a random sample of the data (N = 200,748 concepts). White cells are explained by the null model, shades of blue/red show the distance of observed co-occurrences from the lower/upper border of the confidence interval. Low explained variation (11%, 95% confidence level), suggests that non-random processes are in place.

average and unit variance). I use the Infomap algorithm (Rosvall et al., 2010) to identify language communities that are most similar in their interests (see Section 3.1.2). Additionally, I compare these results with the Louvain clustering algorithm (Blondel et al., 2008) and establish that both methods show high agreement for this dataset.

Results. Cluster analysis suggests that no language community is completely separated from other communities, and in fact, there are significant topics of common interest between almost any two language pairs. I reveal 21 clusters of two and more languages, plus 9 languages that are identified as separate clusters (see Table 3.4). Notably, English forms a self-cluster, and this independent standing means little uniqueness in interest similarity between English and other languages. This is an interesting finding in the light of the recent discussions on whether English is becoming a global language and the most suitable *lingua franca* for cross-national communication (Crystal, 2003). The entire global network of share interests across language communities on Wikipedia is visualised in Fig. 3.5. The links within clusters are weighted according to the amount of positive deviation of z-score per language pair from the threshold of randomness. Stronger

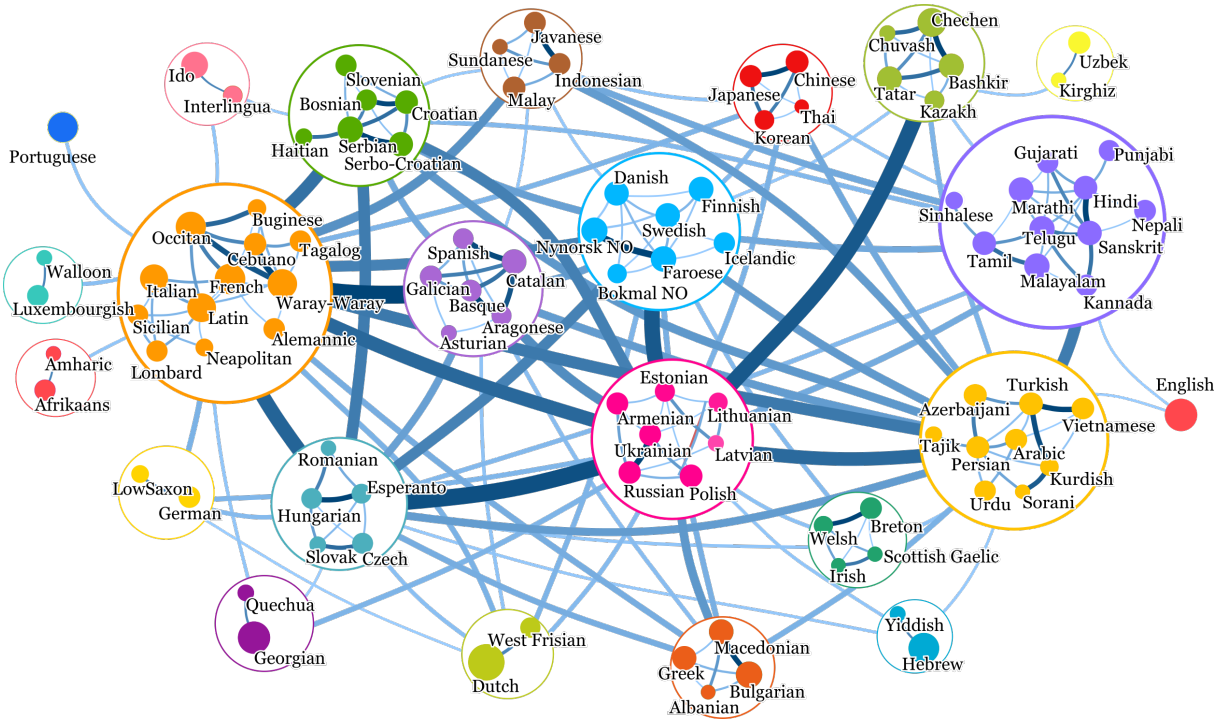


Figure 3.5: The network of significant Wikipedia co-editing ties between language pairs. Nodes are coloured according to the clusters found by the Infomap algorithm (Edler and Rosvall, 2013), and link weights within clusters represents the positive deviation of z-scores from the threshold of randomness; links are significant at the 99% level. For visualisation purposes we display only 23 clusters and the strongest inter-cluster links in the network. The inter-cluster links show the aggregated z-scores between all nodes of a pair of clusters. The network suggests that local factors such as shared language, linguistic similarity of languages, shared religion, and geographical proximity play a role in interest similarity of language communities. Notably, English forms a separate cluster, which suggest little interest similarity between English speakers and other communities.

weights indicate higher similarity. All links are significant at the 99% level. The inter-cluster links should be interpreted with care in the context of this analysis, as they are weighted according to the aggregated strength of connection between all nodes of both clusters. The network is undirected since it depicts mutual topical interest of both language communities, which is inherently bidirectional. For visualisation purposes, I display only the strongest inter-cluster links and 23 language clusters. Full cluster membership information is detailed in Table 3.4.

Cluster interpretation. Visual inspection of language clusters suggests a number of hypotheses which might explain such network configuration. For example, (1) geographical proximity might explain the Swedish-Norwegian-Danish-Faroese-Finnish-Icelandic cluster (light blue), since those are the languages mostly spoken in Scandinavian countries. Other groups of languages form around (2) a local *lingua franca*, which is often an official language of a multilingual country, and include other regional languages which are spoken as second- and even third language within the local community. This way, Indonesian and Malay form a cluster with Javanese and Sundanese (brown), which are the two largest regional languages of Indonesia. Similarly, one of the largest clusters in the network (purple) consists of 11 languages native to India, where cases of multilingualism are especially common, since one might need to use different languages for contacts with the state government, with the local community, and at home (Crystal, 2003). Another interesting example is the cluster of languages primarily spoken in the Middle Eastern countries (yellow), which apart from geographical proximity are closely intertwined due to (3) a shared religious tradition. Finally, some clusters illustrate (4) the recent changes in sociopolitical situation, which can also be partially traced through bilingualism. Following the civil war

of the 1990s in former Yugoslavia, its former official Serbo-Croatian language is now replaced by

Cluster	Language	Weight
1	French	0.01415
	Occitan	0.01372
	Waray-Waray	0.01291
	Latin	0.01219
	Italian	0.01147
	Cebuano	0.00652
	Alemannic	0.00591
	Tagalog	0.00581
	Lombard	0.00566
	Sicilian	0.00536
	Buginese	0.00394
	Neapolitan	0.00355
2	Russian	0.01665
	Polish	0.01584
	Ukrainian	0.01509
	Armenian	0.01190
	Estonian	0.01068
	Lithuanian	0.00994
	Latvian	0.00800
3	Sanskrit	0.01046
	Hindi	0.01032
	Tamil	0.00927
	Malayalam	0.00869
	Telugu	0.00861
	Marathi	0.00826
	Bengali	0.00613
	Kannada	0.00578
	Nepali	0.00578
	Gujarati	0.00537
	Punjabi	0.00528
	Sinhalese	0.00337
4	Norwegian (Bokmal)	0.01818
	Swedish	0.01499
	Norwegian (Nynorsk)	0.01311
	Finnish	0.01093
	Danish	0.01048
	Icelandic	0.00640
	Faroese	0.00432
5	Serbian	0.01884
	Serbo-Croatian	0.01632
	Croatian	0.01417
	Slovenian	0.01133
	Bosnian	0.01036
	Haitian	0.00568
6	Vietnamese	0.01362
	Turkish	0.01022
	Persian	0.00966
	Azerbaijani	0.00965
	Arabic	0.00886
	Urdu	0.00838
	Kurdish	0.00629
	Tajik	0.00523
	Sorani	0.00423
7	Esperanto	0.01717
	Hungarian	0.01552
	Czech	0.01361
	Slovak	0.01159
	Romanian	0.01104

Cluster	Language	Weight
8	Catalan	0.01566
	Galician	0.01011
	Basque	0.00983
	Spanish	0.00903
	Aragonese	0.00864
	Asturian	0.00576
9	Chechen	0.01332
	Bashkir	0.01191
	Tatar	0.01184
	Kazakh	0.01005
	Chuvash	0.00878
10	Indonesian	0.02004
	Malay	0.01410
	Javanese	0.01310
	Sundanese	0.00773
11	Bulgarian	0.01232
	Macedonian	0.01135
	Greek	0.00878
	Albanian	0.00688
12	Chinese	0.00891
	Japanese	0.00792
	Korean	0.00734
	Thai	0.00617
13	Breton	0.00948
	Welsh	0.00648
	Scottish Gaelic	0.00509
	Irish	0.00494
14	Dutch	0.01673
	West Frisian	0.00630
15	German	0.01354
	Low Saxon	0.00520
16	Georgian	0.01042
	Quechua	0.00451
17	Uzbek	0.00797
	Kirghiz	0.00385
18	Hebrew	0.00847
	Yiddish	0.00298
19	Luxembourgish	0.00710
	Walloon	0.00295
20	Ido	0.00612
	Interlingua	0.00390
21	Afrikaans	0.00740
	Amharic	0.00229
22	Portuguese	0.00892
23	Simple English	0.00806
24	English	0.00763
25	Swahili	0.00560
26	Scots	0.00486
27	Yoruba	0.00453
28	Mongolian	0.00349
29	Burmese	0.00238
30	Malagasy	0.00187

Table 3.4: Clusters of languages with shared interest as found by the Infomap clustering algorithm. The weight of each language is the normalized weighted degree of the node. Some languages, including English, do not belong to a larger community and form a self-cluster instead.

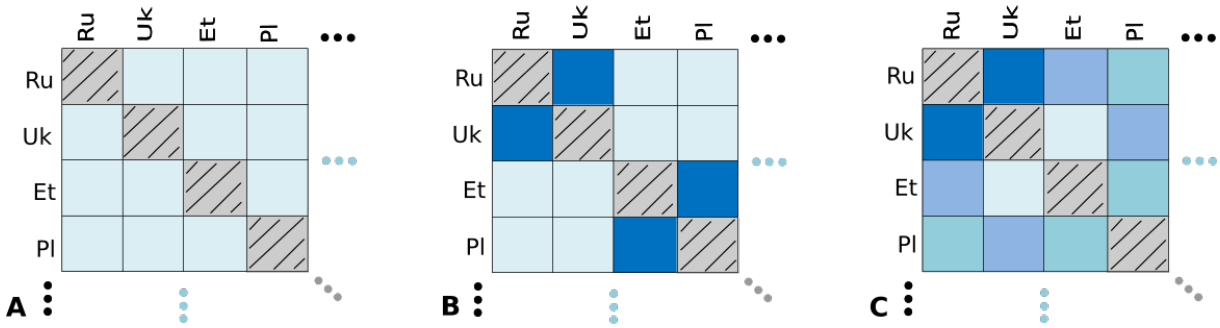


Figure 3.6: A toy example of expressing a hypothesis through a transition probability matrix. The co-editing matrices are symmetrical. The diagonal is empty since the data do not allow self-loops. According to each hypothesis, the cells with more likely transitions are coloured in darker shades of blue. In (a) Uniform hypothesis – all transitions are equally possible, i.e. the editions are co-editing random topics. In (b) Shared religion hypothesis – the dyads Russian-Ukrainian and Polish-Estonian are given more belief on the basis of shared religion. Finally, in (c) Geographical proximity hypothesis – the shorter the distance between languages, the stronger belief in the transition.

three separate languages: Serbian, Croatian, and Bosnian (green cluster). Notably, there is still a separate Serbo-Croatian Wikipedia edition. To give another example, Russian held a privileged position in the former Soviet Union, being the language of the ideology and a priority language to learn at school (Crystal, 2003). Even twenty years after the dissolution of the Soviet Union, Russian remains an important language of exchange between the post-Soviet countries. Similarity of interests between speakers of Russian and the languages spoken in nearby countries, as seen in the magenta cluster, comes as little surprise.

In the following sections, I show how the network of significant shared interests could be used to inform hypothesis formulation. Discussed in the previous sections, the anecdotal interpretations of the clusters are useful to inform the hypotheses about the mechanisms that affect the formation of co-editing similarities. In this section I build upon these initial interpretations and formulate them as quantifiable hypotheses. To evaluate the validity of the hypotheses, I compare their plausibility against one another using two statistical inference approaches. First, I use Bayesian approach and visually compare the strengths of hypotheses. Then I apply frequentist approach to evaluate the explanatory power of different models. I begin by outlining the hypotheses formulation and the necessary methodology and continue with reporting the results.

Explaining the clusters of co-editing interests: Hypothesis formulation

I convert the initial interpretation of the network clusters discussed in Section 3.3.2 into quantifiable hypotheses, which are expressed through transition probability matrices. A toy example of this process is illustrated in Fig. 3.6. The hypotheses aim to explain the link weights in the network of co-editing similarities, which correspond to the obtained z_{ij}-scores. The transition probability matrices are square with dimensions $N = 110$, corresponding to the number of language editions studied. The diagonal is empty, since self-loops are not allowed. The formulae, the definitions, and data sources for hypotheses formulation are summarised for reference in Table 3.5. Below I give more extended explanations on the process of hypotheses construction.

- **H0: Uniform**

  All language co-occurrences are possible with the same probability. A concept can be randomly covered by any language edition. The transition probability t_{ij} for all permutations of languages i and j is

  $$t_{ij} = 1.$$

- **H1: Shared language family**

  I retrieve the whole family tree profile of each language and count the number of branches overlapping between each language dyad. For example,

  - Arabic: *Afro-Asiatic; Semitic; Central Semitic; Arabic languages; Arabic*
  - Hebrew: *Afro-Asiatic; Semitic; Central Semitic; Northwest Semitic; Canaanite; Hebrew*

  Arabic and Hebrew share three levels of language tree hierarchy (*Afro-Asiatic; Semitic; Central Semitic*) and thus will have the transition score of 3 in the hypothesis table. The data on language family classification comes from the infoboxes of the corresponding Wikipedia articles in the English edition, for example, the article 'Hebrew language'. This is retrieved separately for each of 110 studied languages. If f_i is the set of branches describing the full language family profile of language i, the transition probability t_{ij} corresponds to the count of shared branches in the family tree of languages i and j, and is computed as

  $$t_{ij} = |f_i \cup f_j|.$$

  Thus, the more closely related two languages are, the more likely it is for their speakers to share interest in co-editing the same topics.

- **H2: Bilingual population within a country**

  To formalise other hypotheses, I needed to map languages to countries where they are spoken. The data on language usage comes from CLDR Charts (2015). I list all countries where a pair of languages are co-spoken; for each country computing the probability of a person to speak both languages. The hypothesis table contains the average probability of a person to speak both languages computed across all countries where both languages are spoken by more than 0.1% of the population. The transition probability is described by

  $$t_{ij} = \frac{1}{N_{ij}} \sum_A p(i)_A p(j)_A,$$

  where $p(i)_A$, $p(j)_A$ are proportions of speakers of languages i, j in a country A, N_{ij} is the number of countries where i,j are co-spoken. The more bilinguals speaking i and j live in the same country, the higher the transition belief.

- **H3: Geographical proximity of language speakers**

  I assign each country to its primary language (the language that the majority of its population speaks) and compute the average distance between all permutations of countries where language i or j are spoken. All inter-country distances are scaled between 0 and 1. Thus,

  $$t_{ij} = \frac{1}{N_{ij}} \sum_{A,B} \frac{d_{\min}}{d_{AB}},$$

  where N_{ij} is the number of country permutations where i or j are spoken as primary language, d_{AB} is Euclidean distance between each pair of countries, and $d_{\min}$ is the smallest distance between countries in the dataset. The smaller the distance between speakers of i and j living in separate countries, the higher the chances for languages i, j to cover the same concept. The data on distances between countries is taken from CIA (2015).

- **H4: Gravity law – demographic force attracting language communities**

  Like in the previous hypothesis, I allow one (primary) language per country and consider all country permutations where languages i or j are spoken. Demographic attraction is

strongest between large population of speakers who live in separate counties which are located closely. Consider the example of France and Germany, where large numbers of French and German speakers correspondingly, live at close distance. I compute average demographic attraction between all permutations of country pairs. I define

$$t_{ij} = \frac{1}{N_{ij}} \sum_{A,B} \frac{m_{A,i} m_{B,j}}{d_{AB}^2},$$

where $m_{A,i}$, number of speakers of the primary language i in a country A, d_{AB} is Euclidean distance between each pair of counties (in kilometers), N_{ij} is the number of country pairs where i or j are spoken as primary language. The larger the language-speaking population and the smaller the distance between the countries A, B, the more the attraction between i and j.

- **H5: Shared primary religion**

  For each country I identify its primary language and its most widespread religion (from the following list: *Christian, Muslim, Hindu, Buddhist, Folk, other* or *unaffiliated*). The data on world religions was taken from the most recent 2010 Report on Religious Diversity provided by the Pew Research Center (Pew Research Center, 2010). The religion I assign to a language is the most common religion in the list of countries where the language is spoken as primary. For a language pair, if they share the religion, I add 1 to the hypothesis matrix, and 0 otherwise. Thus, the hypothesis formalises the intuition that the linguistic communities which profess the same religion show consistent interest in the same topics.

Bayesian inference – HypTrails

In order to explain why certain languages form communities of shared interest, it is necessary to explain the link weights, or z_{ij}-score values. At first, I formulate multiple hypotheses based on real-world statistical data, and compare their plausibility using HypTrails (Singer et al., 2015), a Bayesian approach based on Markov chain processes. I input the z_{ij}-scores into a matrix, and express hypotheses about their values via Dirichlet priors – matrices of transition probabilities between each possible state (in this case – language edition). I use the trial roulette method to compare different hypothesis. This approach allows to visualise how plausibility of the hypotheses changes with the increasing belief and decreasing allowed variation. Although it was initially designed to compare hypotheses about human trails, in this research I show for the first time that HypTrails approach is also useful in explaining link weights in networks. The Hyptrails algorithm does not output the absolute values for plausibility of hypotheses, but only compares them one to another. Thus, one must always compare the hypotheses to a uniform hypothesis, and discard those hypotheses that are ranked below the uniform. For the upper bound of comparison, I use the z_{ij}-scores data itself, since no hypothesis can explain the data better than the data itself.

Data preparation. Using the formalisations detailed in Table 3.5, I fill out corresponding transition probabilities matrices. I apply Laplacian smoothing of weight 1 to all matrices to avoid sparsity issues and to account for the cases when editions co-edit a topic of a general encyclopedic importance which might be relevant for multiple language communities. All matrices are normalised row-wise; diagonals are zero as no self-loops are allowed.

Results. Fig. 3.7 summarises the results of the HypTrails algorithm. All hypotheses are compared against the uniform hypotheses of random co-occurrence. The results suggest that multiple factors play role in how shared interests are shaped, including geographical proximity, population attraction, shared religion, and especially strongly, linguistic relatedness of the languages and the number of bilingual speakers. No hypothesis explains perfectly all variations in the data, however the Bayes Factors for all pairs of hypotheses are decisive. Geographical proximity only explains

Hypothesis and Formalisation	Notation	Description	Data Source
H0: Uniform hypothesis $t_{ij} = 1$	–	All co-occurrences are equally probable, i.e. every edition i covers the same concept as edition j with a constant probability.	–
H1: Shared language family $t_{ij} = \lvert f_i \cup f_j \rvert$	f_i is the set of branches describing the full language family profile of language i, t_{ij} is the count of shared branches in the family tree of i and j.	Language communities of linguistically related languages will show more co-editing similarity.	The data on language family classification was taken from English Wikipedia infoboxes of articles on each of 110 languages, such as 'Hebrew language'.
H2: Bilingual population within a country $t_{ij} = \frac{1}{N_{ij}} \sum_A p(i)_A p(j)_A$	$p(i)_A, p(j)_A$ are proportions of speakers of i, j in a country A, N_{ij} is the number of countries where i,j are co-spoken.	Multilingual editors belong to multiple cultural communities and might serve as bridges between them. The more bilinguals speaking i and j live in the same country, the higher the transition belief.	Territory–language information was downloaded from (CLDR Charts, 2015), and is based on the data from the World Bank, Ethnologue, FactBook, and other sources, including per-country census data.
H3: Geographical proximity of languages $t_{ij} = \frac{1}{N_{ij}} \sum_{A,B} \frac{d_{\min}}{d_{AB}}$	N_{ij} is the number of country permutations where i or j are spoken as primary language, d_{AB} is Euclidean distance between each pair of countries, and $d_{\min}$ is the smallest distance between countries in the dataset.	The smaller the distance between speakers of i and j living in separate countries, the higher the chances for languages i, j to cover the same concept. We consider one (primary) language per country.	Distance between countries is computed as Euclidean distance in kilometers between country capitals (CIA, 2015).
H4: Gravity law – demographic force attracting language communities $t_{ij} = \frac{1}{N_{ij}} \sum_{A,B} \frac{m_{A,i} m_{B,j}}{d_{AB}^2}$	$m_{A,i}$, number of speakers of the primary language i in a country A, d_{AB} is Euclidean distance between each pair of counties, N_{ij} is the number of country pairs where i or j are spoken as primary language.	The larger the language-speaking population and the smaller the distance between the countries A, B, the more the attraction between i and j. Based on the countries' primary languages.	Country population data is taken from CIA Factbook (CIA, 2015).
H5: Shared religion $t_{ij} = \begin{cases} 1, & \text{if } r_i = r_j \\ 0 & \text{otherwise} \end{cases}$	r_i is the dominating religion of a language community. It is defined as the most common religion in the list of countries whose primary language is i.	Cultures which profess the same religion will show consistent interest in the same topics.	The data on world religions was taken from the most recent 2010 Report on Religious Diversity provided by the Pew Research Center (Pew Research Center, 2010).

Table 3.5: Formalisation of hypotheses to explain the probability of language dyads to co-edit a Wikipedia article about the same concept. The hypotheses aim to explain the values of link weights (z_{ij}-scores) in the network of co-editing similarity (see Fig.3.5). The transition probability matrices are square with dimensions $N = 110$, corresponding to the number of language editions studied. The diagonal is empty, since self-loops are not allowed. The value t_{ij} expresses the hypothesised probability of Wikipedia language editions i and j to cover the same concept. After construction of the hypotheses matrices, the matrices undergo Laplacian smoothing of weight 1 (for HypTrails hypotheses testing only), and are further normalised row-wise. The process is illustrated in Fig.3.6. The results of hypothesis testing are represented in Fig.3.7 for the HypTrails approach, and in Fig.3.6 for the MRQAP approach, and are discussed in sections 3.3.2 and 3.3.2 correspondingly.

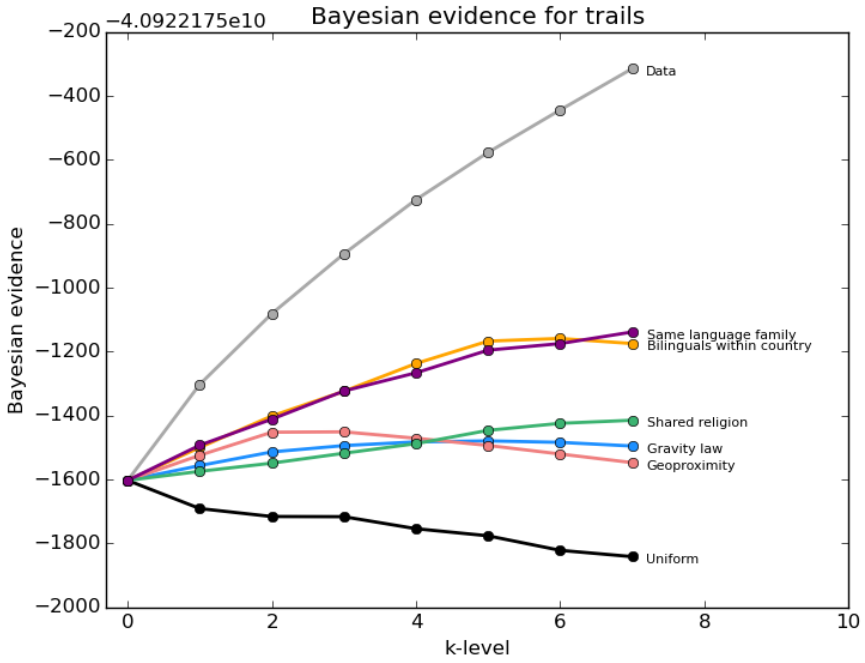


Figure 3.7: HypTrails-computed Bayesian evidence for hypotheses plausibility on shared editing interest Wikipedia data. Higher values of the Bayesian evidence denote that a hypothesis fits the data well. The bottom black line represents the hypothesis of random shared interests and the top grey line is the fit of data on itself – together forming an upper and lower limit for fitting hypothesis. The ranking of hypotheses should be compared for the same k. All hypotheses are significant, but the most plausible ones to explain cultural proximity are the shared language family, the bilingual, the shared religion, and the gravity law hypotheses. The results show that cultural factors such as language and religion play a larger role in explaining Wikipedia co-editing than geographical factors.

the data to a limited extent, and decays for higher values of belief k, while the number of bilinguals in the same country, shared language family, and shared religion hypotheses grow stronger with more belief, which suggests that they explain the data most robustly. The explanatory power of hypotheses should be compared for the same values of k, which expresses how strongly we believe in the hypotheses and how much variation is allowed.

Frequentist approach – MRQAP

In addition to the HypTrails analysis, I use MRQAP (Hubert and Schultz, 1976) to assess statistical significance of association between the concept co-editing network ties and various hypotheses. This method has a long established tradition in social network analysis as a way to sift out spuriously observed correlations (Dekker et al., 2003), and is well-suited for analysing dyadic data where observations are autocorrelated if they are in the same row or column (Krackardt, 1987). I treat the network of concept co-editing as a dependent variable matrix; the independent variable contains the set of hypotheses about the configuration of the network, expressed via hypothesis matrices. Formulation of hypotheses is given in Table 3.5. I normalise the matrices row-wise in order to standardise the values across matrices. MRQAP is a nonparametric test – it permutes the dependent variables to account for dyadic inter-dependencies. It is also robust against various underlying data distributions (Dekker et al., 2007). I used 1,000 permutations, which usually suffices for the procedure (Jackson and Somers, 1989).

Results. The results of the MRQAP are reported in Table 3.6. Different models include variations of hypotheses combinations that explain the variation in language co-editing ties. The results of the test are in agreement with the hypothesis ranking obtained from applying HypTrails. The number of bilinguals, shared language family, shared religion and demographic attraction are the factors significantly contributing to cultural similarity, as suggested by the t-statistic. By including all five hypotheses into Model 1, it is possible to explain 15% of variation in the data. Geographical distance, although a significant factor in several models, is not a very strong one: after excluding the distance hypothesis (Model 2), precision does not decrease. Excluding other hypotheses one by one (Models 3, 4, 5 and 6) lowers precision considerably. Finally, shared language family and bilinguals alone (Models 21 and 22) explain 5% and 7% variation in shared interests correspondingly.

	Model	Bilinguals	Lang. family	Religion	Gravity	Distance†	R^2 adj.	F-stat.	dF	Intercept
1	Estimate	0.0688	0.1074	0.0900	0.0470	-0.0042^*	**0.1458**	410.3	11984	0.0066
	t-statistic	**27.6524**	**23.6158**	**13.4772**	**10.2732**	$\mathbf{-1.3422}^*$				
2	Estimate	0.0676	0.1075	0.0894	0.0464	–	0.1458	512.4	11985	0.0067
	t-statistic	29.1517	23.6428	13.4200	10.1893	–				
3	Estimate	0.0703	0.1129	0.1022	–	-0.0009^*	0.1384	482.3	11985	0.0067
	t-statistic	28.1932	24.8853	15.4831	–	-0.2989^*				
4	Estimate	0.0685	0.1080	–	0.0581	-0.0016^*	0.1329	460.5	11985	0.0074
	t-statistic	27.3119	23.5817	–	12.7773	-0.5225^*				
5	Estimate	0.0716	–	0.0916	0.0598	-0.0055^*	0.1061	356.9	11985	0.0075
	t-statistic	28.1697	–	13.4180	12.8396	-1.7256^*				
6	Estimate	–	0.1134	0.0881	0.0546	0.0272	0.09140	302.5	11985	0.0070
	t-statistic	–	24.2095	12.7958	11.5815	9.0453				
7	Estimate	0.0700	0.1129	0.1020	–	–	0.1386	643.1	11986	0.0067
	t-statistic	30.2487	24.8885	15.5098	–	–				
8	Estimate	0.0703	0.1151	–	–	0.0030^*	0.1212	552.2	11986	0.0076
	t-statistic	27.9237	25.1460	–	–	0.9388^*				
9	Estimate	–	–	0.0898	0.0684	0.0272	0.0470	198.2	11986	0.0079
	t-statistic	–	–	12.7323	14.2619	8.8191				
10	Estimate	0.0700	–	0.0909	0.0590	–	0.1060	474.8	11986	0.0075
	t-statistic	29.5521	–	13.3370	12.7297	–				
11	Estimate	–	0.1140	–	0.0654	0.0296	0.0790	344.0	11986	0.0077
	t-statistic	–	24.1755	–	13.9808	9.7791				
12	Estimate	0.0712	0.1151	–	–	–	0.1212	827.8	11987	0.0076
	t-statistic	30.4703	25.1430	–	–	–				
13	Estimate	0.0738	–	–	–	0.0027	0.0749	486.5	11987	0.0085
	t-statistic	28.6184	–	–	–	0.8295				
14	Estimate	–	–	–	0.0794	0.0296	0.0342	213.4	11987	0.0086
	t-statistic	–	–	–	16.7162	9.5508				
15	Estimate	0.0733	–	0.1072	–	–	0.0940	622.8	11987	0.0076
	t-statistic	30.9368	–	15.9020	–	–				
16	Estimate	–	0.1222	–	–	0.0357	0.0641	411.6	11987	0.0080
	t-statistic	–	25.9063	–	–	11.8512				
17	Estimate	–	–	0.0936	0.0741	–	0.0409	256.8	11987	0.0080
	t-statistic	–	–	13.2534	15.5280	–				
18	Estimate	–	–	–	–	0.0372	0.0118	144.1	11988	0.0090
	t-statistic	–	–	–	–	12.0025				
19	Estimate	–	–	–	0.0861	–	0.0269	333.1	11988	0.0087
	t-statistic	–	–	–	18.2514	–				
20	Estimate	–	–	0.1144	–	–	0.0217	267.1	11988	0.0081
	t-statistic	–	–	16.3447	–	–				
21	Estimate	–	0.1233	–	–	–	0.0532	674.9	11988	0.0081
	t-statistic	–	25.9798	–	–	–				
22	Estimate	0.0746	–	–	–	–	0.0749	972.2	11988	0.0085
	t-statistic	31.1808	–	–	–	–				

† Geographical proximity of language speakers based on countries' primary languages

Table 3.6: MRQAP decomposition of pairwise correspondence between concept co-occurrence and cultural factors. The combination of all hypotheses explains most of the variation in the data (15%). The most plausible explanations are the number of bilinguals and shared religion. The results of MRQAP agree with the ranking of hypotheses by the HypTrails algorithm summarised in Fig. 3.7. All statistics except those labelled with * are significant at the 0.05 level.

3.4 Discussion of empirical results

Culture is a very complex concept, and its very definition is constantly debated by Anthropologists, Social Scientists, and Linguists, among other scientists. Although it is universally agreed that cultural communities exist, their borders are very fuzzy and depend on how the researcher defines the term "culture". In this chapter, I have focused on two such possible definitions of culture. In Section 3.2, I have examined the (1) the geopolitical cultural communities, defined through national borders; and in Section 3.3 I have looked at (2) the relation between language and culture. Both studies have a particular focus on how similarities between cultural communities of Wikipedia editors can be distilled by analysing large-scale co-editing patterns across multilingual editions of Wikipedia.

The current analysis shows that the decision of Wikipedians to write or not to edit an article on a certain topic is not a random one. The statistical analysis of Wikipedia co-editing patterns reveals that the interests of both national and linguistic communities are not universal. Rather, they are constrained to clusters of countries and languages which are more likely to share the interests with other members of the cluster than with the rest of the shared interest network. In particular, the studies presented in this chapter have found 18 national (Section 3.2) and 21 language related (Section 3.3) clusters of shared interests. This finding is similar to the idea of national cultural repertoires in the traditional Cultural Sociology (Lamont and Thévenot, 2000), which in the context of Wikipedia implies that national communities apply different grammars of worth and criteria of evaluation when selecting the topics to document. Additionally, I find a similar pattern when studying how various linguistic communities co-edit Wikipedia articles. The idea that linguistic communities of Wikipedians self-select to contribute to those articles which would appeal to the common interest of the language community, has been reflected in other studies too. In particular, emphasising that each language edition represents a community of shared understanding with unique linguistic point of view (Bao et al., 2012; Massa and Scrinzi, 2011, 2012), its own controversial topics (Yasseri et al., 2014), and concept coverage (Callahan and Herring, 2011).

The structure of the global network of shared information interests can be explained by a variety of socio-economic and historical factors. In case of the geopolitically defined communities, the most significant factors are shared language and religion, along with geographical proximity, the presence of the past colonial ties, and current trade flows. When it comes to linguistic communities, language-related factors (such as the number of bilinguals and linguistic similarity of the languages) and religion, continue to offer the strongest explanatory power, followed by demographic attraction and geoproximity of speakers.

The link between common language and shared information interests is the strongest, and it is not surprising. Language is a fundamental part of identity, self-recognition, and culture (Bloomfield, 1945; Castells, 2011a; Kramsch, 1998; Whorf, 1940). Moreover, it is well known that interests are formed by cultural expression and public opinion, and language is an important platform for these expressions (Usunier and Lee, 2005). It is hard to separate the effects of the number of bilinguals and shared language family from one another, since both might be related: shared vocabulary and grammatical features of the languages from the same language family might explain higher level of bilingualism for these language dyads. Moreover, language choice and bilingualism are an effect of factors galore, such as post-colonial history, education, language and human right policies, free travel, and migration due to political instability, poverty, religious persecutions or work (Crystal, 2000; Rassool, 1998). Finally, cultural similarity defined through Hofstede's four dimensions of values (Hofstede, 1980) has also been found to relate to language (Pfeil et al., 2006; West and Graham, 2004).

Shared religion is another uniting factor for shared interest between communities. This chapter demonstrates that similarity in bilateral and cross-lingual information interests reveal the patterns that echo religious "fault lines". This finding is in line with Huntington's thesis which argues that

cultural and religious identities of people form the primary source of potential conflict in the post-Cold War era (Huntington, 1993). Similar results were found in other studies that analyzed Twitter and email communication worldwide (State et al., 2015).

Population attraction and geographical proximity are the uniting factors that have been extensively discussed in the literature, most relevantly in the context of mobile communication flows (Krings et al., 2009) and migration (Simini et al., 2012). Similar to my results, several studies report gravity laws in online settings, including (Backstrom et al., 2010). Interestingly, not only choice of topics to edit, but also online trade in taste-dependent products is affected by distance. For example, (Blum and Goldfarb, 2006) finds that proximate countries show more similarity in taste. Notably, this effect only holds for culture-related products such as music. This further supports my finding that there is a relationship between geographical distance and culture, and allows to speculate that the Internet fails to defy the law of gravity.

In other words, globalization of technology does not bring globalization of the information and interests. Language, religion, geographical proximity, population attraction, historical background, and trade are potential driving factors that unite or polarize the information interests. These results coincide with earlier works that highlight the impact of the colonization, immigration, economics, and politics on the cultural similarities and diversities (Bleich, 2005; Castells, 2011a; Feldman-Bianco, 2001; Gelfand et al., 2011; Hennemann et al., 2012; Risse, 2001; Tägil, 1995).

Finally, when it comes to globalisation, the question of whether English is becoming the world's *lingua franca* is an intriguing one (Crystal, 2003). Its central, influential position in the global language network has been reported in networks of book translations, multilingual Twitter users, and Wikipedia editors (Hale, 2014a,b; Ronen et al., 2014). On the one hand, such high visibility allows information to radiate between the more connected languages. On the other hand, Section 3.3 shows that global language centrality plays a minor role in shared interests. Moreover, it shows that the domination of English disappears in the network of co-editing similarities, and instead local interconnections come to the forefront, rooting in shared language, similar linguistic characteristics, religion, and demographic proximity. A similar effect has been observed in international markets, where economic competitiveness is linked to the ability to speak a local *lingua franca*, rather than English (Bel Habib, 2011).

3.5 Limitations

The studies presented in this chapter are not free of limitations, some of which are inherent to the nature of the chosen data, while others remind the reader about the limitations of selected setup. Below I list the limitations which are relevant to both studies presented in this chapter, grouped by type.

Cultural communities. Language and location are both inseparable parts of culture, however they represent only a few factors that influence cultural expression. Differences in shared interests and information exchange might also be found in communities related to socio-economic status and upbringing, age, gender, religious standing, secure access to resources, political situation, etc. Thus, more research is needed to explore other dimensions of the shared interest landscape.

Study of culture via Wikipedia: Although there is a mounting evidence in the literature that Wikipedia is a promising and rich data source for those interested in mining cultural relations, I highlight that it is only one of many possible media where culture might find reflection. More studies are needed to explore how other aspects of culture manifest themselves in off- and online worlds, and what difference they make in the global ties of shared interest.

Wikipedia's non-representativeness. Wikipedia itself is not free from structural biases, as it reflects the activity of selected technology-savvy, mostly white and male (Antin et al., 2011; Hill BM, 2013), educated, and economically stable social elites. It by no means is representative of the views of general population. However, it is the elites that often drive the cultural, political, and economic processes (Ronen et al., 2014), and thus Wikipedia editors represent a group worthy of being studied.

Inhomogeneous data. Both of the datasets contain data from the language editions at different growth stages and levels of topical saturation. For example, the proportions of editing data from the largest language editions such as English, Swedish and German are not unexpectedly high. Although this might introduce unforeseen biases, it is likely not a major limitation, since the focus is on aggregated editing activity over a very long period of time, and the presented Model of shared interests accounts for the differences in size and activity levels.

Missing inter-language links. Relying on Wikipedia's inter-language link graph to identify articles on the same concepts in different language editions has some known issues, such as the lack of triadic closure and dyadic reciprocity (Bao et al., 2012). While I extract all inter-language links that exist in the Wikipedia database (see Sections 3.2.1 and 3.3.1 for details), it is possible that some of the links that should exist are missing. As an additional note, my definition of a unique concept[1] is dictated by the database organisation, and allows multiple unique Wikipedia concept instances to exist which cover a similar topic.

Geolocating Wikipedia edits. For the study of geopolitical communities, I limited the analysis to the edits from unregistered editors, because the data on the location for most of the registered Wikipedia editors is unavailable. The exclusion of registered editors is admittedly a drawback. Nevertheless, the resulting dataset contains more than one million (1,069,746) Wikipedia articles and more than 23 million (23,555,117) edits in total. We use these edits as proxies for information interests. This sample is large enough for the purpose of spatial analysis, although it admittedly contains unknown biases and is not representative of activity profile of an average Wikipedia editor.

Hypotheses formulation. While the presented approach is quantitative, it requires some subjectivity in interpreting the clusters and formulating hypotheses. To strengthen the internal validity of the study, the reasoning about the hypotheses is informed by both visual analysis of the clusters and by the previous literature on the subject. Still, I do not claim to have exhausted all possible hypotheses which could explain the data. Moreover, other formalisations of the selected hypotheses might render non-identical results.

Methodological opportunities. The presented approach focuses on the aggregated activity of multiple communities of Wikipedia editors. I leave for future research the interesting task of

incorporating the time dimension in the analysis and examining how interests shape and change over time.

One of the benefits of the presented model of shared interests is that it is free of biases related to topic selection, since it avoids focusing on specific kinds of topics where cultural similarities might be expected. The presented approach to quantifying and understanding large-scale information interests scales well in terms of the number of communities and hypotheses that could be analysed. In case of research on multilingual data, an important benefit of this approach is that it only uses metadata on user interactions, and understanding the language itself is not required. Finally, it is applicable for any example of collaborative production of a common good where individual activity of participants is recorded.

3.6 Chapter summary

In this section, I summarise the results of two case studies presented this chapter presents, and outline the main characteristics of the approach described in Section 3.1.

Approach. Colleagues and I develop a statistical filtering model (see Section 3.1) which extracts a large-scale network of information interests from Wikipedia editing data. This model is shown to successfully identify cultural communities with significantly similar information interests, as well as to quantify the strength of this similarity. This statistical filtering can be generalized to other datasets where the significance of interconnections between system entities is not apparent. This model is distinct from other proposed approaches to evaluating significant ties of correlation in bipartite networks. It is able to bring out significant connections regardless of the communities size and activity levels. Thus, it is especially helpful in studying systems where less active minority groups are present. This model is a part of larger approach which also includes network clustering, and hypothesis testing.

To validate the approach, I apply it to two large multilingual datasets of Wikipedia editing data. The resulting studies operationalise cultural communities through Wikipedia editors' geographical location and the language of contribution.

Validation I: Similarity of bilateral information interests. First empirical study (Section 3.2) examines shared interests of geopolitical communities. It focuses on contributions by unregistered editors, and maps their IP addresses to a specific country. To find the global interest connections, the study analyses how often editors from different countries co-edit articles on Wikipedia. To this end, it examines all edits per country. These edits approximate national information interest profiles. Statistically significant ties of co-edits represent significant connections in a global network.

This analysis provides a world map of shared information interests. Structural analysis of the underlying network shows that information interests are indeed not homogeneous. They split into 18 strongly interconnected country clusters. The results show through regression analysis that this division is driven by factors related to language, religion, trade volumes, geographical proximity, and historical background, such as colonisation past. While technological advances in principle have made it possible to communicate with everyone in the world, current information interests of countries are still constrained by sociocultural and political borders, as well as economic factors.

Validation II: Interest similarity among linguistic communities. Second study (Section 3.3) provides further empirical insights. It explores in detail *interest profiles of linguistic communities*. In particular, it focuses on how different cultural communities select and document their cumulative knowledge in different language editions of Wikipedia. It presents a network of global interconnections between 110 language communities, based on co-editing activity of Wikipedia editors. It also shows how to turn intuitions provided by the network, in order to form and test hypotheses about the mechanisms that explain the observed network architecture. Finally, the statistical robustness of the results is tested with various statistical techniques.

This analysis elicits that linguistic communities of shared interest are polarised into 24 linguistic clusters. Secondly, it demonstrates statistically that the observed network structure is partially explained by several sociocultural factors. These include shared religion, bilinguality, linguistic and geographical proximity of languages, and population attraction. And finally, it shows that the set of language editions covering a concept on Wikipedia is not a random choice.

3.7 Conclusions and implications

This chapter achieves several goals:

- elaborating a computational approach for extracting the ties of significant shared interest;
- validating the approach on two large datasets with multiple communities;
- answering several empirical questions on the intersection of culture, UGC, and collaborative knowledge archiving.

Two empirical studies investigate geopolitical (Section 3.2) and linguistic (Section 3.3) approximations of cultural communities. By considering these two perspectives on cultural communities, this chapter is able to arrive at wide-ranging conclusions about the forces that shape the highways of shared information interests across the globe.

In particular, this chapter sheds light on how culture is reflected in the collective process of archiving knowledge on Wikipedia. In short: we don't live in a global world. Although the means of exchanges are becoming global, what brings us closer together is the information and interests that we share. While technological advances, in principle, have made it possible to communicate with anyone in the world, various factors limit us from doing so. Instead, highways and barriers of information exchange are formed by social and economic factors connected to shared interests. Information interests remain diverse, despite globalization.

These results extend the existing literature on cultural contextualisation of UGC, and Wikipedia in particular. Wikipedia holds an important position not only for information-seeking individuals, but also as a brains of many contemporary computing frameworks and algorithms. This means, that these results have wide-ranging **implications**.

First of all, these studies demonstrate cultural richness of Wikipedia data. They also raise awareness of inter-lingual differences in coverage and attention levels to various concepts. What does this mean? On the optimistic side, these results demonstrate potential for algorithm designers to make use of these differences, and design culturally personalised, more relevant user experiences online. At the same time, they raise challenging questions about the sustainability of Wikipedia data re-use by other systems. In particular, this is interesting because in many cases these data come from the English language edition only.

These concerns may be especially relevant for the Wikimedia Foundation, which currently supports populating peripheral language editions with automatically created or translated content. This policy might be problematic, because the choice of topics to cover on Wikipedia is culturally contextualised. Injecting articles artificially might result in lower quality content due to the lack of interest towards editing them. Some other relevant questions are about user design: Should English Wikipedia aim at becoming an all-inclusive collection of information from other language editions? Should the decision on who and what will be remembered belong to the community of editors, however small, or to an automated algorithm?

To conclude, this work is another stepping stone for the academics wishing to study culture via the Web. The statistical model summarised in this chapter can be useful for computational practitioners and researchers who might wish to benefit from it. Empirical results might provoke critical thoughts among managers, economists and politicians working in multicultural settings. Additionally, I hope that this research will inspire dialogue among enthusiastic Wikipedians on how similarities between language communities can be used to improve participation of editors speaking peripheral languages, and expand the topical coverage of smaller editions. Finally, this chapter has provided an account on quantifying cultural interests via online data, which might interest the general public curious about global, intercultural relationships.

Chapter 4

Quantifying content differences in a specific knowledge domain

In the previous chapter, I established that the set of language editions covering a Wikipedia article on a concept, is not a random choice. Instead, it seems to be connected to complex factors such as bilingualism, geographical distance, and religion of the language speakers. More generally, each Wikipedia language edition is a product of work of a distinct community of editors, and the selection of topics covered by each edition reflects the interests of this community.

In this chapter, I extend my inquiry to an even more interesting aspect. If the selection of covered encyclopedic concepts differs across language editions, what about the presented information? To which extent does the content of the articles differ across editions? Answers to these questions are not intuitive.

On the one hand, it is reasonable to expect that the presented facts should not vary across language editions, because encyclopedic knowledge should be at least to some extent universal. On the other hand, Wikipedia is a product of voluntary work by a multitude of editors, and they each have specific interests and fields of expertise. Thus, some gaps or over-/under-emphasising are possible, and to some degree, even inevitable. This line of inquiry raises further interesting *research questions* which form the core of this chapter:

- Since some topics are covered by multiple language editions, how much content do these editions borrow from each other?
- Do they present a similar view on the topic, or contradict each other? If so, which facts are omitted?

In order to answer these questions, I narrow the context down to a specific knowledge domain. In particular, I focus on how national histories of the last 1,000 years are described across language editions of Wikipedia.

Writing about history – historiography – is an interesting example to look at, since history stands at the center of all social groups. Be it individuals, groups by interest (e.g. clubs, communities), or nations – establishing a consensus on historical background provides a feeling of roots and belonging. It is at the core of building identities. Different communities form different perspectives on the past. And these disagreements eventually find reflection in the collective documenting of historical accounts. For example, various communities might highlight distinct historical periods, events, and persona as important.

Wikipedia is a perfect example of an environment where such differences across communities can be expected and furthermore, quantified. The fact that it is split into language editions allows each group of language speakers to work on their own version of encyclopedia rather independently. That involves selecting topics, resolving emerging conflicts around the content, making decisions on which facts to include (and importantly, to exclude), and finally, establishing consensus on the given historical narrative.

Some borrowing and information transfer are known across language editions due to the work of bots and multilingual editors. Still, multilingual articles are not translations of each other. Likewise, article narratives do not have to be identical across language editions. Rather, every language edition presents a unique subsystem, with its own set of editors and interests. Moreover, its articles present a point of view of the distinct community of editors who worked on them. In this chapter, I present an approach to quantifying these points of view, as well as comparing them across communities.

The rest of the chapter is structured the following way. I start by outlining the general approach to quantifying historiographical narratives (Section 4.1). The approach includes selecting an adequate unit of comparison, and defining a Null model for comparing across communities. Then I present two empirical studies which validate the described approach in different contexts. Each study is presented as a separate section which includes empirical background, details of methodology, results of the analysis, and discussion of the empirical insights.

The first empirical study (Section 4.2) outlines the general historical landscape of national histories on Wikipedia. It looks at the timelines of all UN member states in the last millennium. In particular, it focuses on the European languages perspective on historiography, and compares the descriptions of national histories across 30 language editions of Wikipedia.

The second study (Section 4.3) explores how the content of crowdsourced Wikipedia articles compares to the articles on the same topics written by selected experts. Specifically, it compares English Wikipedia articles on national histories with equivalent articles in Encyclopedia Britannica. This study validates that the proposed approach can be extended to domains other than Wikipedia. On top of that, it can be combined with methods from natural language processing, which allows to discover more fine-grained content differences between the corpora.

4.1 Approach: Quantifying and comparing historiographies

Quantifying and comparing historical narratives in a multilingual environment is a challenging task. In this section, I give a short overview of the general approach that I am proposing. I also demonstrate how to adapt it to operationalise historiography.

4.1.1 Step I: Choosing the unit of comparison

First of all, historiography is a very complex domain that needs to be simplified and re-expressed in terms of **quantifiable units of comparison**. These units should be present in, and have a comparable meaning across all studied communities, regardless of the language. Moreover, there should be a reliable (with an acceptable threshold of precision) way of extracting the data about these units from the community narratives. Before settling on a unit of comparison, one should examine its validity, i.e. If it truly measures the idea/construct that it is designed to measure. This has a direct effect on interpretability of the results. Unit's complexity will also determine the possibilities for the future comparative statistical analysis. These are some examples of how such a unit could be represented:

- mentions of temporal expressions (e.g. dates);
- mentions of named entities (geographical, biographical, historical events, etc.);
- a vector of topics or meanings (using, e.g., topic modelling, representation learning);
- a (temporal) network of relational ties (e.g, between Wikipedia article in-/outlinks, or between the mentioned historical persona).

Once each historical narrative is summarised in terms of such a quantifiable unit, these units can be compared across communities using statistical methods. Significant statistical differences will point to interesting discrepancies between the corresponding language communities, and these cases could be studied further with other methods.

In this work, to quantify and compare historical narratives on Wikipedia, I select *year mentions* as a unit of comparison across language communities. In particular, I extract all date mentions in the format of a four-digit number between 1000 and 2016 (about one millennium of human history).

4.1.2 Step II: Establishing the validity of the unit of comparison

Before commencing the analysis of the extracted data, I evaluate its reliability. For example, it is important to establish whether the extracted 4-digit numbers refer to year mentions, as opposed to numerals, e.g. population counts. This evaluation is performed separately for each of the linguistic datasets, in order to ensure that the validity of the chosen measure holds for all linguistic communities.

Evaluation procedure for the studies presented in this chapter is the following. Volunteers were asked to evaluate a curated random set of extracted 4-digit numbers, each surrounded by text fragments of 40 characters before and after the number. Each case was judged as a True positive (a date) or a False positive. The details of each particular evaluation setup are presented in Sections 4.2.2 and 4.3.2. In both studies, the evaluation has shown low error rates for each linguistic dataset and century. It indicates that the extracted numbers can be interpreted as dates and aggregated into meaningful historical timelines, which makes further analysis possible.

Although such approach to evaluation involves human input and creates a computational bottleneck for scaling the analysis up to larger datasets, it is a crucial part of the analysis. Without first establishing the validity of the selected measure, applying quantitative techniques might bring the result that are hard to interpret and have questionable empirical value.

4.1.3 Step III: Null Model for comparing historiographies

Once the validity of the unit of comparison is established, one can continue with statistical analysis. Like in many cases with empirical data, year counts in my datasets are not distributed equally. Instead, some countries and decades have a much larger total number of counts compared to the rest. This makes date mentions not directly comparable across cases.

To address this challenge, I propose a Null Model as a tool for verifying if national historical timelines display non-random patterns at certain decades. The Null Model reflects the intuition that the national timelines are constructed randomly: by sampling out of the total pool of all dates in the dataset. Thus, each decade or country is mentioned equally often on average, given their total frequency.

This random baseline of expected date mentions is not a constant across the dataset, but instead is scaled to match the total number of dates collected for each decade and country. Whenever empirical date mentions are significantly different from the random baseline, it manifests a non-random signal in the data. When empirical date mentions in some decades are significantly more frequent than expected, I refer to these time periods as **national focal points**. The decades which are mentioned significantly less frequently indicate the periods of lower interest.

The Null model itself is rather intuitive, and is essentially an urn model with replacement and without duplication. First, I create a pool M of all collected dates. Then, for each country i, I randomly draw from the pool a batch of N_i dates, where N_i is the number of collected dates related to the history of country i. Every batch is then split into decades, counting how many of the extracted dates fall within a certain decade d. I repeat the process 1,000 times. This procedure gives a distribution of date frequencies for each of the decades, which consists of 1,000 data points.

Thus way, for each decade I build a distribution of the expected number of dates, within the hypothesis of events randomly distributed in time. Further, I compare the mean of this expected distribution $E[w_i^d]$ with the empirical date count for the country in the same decade, w_i^d. This difference is finally converted into a z-score.

In sum, this procedure allows to identify for each country in which decades the number of observed dates w_i^d differs significantly from the expected number of dates in this decade. Mathematically, the z-score of country i in decade d is given by:

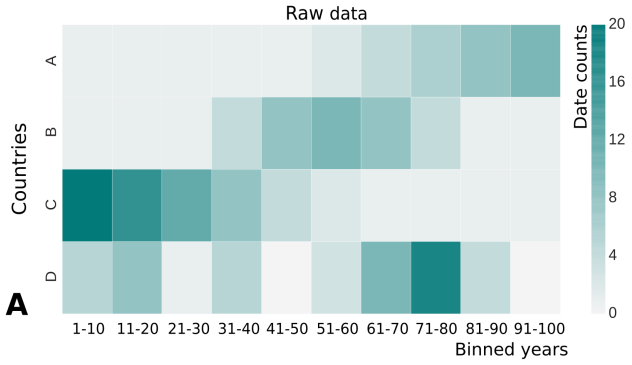
$$z_i^d = \frac{w_i^d - \mathrm{E}[w_i^d]}{\sigma_i^d}, \tag{4.1}$$

where $\mathrm{E}[w_i^d]$ is the mean of the simulated date counts in decade d across 1,000 random draws, and σ_i^d is the standard deviation of the simulated date counts.

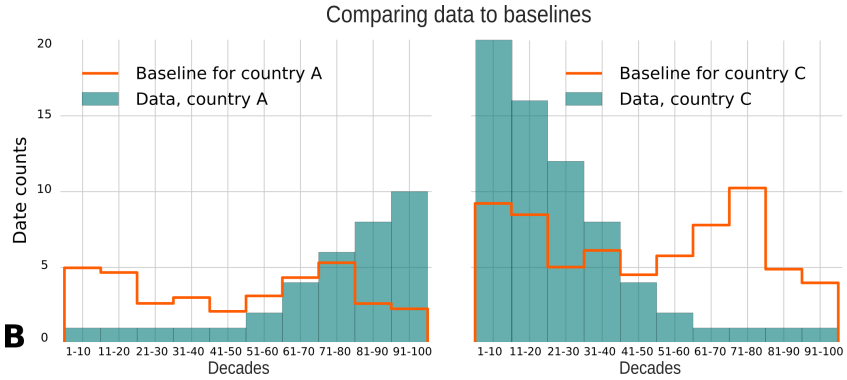
The Null hypothesis that a national timeline contains no statistically significant signal is rejected if the likelihood of the observed data under the Null hypothesis is low. However, if multiple hypotheses are tested (multiple countries and decades in this case), the chance of incorrectly rejecting the Null hypothesis increases. This is known as making a Type I error. The Bonferroni correction Dunn (1961) compensates for this by making the condition (p-value) for rejecting the Null hypothesis stricter.

In particular, for the desired significance level $\alpha = 0.01$, n_i countries and n_d decades, the Bonferroni corrected p_{Bc}-value would equal α/m, where $m = n_i + n_d$ is the number of comparisons. The z_{Bc}-score corresponds to the corrected p_{Bc}-value. This score determines the border between random fluctuations and significant signal. Evidently, all cases where z-score of country i in decade d is larger than z_{Bc} can not be explained by random fluctuations. As such, they fall under the definition of the *national focal points*.

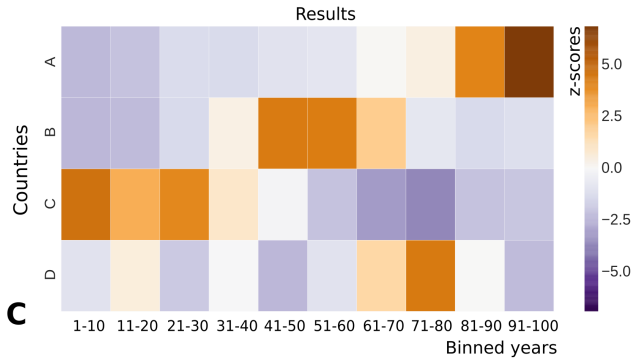
Fig. 4.1 illustrates the method with a toy example. Consider four hypothetical countries A-D with different artificial timelines (Fig. 4.1a). The dates are binned into decades, each matrix cell indicates the number of dates in the corresponding decade. Fig. 4.1b shows histograms of date



(a) Toy data



(b) Comparing toy data to the Null Model baseline



(c) Extracted focal points

Figure 4.1: Illustration of the method. Fig.4.1a shows the initial distribution of toy data for hypothetical countries A-D. The dates are binned into decades, each cell colored according to the number of dates. Fig. 4.1b illustrates the method on countries A and C. I plot a histogram of date counts for each country (green bins), and compare them with expected baselines (orange lines). The baselines are the average over four initial distributions, adjusted to match each country's total date count. Thus, the baselines are unique for every country and decade. Finally, in Fig. 4.1c I convert the differences between the data and the baseline into z-scores.

counts per country. Orange lines (baselines) correspond to the expected distributions given by the Null Model. Theseh are simply the average over four initial distributions, adjusted to match the total country date count. These baselines vary across decades and countries. I then convert the differences between the observed data and the baselines into z-scores (Fig. 4.1c). For each country I can now extract the decades in which the number of collected dates differs significantly from the expected baseline.

This method is especially useful when differences in counts are not directly comparable, as it

is in our case: all row sums in Fig. 4.1a differ. To illustrate, in Fig. 4.1a, the cell with the largest date count is country C in the first decade, but after comparing with the expected baseline (Fig. 4.1b), country A in the last decade stands out most, although its underlying count is smaller.

To summarise, this approach leaves freedom in determining the exact formulation of the expected baseline. Also, it is general enough to be adapted to a variety of settings and datasets. It is particularly useful when the datasets differ in size, and direct comparisons across them are not meaningful. It also scales well across datasets and languages. Thanks to being completely computational, it eliminates the potential bias associated with the cultural background of researchers.

In the subsequent sections of this chapter I validate this approach on two datasets. At first, I apply it to quantify public historiographies in multilingual Wikipedia. Then, I compare public and expert-written narratives on history.

4.2 Validation I: Historical landscapes of multilingual Wikipedia

This section presents an empirical study of multilingual narratives in a specific knowledge domain. Its particular focus is on how national histories are described in various language editions of the online Encyclopedia Wikipedia.

Current section is based on the results published in the Proceedings of the Eleventh international AAAI Conference on Web an Social Media, ICWSM 2017 (Samoilenko et al., 2017) and my contributions to this work are outlined in detail in Section 1.3. In order to reflect the fact that this is a collaborative work, where justified, the narrative switches to plural academic 'we'.

Today, encyclopedia Wikipedia has a vast readership across continents and languages. It offers quick, effortless access to a spectrum of reference information, including historical accounts. These representations might contain errors and false information Potthast et al. (2008), be biased towards specific viewpoints, or differ otherwise from the books written by professional historians. Fortunately, Wikipedia's open and digital nature allows for thorough quantitative analysis of historical narratives, even across a large number of languages – something which is not a typical case for other historiographical sources, such as printed encyclopedias or history textbooks. This study investigates the descriptions of national histories in different Wikipedia language editions, taking a comparative computational approach. In that direction, my co-authors and I pursue two goals: (1) presenting a data-driven approach that enables analysis of historiography through a computational lens, and (2) answering specific research questions on the depiction of history in Wikipedia.

Approach. This study is built on the computational approach to the analysis of textual historiographical data which is presented in Section 4.1.3. I apply it to Wikipedia articles on all UN member states in 30 language editions. I concentrate on *year dates* as accessible representations of more complex historical structures. To be able to compare the descriptions across languages, I retrieve from article texts all date mentions (in the form of 4-digit numbers between 1000-2016), and use them as a language-independent unit of comparison (Rüsen, 1996). Finally, I extract *significant focal points of national histories* – time periods of significantly high mentions, compared to a random expectation model. The study combines visual interpolation and expertise of invited history experts in order to evaluate how the results of the approach compare with the existing historical knowledge. It also uses hierarchical clustering to group countries whose histories are represented similarly on Wikipedia. At last, inter-language agreement on history of each country is computed using the Jensen-Shannon divergence measure.

Empirical questions. This study investigates, what readers of different languages can learn about national histories from their 'home' Wikipedia language editions. The particular focus is on three research questions:

- **RQ1:** What are the most documented periods of history of the last 1,000 years in Wikipedia?
- **RQ2:** What are the temporal focal points in descriptions of national histories in Wikipedia?
- **RQ3:** Are country timelines consistent across language editions?

Empirical findings. Colleagues and I find the presence of recency bias across all language editions and countries – the tendency to document recent events more frequently than those that happened in a more distant past. We also find that the distribution of historical focal points in the analysed articles is inhomogeneous across continents. We discover a multitude of focal points distributed through entire timelines of European countries, while we see much fewer highlights in pre-Columbian Americas and Oceania. Groups of countries with similarly distributed focal points map well to geopolitical blocs. Finally, we find differences in the way national histories are

described in the examined language editions, although on average the cross-lingual consensus is rather high.

Contributions. This empirical study contributes to the pool of computational methods that help to quantify historiographical processes. Colleagues and I combine multiple computational methods into an approach that can be used for quantitative analysis of large textual historical and historiographical datasets, such as demographic and economic records, census data, digitised books, etc. Our approach scales well, and is suited for large comparative studies of multidimensional (e.g. multiple languages and countries) data. By including 193 countries and 30 languages, we step beyond the current state of comparative historiography and allow for a large-scale transnational perspective on similarities, conjunctions, or alternatives in historiography. Although we start from the (limited) concept of the (pre-)histories of nation-states we, (i) methodologically, enable cross-lingual and -national clustering and comparison and, (ii) empirically, show that historiographical focal points transcend national borders, and contribute to existing literature on collective memory and public history as created and perceived through Wikipedia.

4.2.1 Empirical background

Our approach to quantifying historiography carries characteristics of 'the digital turn' that the study of history has envisaged in the recent years: it uses a large (digital) data set, borrows from statistical methods, and conceptually, turns to transnational and global comparative perspectives. Theoretically, our approach lies in the domain of cultural history (analysing the multitude of historical interpretations, e.g. gender-based or post-colonial histories), with a specific focus on the formation and effects of collective/public memories (Conrad, 2007), and the analysis of nations as imagined communities (Anderson, 2016).

Wikipedia as a data source: Many non-academics start seeing history as a venue for active participation, rather than a domain of professional historians (Rosenzweig and Thelen, 1998). The encyclopedia Wikipedia is open for everyone to contribute on any topic. Thanks to this feature, it has become one of the primary venues where 'free-lancer' amateur historians, potentially, alongside with professionals, can participate in history-making and shaping historiographic discourse (Conrad, 2007). A number of professional historians recognise Wikipedia as a place where enthusiasts collaboratively re-think the past (Pfister, 2011), construct public memories (Pentzold, 2009), and write history in an open source manner (Rosenzweig, 2006). Although such popular understandings of the past might differ from those of professional historians (Conrad, 2007), Wikipedia is a popular source of information when it comes to history (Spoerri, 2007) and thus has become an object of research itself.

To the best of my knowledge, only a few researchers have investigated historical narratives of Wikipedians: Luyt compared the articles on the history of two countries, concluding that the history of Singapore recounts the dominant political narrative, while the article on the history of the Philippines contains both traditional and alternative views. Jensen looked into the discussion pages about the article on the war of 1812 and found that the main debate among the editors was on who won the war. Both studies use a traditional descriptive methodology. Finally, Gieck et al. used a data science approach and compared war-related articles across 5 language editions, using methods from sentiment-, network-, and language complexity analysis. The authors found that World Wars I and ii are the most important historical events in these editions.

Quantifying history: Quantitative approaches were integrated into history studies in the last century. Opposite to traditional qualitative interpretations, they relied on statistical methods and a new conceptualisation, in which historical reality was condensed to quantifiable (often socioeconomic) historical facts, whose evolution was traced through longitudinal studies (Furet, 1971). Computational approaches that allow formulating and testing retrospective hypotheses, running historical experiments, and discovering large-scale patterns of the past by processing big data-sets, appear to be the obvious next step that could turn history into an analytical, deductive, predictive

Figure 4.2: Data collection. Parts of the article on Portuguese history and one of its outlinks, as they appear in Slovenian Wikipedia in 2016. I collect all 4-digit numbers from the main text of the article and all its outlinks, and analyse the resulting distribution (bottom part of the figure).

science (Kiser and Hechter, 1991; Turchin, 2011). Mathematical simulations have helped to test historical hypotheses about the evolution of commodity flows across ancient Asia (Malkov, 2014), the influence of agriculture on birth rates in the Old World (Bennett, 2015), and the rise and fall of large-scale societies due to the interplay of geography and military innovations (Turchin et al., 2013). Network approaches have also found popularity among historiographers, due to their possibility to visualise and quantify relational ties that are abundant in written (digitised) historical texts. For example, Sindbæk (2007) has used the text of a 9th century historical novel to build a network of mentioned geo-locations and map the travels and settlement of the Vikings. Using the text written by a historian, Padgett and Ansell (1993) coded a network of Florentine elite families and studied centralisation of political parties in Renaissance Florence. Jackson (2016) studied the ego-networks of the elites of Medieval Scotland based on the mentions of people in a large collection of historical documents. Finally, Schich et al. (2014) applied networks to track the intellectual mobility of notable individuals on a large scale.

The step from statistic to historic interpretation still remains a difficult one, which is one of the reasons quantitative approaches have been slow in gaining support among the traditional historians. Nevertheless, computational studies could contribute evidence to support the existing historical theories, or suggest otherwise unavailable new hypotheses. The recent rise of interest to digital humanities and successes in digitising large collections of historical documents (Michel et al., 2011) allow broad possibilities for historians to select the data relevant to their questions. Still, the pool of available methods remains rather sparse. As it becomes easier for historians to extract data from digitised records, as well as from digitally born sources such as Wikipedia, new methods are in need that will help quantify and map historical processes.

4.2.2 Data collection & validation

In this section I describe the steps of data collection and validation. Both data and code are available online (Samoilenko, 2017). I focus on the history of 193 countries[1] which are the current UN member states[2].

Data collection: For each of the 193 countries I locate an article in the English edition of Wikipedia, titled 'History of X', where X is the country name. Using Wikipedia's inter-language links, I retrieve other language versions of the article from sister editions. The analysis is limited to

[1]Throughout the section I use the terms nation, country, and state as synonyms, being aware of the differences.

[2]List of the UN member states, http://www.un.org/en/member-states/index.html (accessed Nov. 13, 2016)

30 largest Wikipedia editions (more than 125,000 articles[3]) providing these languages are native to Europe. By applying this setup one can avoid issues connected with extraction and alignment of dates from the languages using different calendars and alphabet systems. The limitations related to multilingual data retrieval and the choice of linguistic scope are discussed in detail in Section 4.4.

I retrieve the main text of each article from the English Wikipedia and – if available – from all 30 selected sister editions, as well as the text of all Wikipedia articles to which these pages link. I focus on the out-links because they are embedded in the main articles' texts and thus immediately available for a reader to inspect, unlike, for example, the in-links, which could not be found by reading the article page. I find between 14,927 (italy) and 394 (the Federated States of Micronesia) articles related to the history of each nation. In order to assess the coverage of historical periods, I choose a language-independent measure – the mentions of year numbers in the article text. Since this study is interested in historical events of the last millennium, I retrieve all 4-digit numbers in the range between 1000 and 2016 from the main text of all articles in our collection. Fig. 4.2 illustrates the process with an example of an article on the history of Portugal in Slovenian Wikipedia. In cases when paragraphs consist mostly of hyperlinks (more than 50% of words are hyperlinks), I record no dates from them, since there is little narrative in such paragraphs.

I ran the data collection in July 2016, using the access provided by Wikimedia Tool Labs [4] as well as the Wikipedia API [5], and retrieved approximately 17M dates from 773,121 articles in 30 language editions.

Data validation: In order to ensure internal reliability of our extraction method, I check whether the extracted numbers are years rather than numerals indicating, for example, height. For that, I create a random sample of 3,300 extracted 4-digit numbers evenly split across 30 languages and 11 centuries, and ask 3 independent human coders to evaluate each case, i.e. to say whether a number is a date or not (false positive). For each language there are 110 evaluation tasks, which consist of: the potential date (4 digits), the text surrounding the potential date in the original language (40 characters before and after the number), and the same text translated into English via Google Translate (except for the English edition case). If the coder is unsure about a number, it is treated as a false-positive. Each case is settled by the majority vote. The resulting inter-rater agreement is substantial (Fleiss' kappa = .77). I compute the expected error rates for centuries,

$$< E_c >= \frac{1}{D_c} \sum_l (\frac{n_{c,l}}{10} d_{c,l}), \tag{4.2}$$

and language editions,

$$< E_l >= \frac{1}{D_l} \sum_c (\frac{n_{c,l}}{10} d_{c,l}), \tag{4.3}$$

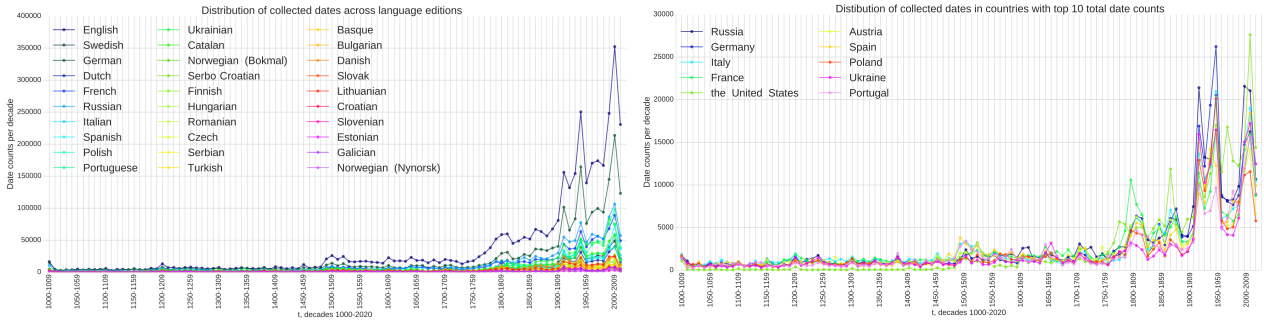
where D_l and D_c are the total count of collected potential dates per language and century, and $n_{c,l}$ is false positives count in the random sample of $d_{c,l}$ numbers collected per language edition l and century c.

The expected error rates for both centuries and language editions are reported in Table 4.1. All language editions and most of the centuries show a very low expected error rate (below .04). The highest estimated error rate is in the 11th century (.24), since a large number of extracted digits turned out to be numerals relating to heights, population counts, etc. This error is present mostly in this century, presumably due to the numeral 1000 being often used for other purposes than mentioning a year. Other false-positives include dates from Before Christ. In the more recent centuries the extraction method is very exact.

[3]Wikipedia: List of Wikipedias, http://en.wikipedia.org/wiki/List_of_Wikipedias (accessed Nov. 13, 2017)

[4]Wikimedia Tool Labs, https://wikitech.wikimedia.org/ (accessed Nov. 13, 2017)

[5]Wikipedia API for Python, https://pypi.python.org/pypi/wikipedia/ (accessed Nov. 13, 2017)



(a) Extracted timelines: 30 language editions. **(b)** Extracted timelines: selected countries

Figure 4.3: Distribution of collected dates. (a) in 30 language editions and (b) in top 10 countries according to the number of collected dates. Across editions of all sizes, and across countries I observe the same strong bias towards dates within the last 100 years, while dates from before 1500 are rarely mentioned.

Reproducibility. All retrieved pages, extracted dates, and evaluation data are stored for reproducibility purposes and could be downloaded from GESIS Datorium service (Samoilenko, 2017) or at a request to the author. Although it is also possible to retrieve this data from future Wikipedia dumps, I believe this preservation effort will save time for the researchers wishing to reproduce the study and conduct further investigations.

4.2.3 Approach and results

In this section, I outline further details of the approach and present the findings regarding interlingual portrayal of national histories in Wikipedia. The analysis consists of three parts: (1) identifying most covered historical periods across countries and languages, (2) extracting the focal points of national histories across all language editions, and (3) quantifying the amount of interlanguage agreement on representation of national histories.

Most covered historical periods

In order to gain a better understanding of the dataset, I first look into timelines of extracted dates. Figures 4.3b and 4.3a present the distribution of collected dates across all 30 language editions and for ten selected countries with the most available dates.

Results. Across language editions, the data show bias towards recent dates, having a large proportion of dates (between 60 and 80 percent) in the more recent decades (since 1800), and very low date counts before 1500. This is partly due to the chosen subject (nation states), but also points to a more general recency bias.Thus, Wikipedia readers can find a more detailed documentation

Language	Exp. error.	Language	Exp. error.	Language	Exp. error.
English	0.0067	Ukrainian	0.0351	Basque	0.0042
German	0.0186	Catalan	0.0096	Bulgarian	0.0262
Swedish	0.0109	Norw. Bokmal	0.0538	Danish	0.0086
Dutch	0.0198	Serbo-Croatian	0.0068	Slovak	0.0040
French	0.0018	Finnish	0.0208	Lithuanian	0.0266
Russian	0.0395	Hungarian	0.0347	Croatian	0.0178
Italian	0.0081	Romanian	0.0378	Slovenian	0.0025
Spanish	0.0069	Czech	0.0076	Estonian	0.0131
Polish	0.0223	Serbian	0.0119	Galician	0.0154
Portuguese	0.0166	Turkish	0.0256	Norw. Nynorsk	0.0246

Century	Exp. error.
11th century	0.2428
12th century	0.0442
13th century	0.0982
14th century	0.0214
15th century	0.0363
16th century	0.0415
17th century	0.0261
18th century	0.0089
19th century	0.0094
20th century	0.0000
21st century	0.0100

Table 4.1: Expected error rates: language editions and centuries

of historic events of the past 200 years, compared to earlier centuries. Apart from intense coverage of the most recent events (2000s), I also observe peaks of date mentions that correspond to some of the most violent recent conflicts: Napoleonic war (1800-10s) and the First (1910s) and the Second (1940s) World Wars.

Historiographic focal points of countries

For this part of the analysis I aggregate all country-related dates across language editions and apply the Null Model for comparing historiographies which is described in Section 4.1.3.

Results. The results of this procedure for selected geopolitical blocs are reported in Fig.4.4. z-scores below -6 and above 6 correspond to p-values < 0.01 (the expected distributions of decade counts are approximately normal), which means the results in all coloured cells are statistically significant. There are noticeable differences in distributions of focal points (in dark orange) across countries. For Western European countries, I observe high coverage of the Medieval and Early Modern periods (until ∼1800). Specific periods of interest for individual countries include, for example, the French Revolution in France (1780-90s) and the Third Reich in Germany (1930-40s). By contrast, in East Asia the focal points are more heterogeneous. For Mongolia, the timeline focuses on the Mongolian Empire in the 13th century. Articles on Japanese and Chinese histories exhibit a strong focus on specific small time frames: the rise of the Tokugawa shogunate (1180-90s), the Kenmu Restoration (1330s) and the beginning of the Edo period (around 1600) in Japan; and the rise of the Jin (1120s), Yuan (1270s), Ming (1360s) and Qing (1640s) dynasties in China. Only with stronger European involvement in the region (starting in the mid-19th century) there is a more steady coverage. For Central America, the timelines focus on the Age of Discovery (late 15th - early 16th centuries), and the Spanish-American Wars of independence (first half of the 19th century). in North America, the eras of the American Revolutionary War (end of 18th century) and the American Civil War (1860s) are most noticeable. For different regions of Africa, historical timelines strongly focus on the periods of its occupation and colonisation (Scramble for Africa in late 19th century), and recent history following its decolonization in the 1960s. in contrast to Southern Africa, North African national timelines focus on the Medieval history (Caliphate era), which is also the time of close interaction with Europe. The coverage seems to seize around 1300, just before the outbreak of the Black Death epidemic. For Australia and New Zealand the peaks in 1760-80s correspond to the expeditions of James Cook discovering Oceania and South Pacific. Over the next centuries, as contacts between Europeans and the local population grew, the coverage remains stable.

Overall, the number of discovered 'focal points' differs across regions. Within 30 examined Wikipedia editions, there is a disproportionate focus on histories of European countries, and the coverage of non-European states seems more intense in the periods when those states had closer interactions with Europe.

Clustering. I use the results reported in Fig.4.4 to group countries based on their historical timelines' similarity. Each country is represented as a vector of z-scores, and grouped together with the countries whose z-score values across decades are similar both in direction and intensity. I compute pairwise cosine similarity between all countries, and apply hierarchical clustering with complete linkage (Müllner, 2011) on top of the obtained values. The resulting dendrogram in Fig. 4.5 shows that the clusters correspond rather well to geopolitical regions. To illustrate this point, I cut the dendrogram at an (arbitrary) level t=.2, and plot the resulting 18 country clusters on a world map (Fig. 4.6). It suggests that focal points of the countries from the same geopolitical regions are similar in the analysed editions.

To sum up, combining information about cluster membership in Fig. 4.6 with the significant focal points presented in Fig. 4.4 facilitates a transnational impression on patterns of similarity among national timelines. One can see, for example, that most of Africa maps to one cluster. Despite individual differences between country histories, in the analysed descriptions, history of

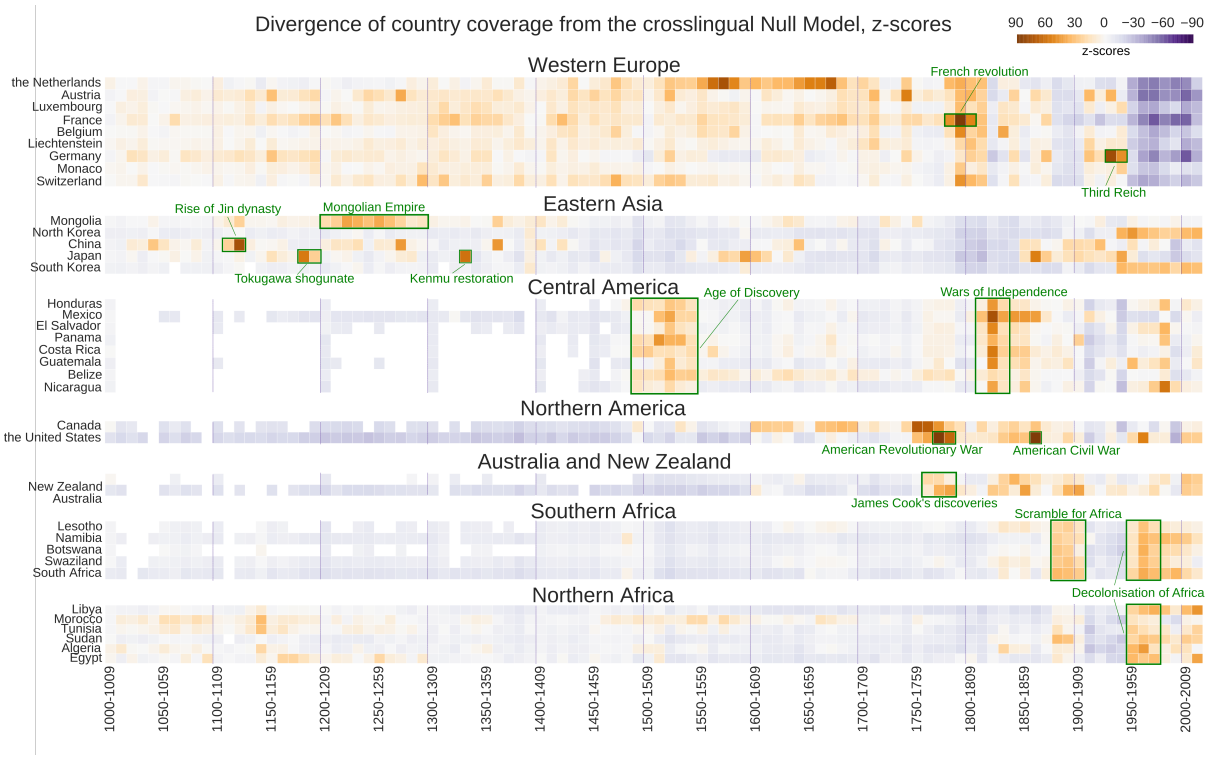


Figure 4.4: Temporal focal points of selected countries. z-scores below -6 and above 6 correspond to p-values < 0.01, which means the results in all coloured cells are statistically significant. Higher z-scores (orange) correspond to positive differences between the observed and the expected date count per decade. Cells with fewer than 30 dates are masked out. interpretation of historical events corresponding to some focal points (in green) is offered by history experts. The distributions of focal points suggests there are similarities across countries within geopolitical blocs.

the entire continent is reduced to the periods of its (de-) colonisation. Similarly, focal points of most of Central and South American countries are limited to the Age of Discovery and their Wars of independence. On the contrary, Europe is separated into several clusters, as here the differences among the individual national timelines are more distinct. This analysis gives an impression of how the entire world groups into regions based on the extracted focal points of individual countries. Also, it illustrates that for some parts of the world (e.g. Africa and parts of Americas), the analysed timelines show a reduced view of history.

Quantifying inter-edition agreement

This section investigates if national historical timelines are consistent across languages. For that, I compute a measure of their divergence across Wikipedia editions.

Method: Jensen-Shannon divergence. Based on the extracted probability distributions of years, for each country I compute a matrix of pairwise inter-language dissimilarities, using the Jensen-Shannon (J-S) divergence (Lin, 1991):

$$J(p \parallel q) = \frac{1}{2}\left[\sum_t p(t)\log_2\left(\frac{2p(t)}{p(t)+q(t)}\right) + \sum_t q(t)\log_2\left(\frac{2q(t)}{p(t)+q(t)}\right)\right], \tag{4.4}$$

where $p(t)$ and $q(t)$ refer to the probability of year t in the language editions p and q. The divergence $J(p \parallel q) \in [0, 1]$, with 0 indicating complete overlap between the compared distributions. Differences across language-specific timelines of each country are summarised by a square $m \times m$ matrix, where m is the number of extracted language editions (up to 30) covering the country's

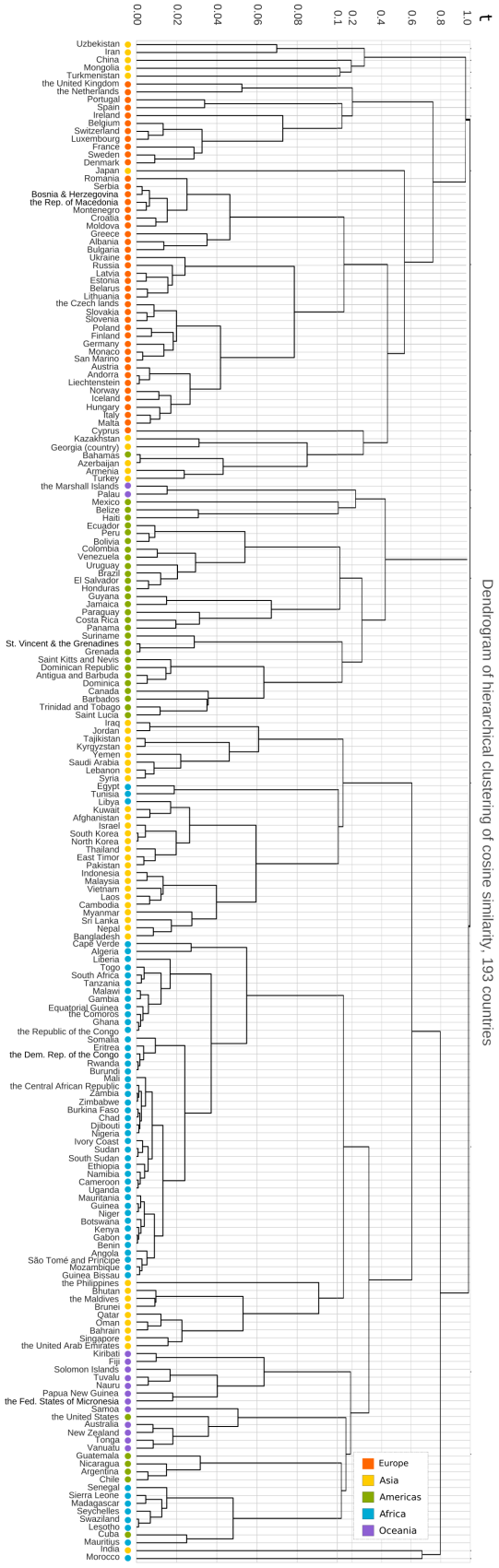


Figure 4.5: Complete dendrogram of hierarchical clustering, based on cosine similarity values for all country pairs. The clusters of similarly described countries largely correspond to geopolitical regions.

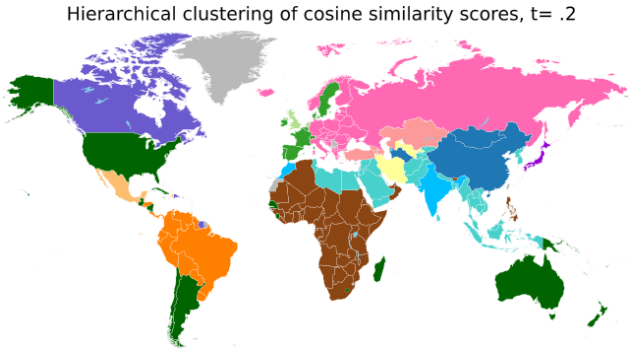


Figure 4.6: World map of country clusters. This results from cutting the dendrogram of hierarchical clustering at a threshold $t = .2$. The countries within clusters have similar temporal focal points, based on the articles in 30 analysed editions, and correspond well to geopolitical regions.

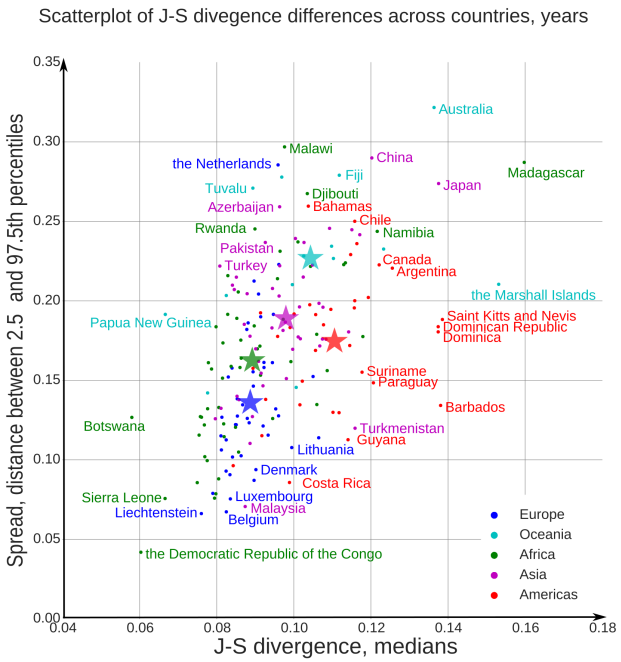


Figure 4.7: Inter-language consensus in Wikipedia articles on national histories, based on pairwise Jensen-Shannon divergence values. Countries in the lower left of the plot show the highest consensus across editions. Stars represent data centroids for countries of the same region. The plot shows a high inter-language consensus on average, though the descriptions are not identical across editions. European countries exhibit the highest amount of consensus.

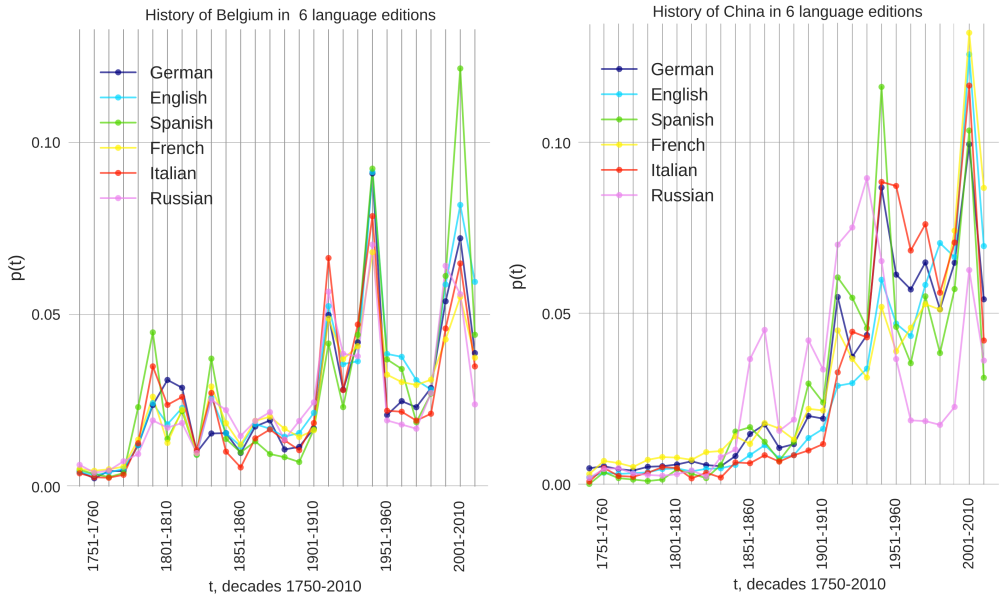
history. Interlingual differences are summarised by two values: median and spread of the distribution of $J(p \parallel q)$ values per country, which are presented in a scatterplot for ease of visual analysis.

Results. The results of this approach are presented in Fig.4.7. Data points in the lower left quarter correspond to countries with the lowest medians and the narrowest distributions of J-S scores (i.e. the smallest differences between the most similar and the most different language pair), and thus, with the highest inter-lingual consensus.

Overall, J-S scores are centred around very low values (medians between .06 and .16), which indicates a high average agreement across language editions. Their spread covers a higher range (up to .35), implying the presence of large differences between some language pairs. Based on the location of data centroids (stars), one observes higher interlingual consensus on the history of European and African countries, compared to Americas, Asia, and Oceania. The largest interlingual disagreement is found in the articles on the history of Australia, Malawi, Madagascar, China, Japan, and the Netherlands; some with the highest consensus are Liechtenstein, Belgium, the Democratic Republic of Congo, and Malaysia. In case of China (Fig. 4.8b), for example, high disagreement is partially driven by the differences between Russian and other language editions. This is especially evident during the period of Sino-Soviet split in 1960-80s, which is less densely covered in the Russian language Wikipedia. Timelines of history of Belgium, on the other hand, are almost identical across all 30 language editions (in Fig. 4.8a I present only 6 largest editions in order not to obstruct the view). Overall, this analysis suggests that country-specific historical timelines differ across language editions, although on average such differences are rather small.

4.2.4 Discussion of empirical results

This section has investigated what readers of different languages can learn about national histories from 30 Wikipedia editions. The empirical results indicate the presence of *recency bias* across



(a) Belgium: high consensus **(b)** China: low consensus

Figure 4.8: Inter-lingual consensus on histories of selected countries. To illustrate cases with very high (a) and very low (b) inter-lingual consensus, I present parts of probability distributions of dates zoomed into 1750-2010s, for 6 large editions. Chinese timeline in Russian Wikipedia differs noticeably from the timelines in all other editions, while for Belgium all timelines are almost identical.

language editions and countries: most retrieved dates belong to the recent decades, while those before 1500 are very sparse. Other studies report similar findings: a survey of students from Europe, the US, and Japan about the events that they perceived most important in the last 1,000 years showed that 60% of the mentioned events happened in the last 300 years (Rovira et al., 2006), while in our case it is between 60 and 80% depending on the language edition. Recency bias is a well-known concept in the fields of social/collective memory and psychology of history, and it is sometimes referred to as genealogical (Candau, 2005), autobiographical (Wertsch, 2002), or, most commonly, communicative memory (Assmann, 2011; Assmann and Czaplicka, 1995). The span of such memory is usually 80-100 years, or three to four generations. It embraces the memories from a recent past, to which there is an immediate connection through a living witness or a personal experience. Through uncovering the recency bias, the findings presented in this section support that Wikipedia is a public space where recent memories are actively negotiated and documented. Possibly, this is because we simply know more about the recent past. However, like (Pentzold, 2009), this study also finds evidence that these narratives stretch beyond the limited domain of communicative memory ('floating gap' of 3-4 generations), and reflect long-term, stabilised cultural memory (Assmann, 2011). While recency bias is common in oral accounts of history, it is novel to demonstrate it for the context of a written encyclopedia.

The analysis of historiographic focal points of countries indicates inhomogeneous coverage across the continents, but high similarity within geopolitical country blocs. This analysis uncovers a multitude of focal points distributed across the whole timelines of European countries, while one observes very sparse coverage and no focal points in pre-Columbian Americas and Oceania. Significant focal points in non-European states appear to relate to the periods which are culturally and historically important for Europe, such as the discovery of Latin America and the Polynesian islands by European travellers, the beginning of European trade with China, and the period of

close interaction between Europe and Northern Africa up until the Black Death epidemic. Co-authors and I interpret this as evidence of *Eurocentric bias*, an issue well-documented in professional historiography (Geyer and Bright, 1995), but also present in public perceptions of history, as cross-cultural surveys show (Liu et al., 2005; Rovira et al., 2006). This indicates that Wikipedia, despite offering a democratised way of writing about history, reiterates similar biases that are found in the 'ivory tower' of academic historiography. Given that the focus of this analysis is on languages spoken in Europe, some dominance of Eurocentric perspectives is expected. Still, some languages that I study (e.g., English and Spanish) are widely spoken in other regions, such as Latin America and Africa. Considering their international reach and the collaborative, global nature of Wikipedia, it is surprising to empirically confirm this imbalance towards European countries.

This study also finds *high consensus across the examined editions* in describing individual country histories. Across language editions, extracted dates also peak in the same decades, which correspond to periods of highly violent conflicts. Although this finding is not immediately intuitive, previous studies have reported high consensus in how different cultures view important historical events (Rovira et al., 2006). The authors explained this by the possible existence of cross-cultural collective memory, dominant hegemonic beliefs about the world history, and the narrowing cultural and interest differences between the communities. The latter has also been studied in Wikipedia context, finding that both linguistic (Samoilenko et al., 2016) and geographic communities (Karimi et al., 2015) of Wikipedia editors are interested in similar article topics. I add to this research by demonstrating that in the case of history, the content (recovered timelines) of articles is also very similar across languages. These similarities in the public perceptions of history might be a result of converging approach to history education, common exposure to media and entertainment (such as popular history TV shows), or lack of exposure to alternative historical material (Conrad, 2007). Additionally, cross-lingual Wikipedia editors and bots might be responsible for inserting similar material in different language versions of the article (Hale, 2014b), which might be a factor in the similarities I find.

4.3 Validation II: Expert vs. crowd perspectives on historiography

This section validates the approach (Section 4.1.3) on a cross-platform dataset. It also extends its scope, this time focusing not only on temporal, but also linguistic aspects of the coverage.

The content of this article is based on the results of the following publication: [(Samoilenko et al., 2018)] Samoilenko A., Lemmerich F., Zens M., Jadidi M., Génois M., Strohmaier M. (Don't) Mention the War: A Comparison of Wikipedia and Britannica Articles on National Histories. My contributions to this work are outlined in detail in Section 1.3. Like before, in order to reflect the fact that this is a collaborative work, where justified, the narrative switches to plural academic 'we'.

The Encyclopedia Britannica is an important authoritative reference on a multitude of topics and subjects. Written by *experts*, it also provides extensive information on the history of countries. With the advent of the World Wide Web and collaborative technologies, Wikipedia has emerged as a *crowdsourced alternative* to traditional encyclopedias, such as Britannica. As of 2017, Wikipedia is among the top five accessed websites globally, while Britannica has a popularity rank of 2,153[6]. Over the years, Wikipedia has also accumulated a rich body of collaboratively written articles on history which are among its top accessed Spoerri (2007) subjects. Just as awareness about history is crucial for developing a sense of national, cultural, and personal identity, understanding the differences offered by various history-related sources is important. In this section, I investigate the ways in which Wikipedia articles about national histories differ from their equivalents in Britannica – thus, taking an important first step towards comparing the views of the past offered by expert- and crowdsourced sources.

Research question: This study asks, *How do the descriptions of national histories in English Wikipedia compare to the corresponding articles in Britannica?* Particularly, I examine the temporal and topical aspects of coverage, and linguistic presentation of the material. The analysis covers the following aspects of representation of national histories in Britannica and English Wikipedia:

- What are the general patterns of temporal coverage of national histories?
- What historical periods are covered most differently across the encyclopedias?
- How do the the distributions of the national focal points compare across the encyclopedias?
- What vocabulary is most distinct for each of the encyclopedias?
- What are the linguistic differences in the presentation of articles on national histories?

Approach: This study aims to offer a first large-scale quantitative investigation of how history articles written by Britannica experts compare to those collaboratively produced by Wikipedians. Colleagues and I take a reader perspective and investigate how the national histories of all UN member states are presented in these encyclopedias. Precisely, we quantify the temporal, topical, and linguistic differences across the articles. Again, *year mentions* are selected as accessible representations of temporal coverage. I retrieve from article texts all date mentions (in the form of 4-digit numbers between 1000-1999), and use them as a unit of comparison Rüsen (1996) across the datasets. To asses temporal coverage differences, we apply the randomisation-based filtering method described in Section 4.1.3 and subsequently, statistical inference. The empirical results are validated by history experts. To compare linguistic features, I compute text statistics, apply a range of well-established readability tests, and run a Part of Speech analysis.

Findings: This study finds that Britannica and Wikipedia exhibit different approaches to historiography, where Britannica leans to a more *spatial and territorial concept of the history of states*,

[6] http://www.alexa.com/siteinfo/britannica.com and http://www.alexa.com/siteinfo/wikipedia.org (accessed 16, October 2017)

and Wikipedia – to presenting their history as *a sequence of political events*. Precisely, Wikipedia puts a disproportional emphasis on periods of conflict and war, with a preference for events well-known to the general public. In comparison, Britannica articles emphasise conflicts with underlying cultural and religious tensions. Semantically, Britannica relies on vocabulary with religious connotations and on geographical terms, while Wikipedia is heavy on political and military words. Finally, both show characteristics of English Academic prose, although Wikipedia's writing is slightly easier to comprehend.

Contributions and implications: This investigation is extensive, and the first to offer large-scale quantitative insights on how the expert-written historiography of Britannica differs from Wikipedia's popular view of the past. Colleagues and I combine computational and linguistic analyses to arrive at a comprehensive account of structure (coverage, timelines, and their focal points), content (historical reference of these focal points, semantic differences), and presentation (readability) of both encyclopedias. Our motivation is that collaborative sources like Wikipedia challenge the authority of traditional encyclopedias, both in popularity and presentation of content, and have become a global facilitator of knowledge.

The rest of the section is structured as follows. I commence by presenting an overview of related work in Section 4.3.1 and outlying the details of data collection and pre-processing in Section 4.3.2. The analysis (Section 4.3.3) is split into several parts examining each research question in detail. I wrap the section up by discussing the findings (Section 4.3.4).

4.3.1 Empirical background

This work draws on several theoretical domains. It directly relates to research on cultural history, collective/public memories (Conrad, 2007), and the analysis of nations as imagined communities (Anderson, 2016). The comparison of crowd-sourced and traditional encyclopedias is related to theoretical studies on how the digital turn and the rise of mass media culture challenge the traditional notion of expertise (Castells, 2011b; Hartelius, 2008; Pfister, 2011).

Wikipedia vs. Britannica comparisons: Comparisons between Britannica and Wikipedia have attracted substantial academic interest in the recent years. Most research has focused on verifying the quality and accuracy of Wikipedia's content by comparing it to authoritative sources. The scepticism about Wikipedia's credibility was mainly due to the new crowdsourced, self-emerging expertise that the encyclopedia draws upon, unlike peer-reviewed, expert-produced content of traditional encyclopedias (Hartelius, 2008; Pfister, 2011).

Although even earlier studies showed little difference in quality, breadth, and validity of the content between Britannica and Wikipedia (Giles, 2005), the claims of Wikipedia's credibility were met with criticism (Editorial, 2006; Magnus, 2006), and inspired a range of follow-up studies examining a range of topical domains. For example, Wikipedia articles on mental disorders (Reavley et al., 2012), military history (Jensen, 2012), and Top Fortune companies (Messner and DiStaso, 2013) have been scrutinised by the field experts, and in every case have been found at least as accurate and broad, or even more up-to-date than Britannica or other authoritative peer-reviewed sources. Other studies, however, suggest that the quality of Wikipedia articles might vary depending on the chosen field (Clauson et al., 2008), and even from article to article within one domain (Holman Rector, 2008). Most of research on Wikipedia's reliability is unfortunately based on very small samples (several articles), and can not be scaled up due to reliance on qualitative methods and field experts.

Several studies looked into the differences in content presentation between the encyclopedias. Messner and DiStaso reported that Wiki-pedia uses a more positive/negative language than Britannica when it comes to articles on large corporations. Greenstein et al. Editorial 2006 computed political slant and bias in 4K Britannica and Wikipedia articles on the US politics, and found that Wikipedia is more biased towards Democratic views. Their results vary depending on the length

of the article and the computation method, though. Finally, the encyclopedias have been compared in terms of content readability, but the results are also controversial (Elia, 2009; Jatowt and Tanaka, 2012; Lucassen et al., 2012).

Although the actual differences between Wikipedia and Britannica in terms of content quality and reliability are not great, Wikipedia suffers from perceived credibility and article selection issues (Chesney, 2006; Lucassen and Schraagen, 2010; Samoilenko and Yasseri, 2014), especially when contrasted with Britannica (Flanagin and Metzger, 2011; Kubiszewski et al., 2011). To sum up, most comparative studies focus only on one dimension (usually, content validity), and do not offer a holistic picture of structural differences between the encyclopedias.

Crowd- vs. expert-written history: While Britannica presents a credible, expert-written resource on history, Wikipedia offers an unsupervised, self-emerging, and multifaceted view of the past. In Social Sciences and History literature, Wikipedia is studied in the paradigms of open source history, participatory/amateur history-making (Rosenzweig, 2006), collective memories (Conrad, 2007; Rosenzweig and Thelen, 1998), and collaborative re-interpretation of the past (Pfister, 2011). While professional historians do not necessarily share the same understanding of the past as Wikipedians (Conrad, 2007), the immense popularity of Wikipedia as a reference source, especially on history (Spoerri, 2007), makes it an attractive object for studying.

When it comes to the history domain, the possible differences between crowd- and expert-written encyclopedic articles largely remain a terra incognita. To the best of our knowledge, only several studies have juxtaposed the accuracy, breadth, and depth of historical articles in Britannica and Wikipedia. Holman Rector, compared the content of nine Wikipedia articles against their equivalents in Britannica, the Dictionary on American History, and American National Biography Online, and found Wikipedia's accuracy to be less reliable (80% compared to 95% in other sources). Luyt and Tan discovered that this weakness is due to many claims in Wikipedia not being verified through citations. A qualitative analysis of the 'War of 1812' article in both encyclopedias (Jensen, 2012) showed that the Britannica article was briefer and focused more on the causes of the war, while lacking in military and naval aspects. The article also concludes that Wikipedia articles on military history are more detailed and easier to read than their Britannica counterparts.

Apart from qualitative research, several approaches have been used to quantify history on a large scale, including network science (Jackson, 2016; Padgett and Ansell, 1993; Schich et al., 2014; Sindbæk, 2007), mathematical modelling and prediction (Kiser and Hechter, 1991; Turchin, 2011), text mining and topic detection (Au Yeung and Jatowt, 2011; Michel et al., 2011), and temporal event extraction (Au Yeung and Jatowt, 2011; Samoilenko et al., 2017). None of them, however, have been applied to compare historical content of online encyclopedias. In this study, I combine computational methods in order to examine, how collaboratively produced Wikipedia articles on national histories compare to the equivalent Britannica articles, both in terms of temporal and topical coverage of events, as well as the linguistic characteristics.

4.3.2 Data collection and validation

In this section, I describe the process of collecting, pre-processing, and validating the data, as well as outline the methodological details. Similarly to the previous study, I continue focusing on the history of 193 countries which are the current UN member states. Although Wikipedia is a multilingual encyclopedia, I limit the analysis to its English edition. This is due to the fact that the Encyclopedia Britannica is only available in the English language, and thus, multilingual comparison is not possible.

Data collection: Wikipedia corpus. For each of the countries I locate an article in the English edition of Wikipedia, titled 'History of X', where X is the country name. I retrieve the article's main text, as well as the text of all Wikipedia articles to which this page outlinks. I focus on the outlinks because they provide readers with an opportunity to follow up and explore topically related material, and thus play a role in shaping user navigation across historical topics. Additionally,

Figure 4.9: Temporal information extraction. Parts of the article on the UK history as they appear on Britannica and English Wikipedia websites in 2017. I collect all 4-digit numbers from the main text of each article, as well as from the texts of all outlinked articles, and analyse the resulting distribution (bottom part of the figure). The data provides insights into the temporal focus and attention of encyclopedic articles.

they are embedded in the main articles' texts and thus are immediately available for a reader to inspect, unlike, for example, the in-links, which could not be found by reading the article page.

Data collection: Britannica corpus. The online Encyclopedia Britannica[7] has a format similar to Wikipedia: the articles are split into topical sections, some contain infoboxes, and the main text incorporates hyperlinks to other Britannica articles. Unlike in Wikipedia, there are no distinct articles on national histories. Instead, this information is embedded as a separate section in the main article about each nation. Usually, this section has multiple subsections focusing on various important events and periods, including the history of pre-states. For this analysis, I identify Britannica articles on all UN member states titled 'X', where X is a country name. For each article, I retrieve the text of the section titled 'History', as well as the text of the outlinks[8]. Other sections, such as 'Economy', 'Land', and 'Cultural life' are excluded as irrelevant.

Pre-processing of corpora. For both datasets, I extract data in HTML format, and clean it with BeautifulSoup parser to exclude text and tags related to, e.g. references, section titles and subtitles, captions, such that both datasets consist only of the main article text. For Wikipedia, I additionally remove (using regular expressions) all instances of citing references (in the format $[n]$, where n is the position of the reference in the article bibliography).

For analysis of language complexity, I prepare several corpora. First, I create a) (main + outlinks) corpus which encompasses all collected text per country, including both seed article and its outlinks. its reduced version b) (main) consists of the text of the seed articles, excluding the text of the outlinks. In these corpora the length of text about country X in Wikipedia might be of significantly different length compared to Britannica. I create and additional c) (main equalised) corpus based on the text of seed articles, but matched in size between Wikipedia and Britannica articles. Matching text size is a relative concept, since it can be measured in characters, words, and sentences, for example. For linguistic analysis, it does not make sense to cut a paragraph at half-sentence or a sentence at half-word. I perform the following procedure to equalise article sized between datasets. For every country, I compare article length (in words) between the two encyclopedias. I keep the shorter article as it is, and randomly remove sentences from the longer article until the word count is equal or lower than the size of the smaller article. As a result, the

[7]The Encyclopedia Britannica, https://www.britannica.com/ (accessed 16 May 2017)

[8]One exception is the article on Monaco, which is not split into sections. In this case, I used the entire text of the article and all of its outlinks for the analysis.

word count per country is the same across Wikipedia and Britannica, rounded up to the sentence boundary.

Extracting temporal expressions. In order to assess the coverage of historical periods, I count mentions of year numbers in article texts. Since I am interested in historical events of the last millennium, I only retrieve the 4-digit numbers in the range between 1000 and 1999. I use the same procedure (illustrated in Fig.4.9) for extracting temporal expressions from both datasets. In Wikipedia I encounter examples of paragraphs that consist mostly (more than 50% of words) of hyperlinks. Since there is little narrative in such paragraphs, I record no dates from them.

I ran data collection for both datasets in February 2017, using the access provided by the Wikipedia API[9], and an HTML scraping script for Britannica. As a result, for Britannica dataset I extracted 326K dates from 27,045 articles including the outlinked articles. In case of Wikipedia, I processed 54,401 pages and retrieved approximately 3M dates. For both datasets, the focus is only on the main text of the articles, excluding infoboxes, section titles, and figure captions.

Validation of extracted time expressions. In order to ensure internal reliability of our extraction method, I check whether the extracted numbers are years rather than numerals indicating, for example, height or distances. For each dataset, I create a random sample of 1,000 extracted 4-digit numbers evenly split across 10 centuries, and ask 3 independent human coders to evaluate each number as a date or a false positive. For each century there are 100 evaluation tasks, which consist of the potential date (4 digits), and the text surrounding it (40 characters before and after the number). If the coder is unsure about a number, it is counted as a false positive. Each case is settled by the majority vote. I compute the expected error rates for centuries as

$$\langle E_{corp} \rangle = \frac{1}{D_{corp}} \sum_{c} \left(\frac{n_{err,c}}{100} D_{corp,c} \right), \tag{4.5}$$

where D_{corp} and $D_{corp,c}$ are the total counts of collected (potential) dates per corpus $corp$ and century c, and $n_{err,c}$ is the count of false positives in our random sample for century c.

The inter-rater agreement is substantial (Fleiss' kappa = .79). Both datasets show very low expected error rates (0.01 per dataset). For Wikipedia, the estimated highest error rate is in the 11th century (.24), since a large number of extracted digits turned out to be numerals relating to heights, population counts, etc. Other false-positives, both for Britannia and Wikipedia, are mostly dates from the Before Christ era. In the more recent centuries the extraction method is very exact (expected error for the 20th century is < .001).

4.3.3 Analysis and Results

The results are presented in several parts. First, I compare Britannica and Wikipedia in terms of the most covered years and historical periods. Then I narrow the analysis down to selected countries, and calculate the decades that are covered most differently across the datasets, as well as extract and compare temporal focal points of nations. Finally, I report on the linguistic presentation of

[9]Wikipedia API for Python, https://pypi.python.org/pypi/wikipedia/ (accessed 16 May 2017)

Exp.error	WP	BR	Exp.error	WP	BR
11th century	0.23	0.03	16th century	0.00	0.02
12th century	0.07	0.07	17th century	0.00	0.00
13th century	0.06	0.02	18th century	0.02	0.00
14th century	0.02	0.01	19th century	0.01	0.00
15th century	0.03	0.03	20th century	0.00	0.00
Total				0.01	0.01

Table 4.2: Expected error rates in extracted dates. Our extraction method is on average very exact both in Wikipedia and Britannica.

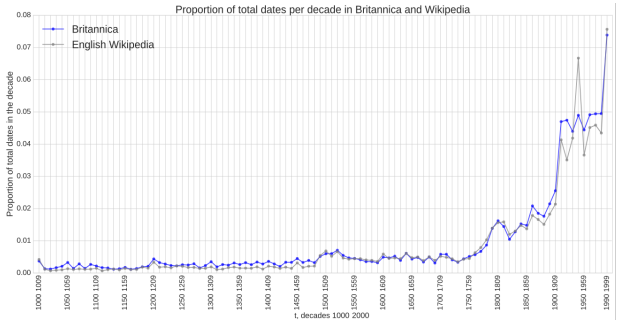


Figure 4.10: Normalised distribution of collected dates. All collected years are binned into decades, and normalised by the total number of collected dates per dataset. Both Wikipedia and Britannica show a strong bias towards covering the last 100 years. Wikipedia demonstrates a visibly higher peak of coverage in the decade corresponding to the WWii.

the articles and compare the most distinctive vocabulary characterising each dataset. I conclude the analysis by presenting an overall comparison of readability and language complexity of the encyclopedias.

General patterns of coverage

Before diving into the computational analysis, I compare the datasets in terms of the number of collected dates and their distribution across the national timelines. There is a startling difference in the number of dates collected from both encyclopedias: while Britannica has a total of 326,021 year numbers between 1000 and 1999, Wikipedia is a tenfold as large with 3,325,946 dates. Some of the most covered countries in both encyclopedias are large European economies (e.g. the UK, Germany, France) accompanied by Australia and the US. The least covered tail of Wikipedia is dominated by the African countries and island states of Oceania. This trend is visible in Britannica too, although it also includes some Asian nations. Overall, there are only 98 countries for which I extract more than 1,000 dates from Britannica articles. In the Wikipedia dataset, even the least covered country has about 1,500 dates.

In order to compare the distribution of dates across the corpora, I bin all collected dates into decades, and normalise them by the total number of dates collected per dataset. Both Wikipedia and Britannica show an uneven distribution of temporal coverage (Fig. 4.10) with small peaks around 1500 (possibly related to the Age of Discovery) and 1800 (Napoleonic war). A particularly strong peak falls on the 20th century, where the periods of First and Second world wars are most visible. Overall, for both encyclopedias I observe a strong bias towards covering the last century. Additionally, Wikipedia demonstrates a visibly higher peak of coverage in the decade corresponding to the WWII.

National temporal distributions

I first explore the overall similarity between Wikipedia and Britannica timelines for each country. For that, I present each country as a vector of 100 values (equal to the number of examined decades), each value being the normalised date count. I then compute cosine similarity between the Wikipedia and Britannica country vectors. Overall, the similarity values range between .59 (San Marino) and .98 (Botswana, Rwanda, Australia), with an average of .88. Thus, the timelines are on average very similar.

To continue, I explore how focused the national timelines are on covering particular periods, as opposed to covering every decade to a similar extent. I take an information theory approach and treat each decade bin of a national timeline as a separate information channel, and compute the entropy across all channels. Thus, the country with an equal number of dates in each decade

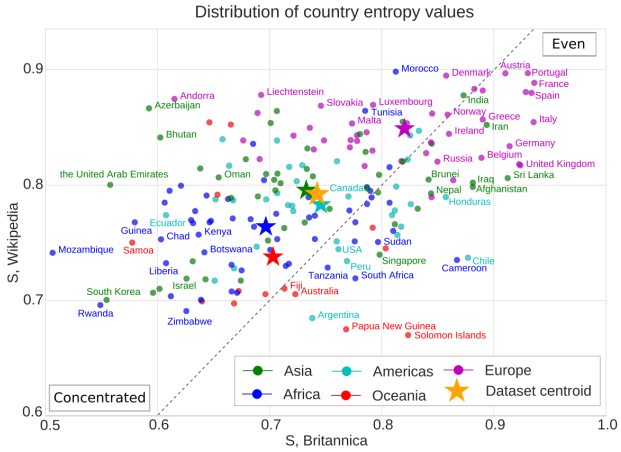


Figure 4.11: Distribution of country entropy values. The scores are normalised to range between 0 (all country dates in one decade) to 1 (all decades have equal number of dates). Based on the location of centroids (stars), in both encyclopedias the timelines of European states are presented in the most equalised manner, while for the countries of Africa and Oceania, Wikipedia and Britannica articles are more biased towards covering a limited number of decades. On average, this bias towards covering certain decades is more pronounced in Britannica (all centroids are above the diagonal).

will have the maximum entropy. Evidently, the minimum entropy corresponds to the case when all country dates are concentrated in just one decade. Country entropy is computed as $S_c = -\sum p_d \ln(p_d)$, where p_d is the normalised frequency of dates in decade d.

Fig. 4.11 demonstrates the distribution of the entropy scores. Based on the location of centroids, one can conclude that in both encyclopedias the timelines of European states are presented in the most equalised manner, while for the countries of Africa and Oceania, Wikipedia and Britannica articles are more biased towards covering a limited number of decades. This bias towards covering certain decades is more typical in Britannica (all centroids are above the diagonal). A few exceptions from this rule are large European states (the UK, Germany, Italy, Spain, France), whose historical timelines are presented much more equally on Britannica, compared to Wikipedia. In the next subsection I continue to examine the cases where temporal coverage differs substantially across the encyclopedias.

Most differently covered periods

As demonstrated in the previous subsection and in the Fig. 4.10, the shapes of national timelines in Britannica and Wikipedia are on average very similar. However, discrepancies are also present. In this section I automatically extract and highlight the decades which are covered differently by the encyclopedias. In particular, I explore in which decades the number of dates in one dataset is noticeably higher (or lower) than a fixed expected baseline. It makes intuitive sense to define the baseline R_c as the ratio of Britannica to Wikipedia total dates for a given country c. I assume that this ratio remains constant for each decade in a timeline of a country. Thus, I test the assumption that regardless of decade, Britannica will always have R_c times fewer dates compared to Wikipedia.

Results: Fig. 4.12 visualises the outcome of this simulation for a selection of countries. It is visible that the encyclopedias have data sparsity issues, especially pronounced in earlier decades and in non-European countries. This issue affects Britannica to a much greater extent than Wikipedia. Nevertheless, based on the decades where both encyclopedias have enough dates, Britannica pays proportionally more attention to the earlier periods. Precisely, in most of the decades before the 20th century the ratio of Britannica to Wikipedia date counts exceeds the expected R_c ratio for

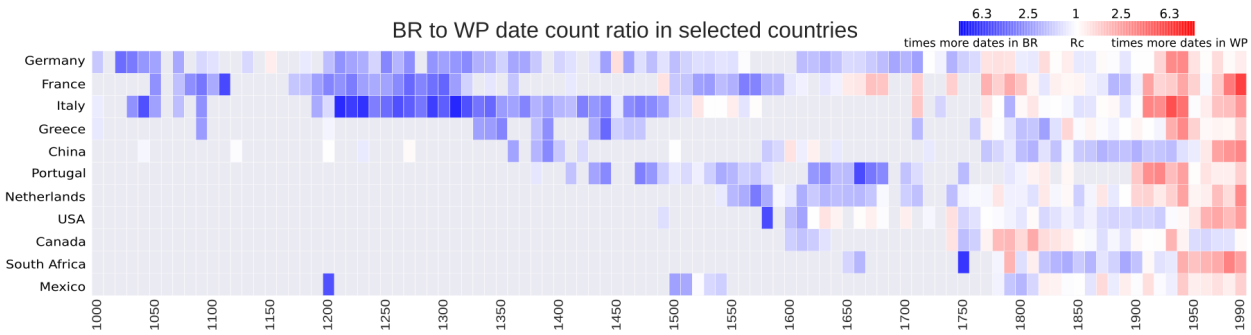


Figure 4.12: Comparison of Wikipedia and Britannica timelines. For each row I compute R_c, a ratio BR/WP based on the number of collected dates for that country. I then compare this country ratio to each BR to WP datecount ratio in each decade. Cell colour shows by how many times the decade ratio is different from the country ratio. The cells where Britannica has more decade dates than predicted by the country ratio are coloured in blue, otherwise – red. Cells with fewer than 30 dates in either BR of WP are masked out (grey). The cells are white when country and decade ratios are equal. The plot shows that for the decades where there is enough data, Britannica pays proportionally more attention to earlier decades, and Wikipedia focuses on the recent periods related to political instabilities, e.g. WWII.

that country. Wikipedia, on the other hand, has a strong bias towards more recent events. It is also noticeable that Wikipedia puts an overproportional emphasis on the times of conflict and war, which is true not only for the 20th century's First and Second world wars, but presumably also adds up to the red Wikipedia-cells in earlier periods (as shown in Fig.4.12). Some evident examples identified by historians include: the Franco-italian wars (1490s to 1550s), the Franco-Dutch war (1670s), the French War of Devolution (1667-68), the War of the Spanish Succession (1701-14), the history of Canada between its invasion (1775) and the War of 1812, the insurrection of Otto of Greece (1843), and the Crimean war (1853-56). Another focus of Wikipedian writing seems to fall on what might be called popular periods: the times that are well known not only to history experts but also to a wider audience, e.g. the reign of Louis XIV or the French Revolution in France, Reformation or the Age of Enlightenment, and the period of Weimar classicism in Germany. Britannica, in comparison, highlights times of conflict to a much smaller extent: the Wars of Religion (1560s, settled by the Edict of Nantes in 1598) in France, Restoration wars in Portugal (1640-48), or the Greek war of independence (1820s). It also shows a noticeable focus on the periods of African (de-)colonisation. Finally, this analysis provides additional evidence to support that on a national level, for Britannica the recency bias is less pronounced than for Wikipedia.

Historical focal points

To continue the investigation of temporal coverage patterns in Britannica and Wikipedia, I extract and compare *the focal points of national timelines*, i.e. the decades which are mentioned significantly more (or less) compared to what is expected by a Null Model. The extraction method is presented in Section 4.1.3, and applied separately to each of the encyclopedic datasets.

Comparison of extracted focal points. As a result, I obtain two timelines of focal points (Wikipedia and Britannica versions) for each country in the format of vectors. I summarise the differences between them by computing cosine similarity between each of the country vectors. The values of cosine similarity range between .92 for Argentina (both encyclopedias offer practically identical timelines) and -.55 for Morocco (focal points in one timeline are of low interest in the other), and are centred at .45. Thus, in terms of focal points, the encyclopedias offer rather diverging versions of national histories. Evidently, low average similarity is partially related to the missing data in Britannica. (For example, Morocco timeline has less than 20 decades with at least 30 dates.) However I also find dissimilarities between the decades for which data sparsity is not

an issue. To illustrate them, I plot the distribution of focal points obtained from each encyclopedia, one under another for 10 top covered countries (Fig. 4.13).

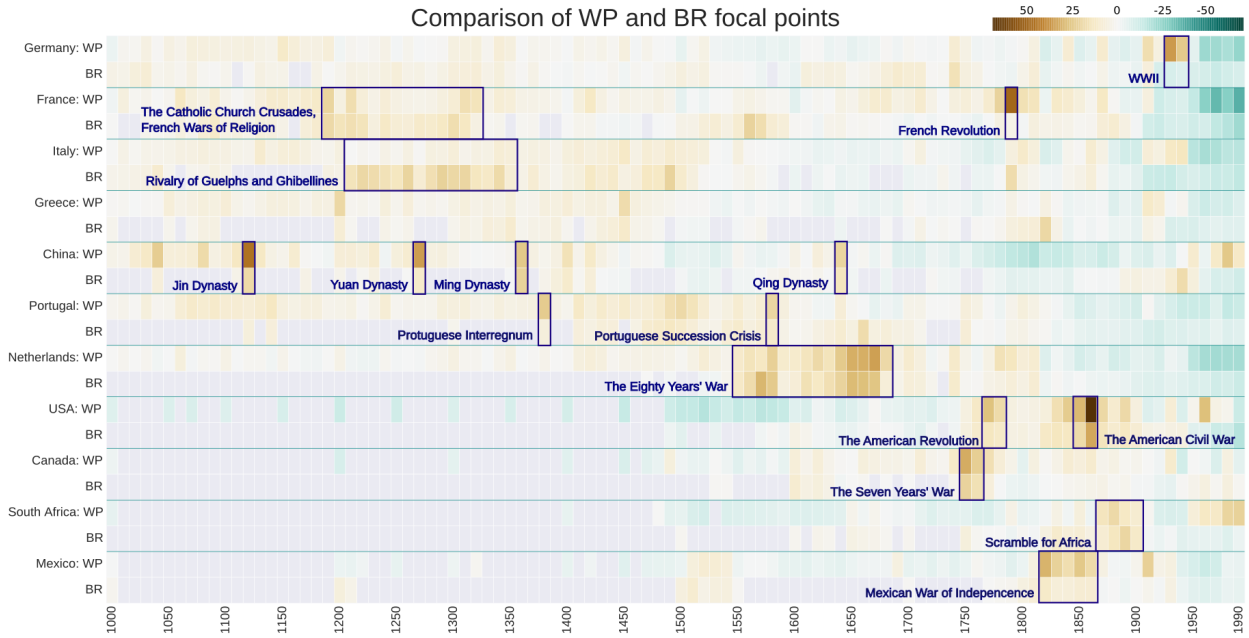


Figure 4.13: Temporal focal points of selected countries: comparison between Wikipedia and Britannica. z-scores below -4 and above 4 correspond to Bonferroni-corrected p-values < 0.01, which means the results in all coloured cells are statistically significant. Higher z-scores (orange) correspond to positive differences between the observed and the expected date count per decade, and could be interpreted as focal points of the timelines. Cells with fewer than 30 dates are masked out (grey). z-scores of Britannica vary between $[-50; 50]$, and of Wikipedia – between $[-70; 70]$. The annotations are produced by history experts. While overall the similarities between the distributions of focal points in Britannica in Wikipedia are evident, the differences are indicative of diverging approaches to historiography.

Two types of signal are evident. Even though I applied the method independently on each dataset, the *agreement in some focal points* is obvious. For Mexico, both encyclopedias focus on the Mexican War of independence (1820s). In the US timeline, the focal events are the American Revolution (1760-90s), and the American Civil War (1860s). Articles on Canadian history highlight the decades associated with the struggle between France and Britain for dominance in the North America (Seven Year's War, 1756-63). The history of South Africa in both encyclopedias mostly highlights the colonisation period (Scramble for Africa in late 19th century). For the Netherlands, the specific period of interest between 1560s and 1670s is likely related to the Eighty Years' War, or as it is also called, the Dutch War of independence against the Spanish political and religious hegemony. The history of Portugal focuses on the dynastic crises: Portuguese interregnum (1380s), a period of civil war triggered by the death of King Ferdinand I who left no male heir; and the succession crisis of 1580s. A similar trend shows up in the articles on the history of China, where both encyclopedias highlight the formation of the Jin (1130s), Yuan (1270s), Ming (1360s), and Qing (1640s) royal dynasties.

Perhaps even more interestingly, another signal in the data is the *disagreements between the encyclopedias*. This is pronounced most strongly in the articles on history of Germany. While Wikipedia narratives strongly focus on the WWII, Britannica is disinterested in the 1930-40s. Similarly surprising, the French Revolution (1780s) is pronounced on the Wikipedia's timeline, but it does not show up on the Britannica's timeline of history of France. Instead, Britannica focuses on the French Wars of Religion (Huguenot Wars of 16th century), and the extension of the Crown Lands of France (1180s to early 14th century), which coincided with the crusades by the Catholic Church

against the Cathards. Britannica articles on italian history focus on the Medieval period between 12th and 13th centuries, which is characterised by the rivalry of the Guelphs and Ghibellines, supporting the Pope and the Holy Roman Emperor. Wikipedia, on the other hand, shows no such emphasis.

Most distinctive topics and vocabulary.

After looking into some of the temporal coverage characteristics of the datasets, I move to a textual analysis of the articles in order to get a first understanding of the themes covered in the articles. I start by extracting the words that are most distinctly used in one dataset, compared to their usage in the other. For that, I extract the union between the top 1000 most frequent words from Wikipedia and Britannica (main corpus) (1219 words), and compare word frequencies using a χ^2 test of independence of variables in a contingency table. The results are reported in in Table 4.3. The words are ranked according to the value of χ^2 statistic, which reflects how significantly biased the usage of the word is towards Britannica (left column) or Wikipedia (right column). Among the analysed words, Britannica relies most distinctly on vocabulary with religious or philosophical connotations, such as *Christ, faith, Jesus, God, spirit, divine; idea, doctrine, systems* and geographical terms (*rivers, plain, basin, mountain, rocks*). Wikipedia, on the other hand, relies heavily on political and military vocabulary, such as *war, killed, colony, soldiers, army, empire, ships, armed, captured*.

Text complexity and readability

Both encyclopedias aim at a wide range of readership, and thus should be written in a way that is accessible to a diverse audience. In this section, I explore this intuitive hypothesis by computing various language complexity measures. Below I report on how two corpora compare in terms of simple text statistics, article readability, and part of speech usage. Depending on the analysis, I use either the entire Wikipedia and Britannica corpora (main + outlinks), or their reduced versions (main) and (equalised main). Section 4.3.2 describes how these corpora are constructed.

Text statistics. At first I report the descriptive text statistics for the (main) corpus. These are computed for each country article separately, averages over each dataset are summarised in Table 4.4. I use Welsch's t-test to compare the means. On average, Wikipedia articles about history use longer sentences (21.6 words vs. 19.9 in Britannica, $p < .001$), and slightly longer words (5.2 characters vs. 5.1, $p = .005$); the differences are statistically significant. To put the numbers in perspective, note the average sentence length in spoken speech (18 words on average) and academic writing (24 words) (Chafe and Danielewicz, 1987). Longer unites of text indicate that Wikipedia uses a slightly more formal writing register. Based on the average word length, both encyclopedias score higher than Academic prose (4.8 characters (Biber, 1995)), and thus belong to the most formal text genre.

Finally, I report the average article length, measured in the number of sentences and words per article (see Table 4.4). The comparison reveals no significant differences. However, there are interesting particularities in the way both datasets reference temporal information. Precisely, Wikipedia texts cite dates (years) significantly more often. The differences are significant both measured as number of dates per 100 words (1.7 dates in Wikipedia vs. 1.3 in Britannica, $p < .001$) and per 100 characters. This might indicate that Wikipedia leans towards factual, rather than descriptive narratives.

Readability. Text readability is usually estimated as the minimal number of education years needed to understand the text at first reading, and is often interpreted using the US grade level system. Readability scores are commonly based on surface characteristics of text, such as the number of its units (syllables, words, and sentences). Some of the tests also include semantic features, such as word difficulty estimated by the word length (in characters (Coleman and Liau, 1975; Senter and Smith, 1967) or syllables (Flesch, 1948; Gunning, 1952; Mc Laughlin, 1969), or

by comparison with pre-computed dictionaries of easily understandable words (Dale and Chall, 1948). Most of the scores make use of similar text statistics, but the coefficients of weighting these statistics are derived from different trials and application contexts, and thus vary. It is not clear, which readability score is more suited for which type of writing. This is why I use a range of established readability scores, in order to benefit from various approaches. The analysis is performed on the (equalised main) corpus to compensate for the article length differences.

The results are summarised in Table 4.5, all differences are statistically significant (Welsch's t-test, $p < .001$). FRE[10] ranges between 0 (very hard to understand) and 100 (understandable to a 5th grader). For both Wikipedia and Britannica the score is around 40, or appropriate for an average high school graduate. While the practical difference between the scores is not large, Wikipedia appears slightly easier to comprehend. Other measures concur with this result, always mapping Britannica's readability to a higher required US grade level (and thus, lacking readability).

While there is variation across the scores as to which graduate level to map each encyclopedia, between the datasets the signal is clear. Wikipedia consistently shows lower readability scores than Britannica, i.e. its articles are written in a language that is accessible to a wider audience. As a note, these grade scores should not be considered as precise values. Depending on a socio-economic and cultural background of the reader and their motivation to read the text, readability formulae are known both to over- and under-estimate comprehension difficulty (Klare, 1976).

Part of speech analysis. Another measure for assessing intrinsic linguistic differences across datasets is based on comparing the distributions of part of speech (POS) frequencies. For this analysis, I use the (main equalised) corpus. To tokenise the texts, I applied the Penn Treebank POS tokeniser (Penn, 2017). It erroneously counts multiword proper nouns as separate entities (e.g. *New York* results is two single proper noun tokens, rather than one multiword proper noun token). Thus, I added a layer of post-processing, merging into one token all instances of adjacent proper nouns which are not separated by punctuation or other parts of speech. The results of the analysis for the most frequent POS[11] are summarised in Fig. 4.14. Both encyclopedias show incredible similarity (cosine similarity = .99) in their patterns of POS usage.

The most used POS are nouns and adjectives, which is a general property of written Academic English (Biber et al., 1999). Since the focus of both corpora is on describing the past, verbs in past tenses are also frequent. I discover some interesting statistical differences between the datasets, for example, in usage of proper nouns and numerals. On average, Wikipedia tends to mention proper nouns and named entities (e.g. unique entities, people, well-known events) significantly more often than Britannica. It also uses cardinal numerals (indicating countable quantities, including dates) with much higher frequency. This hints that Wikipedia might be more focused on writing about famous events, entities, and biographies. Britannica, on the other hand, shows a notably high frequency of nouns, WH-determiners (*that, what, which*), and coordinating conjunctions (*therefore, and, but, so*). Thus, it may exhibit a more didactic and impersonal style, as well as an organised and logical flow of narrative with a focus on explaining structural connections between entities.

[10]The acronyms are abbreviated as follows: FRE - Flesch reading ease; FKG - Flesch-Kincaid grade; CLI - Coleman-Liau index; ARI - Automated readability index; DCRS - Dale-Chall readability score; G-FOG - Gunning FOG index; HS - High school.

[11]POS are defined as follows: NN: noun, common, singular or mass; iN: preposition or conjunction; DT: determiner; NNP: noun, proper, singular; JJ: adjective or numeral, ordinal; NNS: noun, common, plural; VBD: verb, past tense; CC: conjunction, coordinating; VBN: verb, past participle; CD: numeral, cardinal; RB: adverb; TO: to; VB: verb, base form; VBG: verb, present participle or gerund; PRP$: pronoun, possessive; VBZ: verb, present tense, 3rd person singular; PRP: pronoun, personal; VBP: verb, present tense, not 3rd person singular; WDT: WH-determiner; NNPS: noun, proper, plural; WP: WH-pronoun; JJR: adjective, comparative; JJS: adjective, superlative; WRB: Wh-adverb; MD: modal auxiliary; RP: particle; EX: existential there.

Biased towards Britannica			Biased towards Wikipedia		
Word	**BR**	**WP**	**Word**	**BR**	**WP**
feet	18474	22	due	53	107590
miles	16615	53	british	22766	377875
metres	15960	17	war	47661	599373
christ	14565	12	government	44254	561361
faith	13855	57	killed	275	84039
jesus	13330	17	japanese	240	79731
god	35350	36828	colony	288	77958
toward	11615	151	soldiers	178	73302
square	9884	62	started	80	65504
spirit	9757	44	anti	186	66960
divine	9369	16	army	17024	260246
rivers	8706	107	campaign	301	65399
football	8410	7	forces	13409	219442
plain	8440	54	empire	25631	342131
idea	8183	97	ships	93	60593
doctrine	7986	48	president	11871	202456
mountain	7728	70	police	171	57468
systems	7708	69	towards	1	54365
beyond	7367	98	portugal	201	57368
complex	7234	93	armed	296	58780
rocks	7029	8	french	23496	310546
basin	7081	78	captured	222	55474
games	6826	12	arrived	145	54157
extensive	6698	154	around	9159	162148
importance	6590	120	post	155	52273

Table 4.3: Top word usage in the main articles of Wikipedia and Britannica. On the left, top 25 words that appear most distinctly in Britannica (ranked by χ^2 values), compared to Wikipedia, and on the right – most distinct words in Wikipedia. The values correspond to word frequencies in the (main) corpus. While Britannica is distinct in using religious, philosophical and geographical vocabulary, Wikipedia is heavy on political and military terms.

Statistic	Wikipedia	Britannica
Av. word length**	5.2±.1	5.1±.1
Av. sentence length (char.)***	156.9±17.1	140.6±12.7
Av. sentence length (words)***	21.6±2.1	19.9±1.5
Av. lexicon count	6,831±8,860	7,040±5,535
Av. dates per 100 chars.***	.33±.11	0.25±.09
Av. dates per 100 words***	1.68±.57	1.28±.46

Table 4.4: Descriptive text statistics compared for Wikipedia and Britannica datasets. Texts of outlinked articles are excluded. Results are computed per article, averages are reported with standard deviations. Rows with statistically significant differences are starred: *** corresponds to $p < .001$, and ** corresponds to $p = .005$. Comparison suggests that Wikipedia uses a slightly more formal writing style, and on average cites dates more often than Britannica.

Av. read.	Wikipedia		Britannica	
FRE	46.67 ± 6.3	[HS]	42.9 ± 5.9	[HS]
FKG	11.92 ± 1.4	[12th gr.]	12.7 ± 1.4	[13th gr.]
CL*I*	13.8 ± 1.1	[14th gr.]	14.5 ± 1.2	[15th gr.]
AR*I*	14.5 ± 1.5	[15th gr.]	15.6 ± 1.7	[16th gr.]
DCRS	$8.8 \pm .8$	[12th gr.]	$9.1 \pm .8$	[13th gr.]
G-FOG	10.4 ± 1	[10th gr.]	10.9 ± 1.2	[11th gr.]
SMOG	8.8 ± 1.6	[9th gr.]	9.5 ± 1.3	[10th gr.]

Table 4.5: Readability scores[10]. Averages are computed on (equalised main) corpus, estimated grade levels are reported in brackets. All differences are statistically significant at $p < .001$. Across several readability scores, the educational requirements for reading articles about national histories on Wikipedia are lower than the corresponding articles on Britannica.

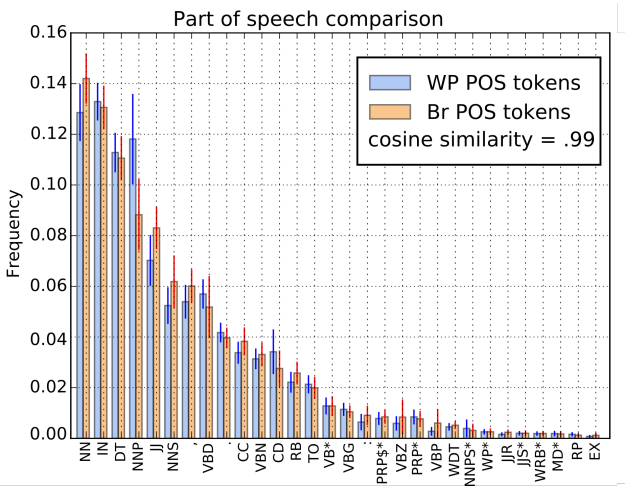


Figure 4.14: Part of speech[11] analysis of Britannica and Wikipedia. Mean frequencies and deviations are computed on the (main equalised) corpus. All differences are significant ($p < .001$, Welsh's t-test) except for the starred labels. The encyclopedias demonstrate nearly perfect overall similarity, but Wikipedia tends to mention proper nouns, and cardinal numerals (e.g. dates) significantly more often than Britannica. The notably high frequency of WH-determiners, and coordinating conjunctions in Britannica might indicate that its articles have a more structured and logical flow.

4.3.4 Discussion of empirical results

The empirical results indicate that both encyclopedias are biased towards covering the most recent periods more extensively than the remote past. This recency bias is especially pronounced in Wikipedia, which has a particularly strong emphasis on the First and Second world wars. The results presented in Section 4.2 have shown that this holds across other language editions of Wikipedia, which is partially attributed to the general psychological tendency to perceive recent events as more important (Rovira et al., 2006). This phenomenon is extensively discussed in the literature on collective/social memory (Assmann, 2011; Assmann and Czaplicka, 1995; Candau, 2005; Wertsch, 2002), however it is mainly associated with public, non-professional narratives. It is new to demonstrate the indication of the same bias in the expert-produced Encyclopedia Britannica. Co-authors and I also observe a more detailed and equalised temporal coverage for large economies, including mostly European states (the UK, Germany, France), the US, and Australia. On average, the history of the European region is comparably detailed across their entire timelines, while for African countries and the small island states of Oceania the timelines are skewed towards covering only a limited number of decades. This Eurocentric bias in professional historiography has been criticised from within the community (Geyer and Bright, 1995), but has not yet been discussed in the context of the Encyclopedia Britannica.

When it comes to language, both encyclopedias exhibit general properties of Academic English prose (Biber et al., 1999). In terms of text readability, Wikipedia articles are more accessible to a wide audience. The scores I report are similar to the results of the 2009 comparative analysis (Elia, 2009), however both findings should be interpreted with care (Klare, 1976). To add, the analysis of part of speech usage suggests that Britannica might offer an overall more didactic and impersonal style, together with a more organised and logical writing flow. At the same time, Wikipedia might lean towards stating numerical facts and focusing of famous name entities.

Juxtaposing the characteristics of temporal coverage across the encyclopedias, one notices that Britannica and Wikipedia might exhibit different approaches to historiography. Our temporal analysis reveals that Wikipedia puts an overproportional emphasis on periods of conflict and war, with a specific preference to the violent events well-known to a general public, such as French

Revolution, First- and Second Word Wars. Britannica articles do not show focal points associated with these events, but instead emphasise the conflicts with underlying religious tensions, for example, the French Wars of Religion, and the Crusades of the Catholic Church. Finally, the overall similarity between the distributions of focal points is rather low (cosine similarity of .45).

Comparing the most distinctly used words shows that Britannica relies on vocabulary with religious connotations and on geographical terms, while Wikipedia is heavy on political and military vocabulary. Moreover, Wikipedia's history articles cite numerals (including dates) significantly more frequently, and have an order of magnitude more dates compared to Britannica. This might indicate that the historiography on Wikipedia is oriented towards outlining facts rather than descriptive narratives. Finally, higher frequency of proper noun usage in Wikipedia supports the earlier observation that Wikipedia is more biased towards covering famous named entities, such as, e.g. well-known events and biographies. Overall, the data seem to suggest that Britannica leans to a more *sociocultural, spatial and territorial concept of history*, whereas Wikipedia – to presenting *a sequence of political events*. These computational results concur with some of the earlier qualitative observations. For example, a case study of the coverage of the Canadian War of 1812 pointed out Wikipedia's detailed focus on battles, military, and naval affairs, and sparsity regarding social and cultural historical aspects (Jensen, 2012). Britannica, on the other hand, was characterised as focusing on the national border line, and limited in the war thematic.

As a direction for future work, it would be interesting to analyse the accentuation of conflict in the encyclopedias. For instance, whether Wikipedia has a stronger interest in inter-nation conflicts, and Britannica – in sociocultural and intra-nation ones.

4.4 Limitations

In this section, I reflect on the methodological choices, such as the operationalisation framework, selected datasets, and analytical methods. I also discuss the effect that these choices have had on the conclusions of the analysis. These are the limitations of the studies presented in this chapter, grouped by type.

History of pre-states: The data might be lacking historical narratives about the history of territories before they reached the current shape. I focus on the history of the current UN member states, however the political map of the world has changed many times throughout history (e.g., post-Soviet bloc). Most Wikipedia and Britannica articles have sections on the history of pre-states in the text of the main article, or outlink to relevant articles. I include the text of outlinked articles to partially solve the issue. Still, some information on pre-history might be lost due to missing links. Also, inclusion of outlinks potentially makes the datasets noisier.

Unit of analysis: One of the main limitations of any computational approach is the fact that it is reductional: in both presented studies, I reduce the complexity of historiography to one quantifiable unit of analysis – year mentions. It is a very fine-grained unit when examining a millennium of human history. Year mentions might be less reliable for earlier periods and countries where the exact dates of events are not well-known or documented. Thus, mentions of decades or centuries might be accounted for in future analyses. Additionally, when it comes to multilingual analysis, different languages might also have different standards on mentioning dates. The advantage of this approach is in focusing on an objective, quantifiable unit that can be compared across linguistic datasets without the biases introduced by translation or cultural background of the researcher. It also scales well across large datasets.

Data validity: Data validation has shown high accuracy of our date extraction method. This is possible because the analysis is limited to the articles evidently related to history. The precision of the method might suffer when analysing texts of broader scope or focusing on the dates from Before Christ era. Already in the selected sample, evaluation finds small numbers of false-positives, e.g. 4-digit numerals expressing heights, lengths, or population counts. Although suitable for the current setup, the presented dates extraction method might need improvement if applied to a different dataset.

Linguistic scope: I focus on 30 languages native to the geographic region of Europe, which are also the largest editions of Wikipedia. For these editions, year is an acceptably robust unit of analysis, since these languages generally share date and time notation standards. I exclude other large editions such as Chinese, Arabic, and Farsi since their distinctive calendar- and numeral systems require developing language-specific methods of dates extraction, and this task goes beyond the scope of this study. The conclusions of these studies should not be generalised to the whole Wikipedia, and are only valid for the studied editions.

Generalisability to other domains: The present findings are valid for the chosen knowledge domain and the selected languages. It is problematic to generalise how Wikipedia articles on history in other language editions compare to Britannica. Additional research is needed to evaluate if these findings hold for articles with other themes than History.

Focal points: I define focal points as time periods of significantly high mentions, compared to a random expectation model. Other formulations of Null Model are possible, which could describe a random process otherwise, and potentially result in non-identical outcomes. Interpretations of historical events related to some extracted focal points depict a viewpoint of selected history experts and are subjective.

Article disambiguation: This analysis focuses on the articles with the specific title wording, 'History of X'. To solve title disambiguation issues in the English edition, I manually map all countries to corresponding Wikipedia articles on their modern history. In cases when a territory has changed names several times (having been a part of several countries, e.g., post-Soviet bloc),

there might be multiple Wikipedia articles related to its history. I partially tackle this issue by including out-linked articles in our dataset.

Inhomogeneous linguistic data: In Section 4.2 I compare Wikipedia language editions at different ages, states of saturation, and sizes of underlying potential editor populations. This unavoidably leads to an overrepresentation of larger editions, for example, the pool of dates is heavily influenced by German and English editions.

Multilingual data retrieval: Our method of retrieving sister-articles from non-English language editions relies on Wikipedia's inter-language links (ILLs). Although the quality of ILLs is a debatable issue, studies have shown that the proportion of bidirectional ILLs between English and the largest European languages is around 98% (Rinser et al., 2013). I do not exclude the possibility that some of the multilingual articles might have been missed, however it is reasonable to assume that their absolute share will not have dramatically affected the results of the study.

Text analysis: The outcomes of text and readability analyses are sensitive to the tokenisers and text pre-processing (Palotti et al., 2015). Slightly different results might be expected if applying other methods.

4.5 Chapter summary

In this section I reflect on the characteristics of the proposed computational approach, and briefly summarise the empirical findings. This section finishes the chapter with concluding remarks.

Approach. I have chosen year mentions as a language-independent unit of analysis, to measure narrative's emphasis on certain historical events. This formalisation has been used before by Michel et al.. In principle, it is possible to choose any other quantifiable unit, such as mentions of geographical locations, persons, or events, providing there is an extra step for tackling interlingual entity disambiguation. The building blocks of this approach are not new in computational fields. However, they are new to the community of quantitative historians, which faces challenges addressing this hitherto unseen inflow of newly digitised historical records.

This three-step approach is general enough to be applied to many large digitised datasets, such as: demographic and economic records, census data, books, etc. Additionally, it is suitable for comparative analysis of any number of countries across languages (as long as the languages are applying the same system for counting and mentioning year dates, see Section 4.4). Importantly, by applying purely computational and data-driven methods, this work aims to eliminate the bias that could be posed by the researcher's cultural background (Ailon, 2008). Finally, it performs transnational analysis on a scale previously unknown to comparative historiography.

The approach is validated through two empirical studies focusing on (multilingual) narratives in a specific knowledge domain of historiography.

Validation I: Historical landscapes of multilingual Wikipedia: First empirical study (Section 4.2) compares historiographical writing across 30 large language editions of Wikipedia. It elicits that all studied timelines are skewed towards more recent events (*recency bias*). Additionally, it finds that Wikipedia narratives about national histories are distributed unevenly across the continents with significant focus on the history of European countries (*Eurocentric bias*). Moreover, the clusters of countries with similarly distributed focal points map well to geopolitical blocs. Finally, mapping countries according to their Jensen-Shannon divergence scores shows that the national historical timelines vary across language editions, although average interlingual consensus is rather high.

Validation II: Britannica vs. Wikipedia perspectives on historiography: The second case study (Section 4.3) compares how historiographical narratives of Wikipedians compare to those of Encyclopedia Britannica experts. The Null model for extracting the national focal points is modified to fit the new dataset. Analysis is extended to include linguistic features comparison: we apply a range of established readability scores, compare the most distinctly used words in each corpora, and run a Part of Speech analysis.

The combination of temporal and linguistic analyses allows us to arrive at a comprehensive account of differences between the encyclopedias, which include structure (temporal coverage and focal points distribution), content (semantic differences and the historical significance of the extracted focal points), and presentation (readability and part of speech usage). Particularly, Wikipedia leans to presenting history as a sequence of political events, putting a disproportional emphasis on periods of war and violent conflicts, as well as the events well-known to the general public. At the same time, Britannica is concerned with a more spatial and territorial concept of the history of states, emphasising the conflicts with underlying religious or cultural tensions. These differences are also reflected in the semantic analysis, which shows that Wikipedia relies on political and military words, while Britannica is heavy on vocabulary with religious connotations and on geographical terms. The analysis concludes that Wikipedia and Britannica exhibit different approaches to historiography.

4.6 Conclusions and implications

Overall, this chapter achieves the following goals:

- elaborating a data-driven approach that enables studying of narratives through a computational lens;
- validating it on datasets from Encyclopedia Britannica and Wikipedia;
- delivering empirical insights by comparing historiography across linguistic-, national-, and expert vs. amateur communities of writers.

This chapter shows how to reduce and juxtapose large, multlingual historiographical corpora. This is done on the example of two data sources with comparable internal organisation, i.e. encyclopedias Wikipedia and Britannica. The approach assembles in a creative way a variety of statistical and linguistic analysis techniques in order to arrive at comprehensive conclusions about the compared corpora. The empirical studies in Sections 4.2 and 4.3 also demonstrate that it is easy to extend and adapt the approach for various data contexts.

To conclude this chapter, I discuss the **implications** of this research.

These case studies have a strong empirical value. For the first time, they present large-scale and multilingual quantitative investigation of the differences between expert-written historiography of Britannica, and Wikipedia's popular view of the past. This concerns not only the community of professional historians. Awareness of these structural differences between the encyclopedias might also be useful for Wikipedia editors who wish to expand the scope of the articles on world history.

The undisputed popularity and outreach of Wikipedia make it a worthwhile object of study, because its images of history may distort our view back. In particular, public awareness of the past might already be skewed by focusing on already popular and well-known periods, as well as on violent conflicts and political events. In revealing blank spaces or biases, this research contributes to fostering richer and more balanced accounts of history.

Moreover, these observed 'peaks' and 'lows' of interest to certain time periods, as well as cross-lingual differences in national timelines, might have different explanations. If these dissimilarities are *intentional*, they might be a reflection of cultural differences. In this case, these results could be interesting to historians and culture scholars who might wish to explore the topic in greater detail and with other methods.

If these differences are *accidental* or could be reduced to 'missing data', these findings could be actionable for the Wikimedia community and enthusiastic editors. For example, this work might motivate them to improve the quality of historiographical articles in various language editions.

Furthermore, the results presented in this chapter show that neither Britannica's nor Wikipedia's historical reference articles are free from gaps and biases. I hope that History teachers and students, as well as lay readers who use online sources, and Wikipedia in particular, to enrich their knowledge about world history, would benefit from this awareness.

Finally, for the computational scholars and practitioners, this work might serve as an invitation to explore other units of comparison, computational techniques, and application domains. The success of these initial studies warrants for further testing of the approach in other environments. It also indicates that the proposed framework can be useful in diverse multidimensional settings where large-scale comparisons are needed. I hope that that this research provides a starting point for a broader computational analysis of written history on Wikipedia and elsewhere.

Chapter 5

Conclusion

UGC is an important part of the modern "big data" revolution, and culture is an important factor which implicitly shapes UGC in many non-obvious ways. This thesis has focused on the following conceptual paradox.

On the one hand, modern computing technologies rely on UGC to act as the brain and blood of the algorithms, with a certain assumption that these data are pristine and somehow represent a certain degree of trustworthiness. On the other hand, mounting evidence suggests that multiple factors, and in particular, those related to the cultural background of the users, play role in what kind of content is generated online. While much effort has gone into integrating UGC data into the workflow of the modern algorithms, not nearly enough has focused on examining these data in the first place, and measuring the biases and imbalances that they potentially introduce into the modern computational systems.

This work has examined a particularly important example of UGC, the Encyclopedia Wikipedia. A review of the literature in Section 2.2, has shown that Wikipedia already plays a tangible role in Business, Academia, Public Policy, and importantly, in Computer Science and Engineering, and this relationship is unlikely to weaken in the near future. However, as much as Wikipedia acts as "the sum of all human knowledge", this thesis has shown that like all humans, it is biased in certain ways.

In particular, we saw in Chapter 3 that the selection of the topics covered in each Wikipedia edition is not a random choice. It is rather is explained by geopolitical, historical, and economic factors that have shaped its underlying language community of editors, in real, offline life. Similarly, Chapter 4 demonstrated that these differences show up not only in the selection of topics, but also in the narratives themselves.

These findings, while to some degree expected, pose important challenges. In particular, these challenges concern Engineers and Computer Scientists who wish to use Wikipedia data to improve their systems. One example of such improvement might be using the Wikipedia database in order to solve the cold start problem in some of the algorithms. Which data should they choose for that? There seems to be no universally right option. As this thesis has illustrated, Wikipedia is culturally contextualised, and all these perspectives are biased in certain ways.

Rather than warn against incorporating such data into the algorithms all together, this work aims to raise awareness of the presence of culture-related differences in UGC, so that the right data is selected for the right use case.

More generally, this thesis views the phenomenon of cultural contextualisation in UGC, and in Wikipedia in particular, both as a challenge, and an opportunity for the Computer Science community. On the one hand, the risk of "injecting" cultural bias into "big data" technologies is practically unavoidable, and has unintended, and truly understudied consequences. On the other hand, cultural contextualisation demonstrates the richness of UGC, and thus, its usefulness in designing more personalised, more relevant user experiences online. Consider, for example, the potential opportunity for the users to shape their own experience by deliberately switching between the points of view and cultural perspectives offered by a variety of culturally contextualised UGC.

This work demonstrates how complex it is to quantify cultural effects in user-generated content. Working with such content is a major challenge, but it is also a source of great opportunities for the Computer Science community. To a large extent, it is a responsibility of our community to continue maintaining and developing technologies that encourage cultural diversity in content production, instead of always putting one (English) perspective in the forefront.

5.1 Implications

Present work has several important implications:

- **Methodological implications.** Cultural context is a natural part of UGC, and being aware of this context is an important part of making the right research design decisions. Sometimes, more data is not necessarily better. Instead, more reliable results might come from differently sampled, theoretically relevant data. Moreover, involving domain experts in research process enriches the interpretative part of the analysis and helps design more relevant indicators and units of analysis.

- **Empirical implications.** Wikipedia editing is done (substantially) by humans, and in such, the encyclopedia contains human biases. Since the applications of such data are diverse, so are the implications. Through Wikipedia content re-use, these biases are potentially injected into large-scale, high impact algorithms. As such, they have a potential to reinforce information inequality, for example, by driving to extinction points of view of cultural minorities, and producing other unintended and unpredictable social spirals.

- **Data applications.** This research demonstrates that there is a huge, and so far unexplored potential for developing smarter algorithms. Algorithms, which would be aware of the cultural diversity of the online content, tailored to specific linguistic communities, and rely on the culturally relevant data. Additionally, there is a high potential in developing Web tools which would allow the user to switch between the cultural viewpoints that they would like to see, as well as compare such cultural perspectives side by side.

- **Considerations for Wikimedia Foundation.** It is a common practice in Wikipedia to use bots for automatic content creation, especially in the language editions where the community of editors is small. While such inorganically created content is often indistinguishable for human readers, these practices might backfire on the encyclopedia's quality. For example, this happens in those cases when automatically injected content is culturally too foreign for the editors to pick up on the writing. As a result, these articles have a high chance to remain incomplete stubs, have high susceptibility to vandalism, and low quality of the content. Finally, automatic translations usually happen from larger editions to smaller, thus diluting the cultural content in the smaller editions, and likely, decreasing the potential of such data.

Additionally, implications of each empirical study are outlined in major detail in Sections 3.7 and 4.6.

5.2 Limitations and future work

By way of concluding, I point to the limitations of this work, and discuss several directions for future research.

This thesis has proposed and validated two approaches to quantifying cultural context in UGC. The limitations of these approaches, as well as of the experimental setup and the chosen data, are provided in detail in Sections 3.5 and 4.4. Instead af re-iterating those, below I summarize some of the more general limitations that run through this work. While doing so, I also propose how they can be tackled in future research.

First of all, presented observations warrant more rigorous investigation of cultural borders. Current work has focused on linguistic and geopolitical operationalisation of cultures, but endless opportunity lies in exploring other definitions. These include, but are not limited to, religion-specific cultures, gender-specific cultures, urban/rural cultures, as well as the cultures specific to groups with a certain occupation, political inclination, or economic situation.

Additionally, it would be valuable to examine the phenomenon of cultural contextualisation of UGC in a longitudinal paradigm, mapping changes over time. A particularly intriguing question here is whether over time the trend goes towards converging cultural perspectives, or alternatively, cultural segregation further intensifies.

To continue, the approaches proposed in this thesis could be validated in the contexts other than UGC. Interesting examples which are likely to be rich in cultural content are digitised historical records, book archives, and art collections. Empirically, it would be interesting to retrace the development of cultural segregation in historical past, as well as quantify and compare cultural manifestations of human thought in the pre-Internet era.

It is equivalently interesting to explore the sources of digital data beyond the domain of collective knowledge documentation. For example, news media present a challenging, multilingual, and culturally rich object of study. Moreover, it has not been explored much within the large-scale computational paradigm. The approaches outlined in this thesis could be easily extended to compare multilingual perspectives, for example, in such large collections as the GDELT Project (Global Database of Events, Language, and Tone).

Bibliography

Agar, M.
1994. *Language shock: Understanding the culture of conversation*. William Morrow & Company.

Ahn, B. G., B. Van Durme, and C. Callison-Burch
2011. Wikitopics: What is popular on Wikipedia and why. In *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*, Pp. 33–40. Association for Computational Linguistics.

Ailon, G.
2008. Mirror, mirror on the wall: Culture's consequences in a value test of its own design. *Academy of Management Review*, 33(4):885–904.

Aitken, M., T. Altmann, and D. Rosen
2014. Engaging patients through social media. *IMS Institute for healthcare informatics, Tech. Rep*.

Allik, J. and R. R. McCrae
2004. Toward a Geography of personality traits: Patterns of profiles across 36 cultures. *Journal of Cross-Cultural Psychology*, 35(1):13–28.

Althoff, T., D. Borth, J. Hees, and A. Dengel
2013. Analysis and forecasting of trending topics in online media streams. In *Proceedings of the 21st ACM international conference on Multimedia*, Pp. 907–916. ACM.

Anderson, B.
2016. *Imagined communities: Reflections on the origin and spread of nationalism*. Verso London.

Antin, J. and C. Cheshire
2010. Readers are not free-riders: reading as a form of participation on Wikipedia. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, Pp. 127–130. ACM.

Antin, J., R. Yee, C. Cheshire, and O. Nov
2011. Gender differences in Wikipedia editing. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, Pp. 11–14, New York, NY, USA. ACM.

Apic, G., M. J. Betts, and R. B. Russell
2011. Content disputes in Wikipedia reflect geopolitical instability. *PloS one*, 6(6):e20902.

Aragon, P., D. Laniado, A. Kaltenbrunner, and Y. Volkovich
2012. Biographical social networks on Wikipedia: A cross-cultural study of links that made history. In *Proceedings of the eighth annual international symposium on Wikis and open collaboration*, P. 19. ACM.

Arazy, O., F. Ortega, O. Nov, L. Yeo, and A. Balila
2015. Functional roles and career paths in Wikipedia. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, Pp. 1092–1105. ACM.

Assmann, J.
2011. Communicative and cultural memory. In *Cultural Memories*, Pp. 15–27. Springer.

Assmann, J. and J. Czaplicka
1995. Collective memory and cultural identity. *New German Critique*, (65):125–133.

Au Yeung, C.-m. and A. Jatowt
2011. Studying how the past is remembered: Towards computational history through large scale text mining. In *CIKM'11*, Pp. 1231–1240. ACM.

Auer, S. and J. Lehmann
2007. What have Innsbruck and Leipzig in common? Extracting semantics from Wiki content. In *The Semantic Web: Research and Applications*, Pp. 503–517. Springer.

Azzam, A.
2017. Embracing Wikipedia as a teaching and learning tool benefits health professional schools and the populations they serve. *Innovations in Global Health Professions Education*.

Backstrom, L., E. Sun, and C. Marlow
2010. Find me if you can: Improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, Pp. 61–70, New York, NY, USA. ACM.

Baeza-Yates, R.
2009. User generated content: How good is it? In *Proceedings of the 3rd workshop on Information credibility on the web*, Pp. 1–2. ACM.

Baker, A. R. and G. Biger
2006. *Ideology and landscape in historical perspective: Essays on the meanings of some places in the past*, volume 18. Cambridge University Press.

Bao, P., B. Hecht, S. Carton, M. Quaderi, M. Horn, and D. Gergle
2012. Omnipedia: Bridging the Wikipedia language gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, Pp. 1075–1084, Austin, Texas, USA. ACM.

Bardak, B. and M. Tan
2015. Prediction of influenza outbreaks by integrating Wikipedia article access logs and Google flu trend data. In *Bioinformatics and Bioengineering (BIBE), 2015 IEEE 15th International Conference on*, Pp. 1–6. IEEE.

Barnett, G. A. and G. A. Benefield
2015. Predicting international Facebook ties through cultural homophily and other factors. *New Media & Society*.

Bartlett, F. C.
1932. Remembering: An experimental and social study. *Cambridge: Cambridge University*.

Bel Habib, I.
2011. Multilingual skills provide export benefits and better access to new emerging markets. *Sens-Public*.

Bellomi, F. and R. Bonato
2005. Network analysis for Wikipedia. In *proceedings of Wikimania*.

Benkler, Y., H. Roberts, R. Faris, A. Solow-Niederman, and B. Etling
2015. Social mobilization and the networked public sphere: Mapping the sopa-pipa debate. *Political Communication*, 32(4):594–624.

Bennett, J.
2015. Modeling the large-scale demographic changes of the Old World. *Cliodynamics: The Journal of Quantitative History and Cultural Evolution*, 6(1).

Biber, D.
1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan
1999. *Longman Grammar of spoken and written English*. London: Longman.

Bleich, E.
2005. The legacies of history? Colonization and immigrant integration in Britain and France. *Theory Soc.*, 34(2):171–195.

Bloch, M.
1991. Language, anthropology and cognitive science. *Man*, Pp. 183–198.

Blondel, V. D., J.-L. Guillaume, R. Lambiotte, and E. Lefebvre
2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Bloomfield, L.
1945. About foreign language teaching. *Yale Review*, 34:625–41.

Blum, B. S. and A. Goldfarb
2006. Does the Internet defy the law of gravity? *Journal of international economics*, 70(2):384–405.

Blumenstock, J. E.
2008. Size matters: Word count as a measure of quality on Wikipedia. In *Proceedings of the 17th international conference on World Wide Web*, Pp. 1095–1096. ACM.

Bond, M. H., K. Leung, A. Au, K.-K. Tong, S. R. De Carrasquel, F. Murakami, S. Yamaguchi, G. Bierbrauer, T. M. Singelis, M. Broer, et al.
2004. Culture-level dimensions of social axioms and their correlates across 41 cultures. *Journal of cross-cultural psychology*, 35(5):548–570.

Borra, E., E. Weltevrede, P. Ciuccarelli, A. Kaltenbrunner, D. Laniado, G. Magni, M. Mauri, R. Rogers, T. Venturini, et al.
2014. Contropedia – the analysis and visualization of controversies in Wikipedia articles. In *OpenSym*, Pp. 34–1.

Boulton, T. J., B. B. Francis, T. Shohfi, and D. Xin
2018. Investor awareness or information asymmetry? Wikipedia and ipo underpricing. *Available at SSRN*.

Brandes, U., P. Kenis, J. Lerner, and D. Van Raaij
2009. Network analysis of collaboration structure in Wikipedia. In *Proceedings of the 18th international conference on World wide web*, Pp. 731–740. ACM.

Breuing, A.
2010. Improving human-agent conversations by accessing contextual knowledge from Wikipedia. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 3, Pp. 428–431. IEEE.

Breuing, A., U. Waltinger, and I. Wachsmuth
2011. Harvesting Wikipedia knowledge to identify topics in ongoing natural language dialogs. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on*, volume 1, Pp. 445–450. IEEE.

Brown, A. R.
2011. Wikipedia as a data source for political scientists: Accuracy and completeness of coverage. *PS: Political Science & Politics*, 44(2):339–343.

Bruns, A.
2005. *Gatewatching: Collaborative online news production*. Peter Lang.

Bruns, A.
2008. *Blogs, Wikipedia, Second Life, and beyond: From production to produsage*. Peter Lang.

Bryson, L. and H. M. Jones
1948. Science and freedom.

Bucholtz, M. and K. Hall
2008. All of the above: New coalitions in sociocultural linguistics. *Journal of Sociolinguistics*, 12(4):401–431.

Butts, C. T.
2008. Social network analysis with SNA. *Journal of Statistical Software*, 24(6):1–51.

Cairncross, F.
2001. *The Death Of Distance: How The Communications Revolution Is Changing Our Lives*. Harvard Business Press.

Callahan, E. S. and S. C. Herring
2011. Cultural bias in Wikipedia content on famous persons. *J. Am. Soc. Inf. Sci.*, P. 1899–1915.

Candau, J.
2005. *Anthropologie de la mémoire*. Armand Colin.

Castells, M.
2011a. *The power of identity: The information age: Economy, society, and culture*, volume 2, second edition. Oxford: Wiley-Blackwell.

Castells, M.
2011b. *The rise of the network society: The information age: Economy, society, and culture*, volume 1. John Wiley & Sons.

Cergol, B. and M. Omladič
2015. What can Wikipedia and Google tell us about stock prices under diferent market regimes? *Ars Mathematica Contemporanea*, 9(2).

Chadwick, A. and P. N. Howard
2009. Introduction: New directions in Internet politics research. *Routledge handbook of Internet politics*, Pp. 1–9.

Chafe, W. and J. Danielewicz
1987. *Properties of spoken and written language.* Academic Press.

Chen, D., A. Fisch, J. Weston, and A. Bordes
2017. Reading Wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Chesney, T.
2006. An empirical examination of Wikipedia's credibility. *First Monday*, 11(11).

CIA
2011. CIA – The world factbook. https://www.cia.gov/library/publications/the-world-factbook/fields/2011.html. Accessed: 6 Jan 2015.

CIA
2015. CIA – The world factbook. https://www.cia.gov/library/publications/the-world-factbook/. Accessed: 24 Sept 2015.

Ciffolilli, A.
2003. Phantom authority, self-selective recruitment and retention of members in virtual communities: The case of Wikipedia. *First Monday*, 8(12).

Ciglan, M. and K. Nørvåg
2010. Wikipop: personalized event detection system based on Wikipedia page view statistics. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, Pp. 1931–1932. ACM.

Clark, E. V.
1996. *Using language*. Cambridge University Press.

Clauson, K. A., H. H. Polen, M. N. K. Boulos, and J. H. Dzenowagis
2008. Scope, completeness, and accuracy of drug information in Wikipedia. *Annals of Pharmacotherapy*, 42(12):1814–1821.

CLDR Charts
2015. Territory-language information. Unicode Common Locale Data Repository,http://www.unicode.org/cldr/charts/latest/supplemental/territory_language_information.html. Accessed: 24 Sept 2015.

Coan, R. W.
1979. *Psychologists: Personal and theoretical pathways.* Irvington Publishers.

Coleman, M. and T. L. Liau
1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Collier, N., M. T. Pilehvar, and M. Gritta
2018. Which Melbourne? Augmenting geocoding with maps. In *ACL (1)*, Pp. 1285–1296.

Conrad, M.
2007. 2007 Presidential Address of the CHA: Public history and its discontents or history in the age of Wikipedia. *Journal of the Canadian Historical Association/Revue de la Société historique du Canada*, 18(1):1–26.

Coursey, K. and R. Mihalcea
2009. Topic identification using Wikipedia graph centrality. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, Pp. 117–120. Association for Computational Linguistics.

Crystal, D.
2000. *Language death*. Cambridge: Cambridge University Press.

Crystal, D.
2003. *English as a Global Language*, second edition. Cambridge: Cambridge University Press.

Dale, E. and J. S. Chall
1948. A formula for predicting readability: Instructions. *Educational research bulletin*, Pp. 37–54.

Dale, R.
2015. The limits of intelligent personal assistants. *Natural Language Engineering*, 21(2):325–329.

de Silva, B. and R. Compton
2014. Prediction of foreign box office revenues based on Wikipedia page activity. *arXiv preprint arXiv:1405.5924*.

Dekker, D., D. Krackhardt, and T. Snijders
2003. Multicollinearity robust QAP for multiple regression. In *1st Annual Conference of the North American Association for Computational Social and Organizational Science*, Pp. 22–25.

Dekker, D., D. Krackhardt, and T. A. Snijders
2007. Sensitivity of MRQAP tests to collinearity and autocorrelation conditions. *Psychometrika*, 72(4):563–581.

Deuze, M., A. Bruns, and C. Neuberger
2007. Preparing for an age of participatory news. *Journalism practice*, 1(3):322–338.

Devlin, M.
2015. Google and Wikipedia: Best friends forever. https://newslines.org/blog/google-and-{W}ikipedia-best-friends-forever/. Accessed: 1 Sept 2018.

Dickerson, A.
2018. Algorithmic trading of Bitcoin using Wikipedia and Google Search volume. *Available at SSRN*.

Dong, X., E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang
2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, Pp. 601–610. ACM.

Dubin, F.
1989. Situating literacy within traditions of communicative competence. *Applied Linguistics*, 10(2):171–181.

Dumont, L.
1979. The anthropological community and ideology. *Information (International Social Science Council)*, 18(6):785–817.

Duncan, J. S.
2005. *The city as text: The politics of landscape interpretation in the Kandyan Kingdom*. Cambridge University Press.

Dunn, O. J.
1961. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64.

Dylko, I. and M. McCluskey
2012. Media effects in an era of rapid technological transformation: A case of user-generated content and political participation. *Communication Theory*, 22(3):250–278.

Eckles, D. and E. Bakshy
2017. Bias and high-dimensional adjustment in observational studies of peer effects. *arXiv preprint arXiv:1706.04692*.

Editorial
2006. Britannica attacks... and we respond. *Nature*, 440(582).

Edler, D. and M. Rosvall
2013. The MapEquation software package. http://www.mapequation.org. Accessed: 24 Sept 2015.

Elder, D., R. N. Westbrook, and M. Reilly
2012. Wikipedia lover, not a hater: Harnessing Wikipedia to increase the discoverability of library resources. *Journal of web librarianship*, 6(1):32–44.

Elia, A.
2009. Quantitative data and graphics on lexical specificity and index readability: the case of Wikipedia. *RAEL: revista electrónica de lingüística aplicada*, (8):248–271.

Eom, Y.-H., P. Aragón, D. Laniado, A. Kaltenbrunner, S. Vigna, and D. L. Shepelyansky
2015. Interactions of cultures and top people of Wikipedia from ranking of 24 language editions. *PloS ONE*.

Ethnologue
2015. Ethnologue. http://www.ethnologue.com/. Accessed: 6 Jan 2015.

Fagiolo, G., J. Reyes, and S. Schiavo
2010. The evolution of the world trade web: A weighted-network analysis. *J. Evol. Econ.*, 20(4):479–514.

Fasold, R.
1984. The sociolinguistics of society: Introduction to sociolinguistics volume i. *Language in society*, 5.

Feldman-Bianco, B.
2001. Brazilians in Portugal, Portuguese in Brazil: Constructions of sameness and difference 1. *Identities. Glob. Stud.*, 8(4):607–650.

Filatova, E.
2009. Directions for exploiting asymmetries in multilingual Wikipedia. In *Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies*, Pp. 30–37. Association for Computational Linguistics.

Fischer, S.
2003. Globalization and its challenges. *American Economic Review*, Pp. 1–30.

Flanagin, A. J. and M. J. Metzger
2011. From Encyclopaedia Britannica to Wikipedia: Generational differences in the perceived credibility of online encyclopedia information. *Information, Communication & Society*, 14(3):355–374.

Fleischhacker, D., H. Paulheim, V. Bryl, J. Völker, and C. Bizer
2014. Detecting errors in numerical linked data using cross-checked outlier detection. In *International Semantic Web Conference*, Pp. 357–372. Springer.

Flesch, R.
1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Freud, S.
1933. *New introductory lectures on psycho-analysis*. Norton.

Friedman, T. L.
2000. *The Lexus And The Olive Tree: Understanding Globalization*. Macmillan.

Furet, F.
1971. Quantitative history. *Daedalus*, Pp. 151–167.

Galloway, E. and C. DellaCorte
2014. Increasing the discoverability of digital collections using Wikipedia: The Pitt experience. *Pennsylvania Libraries: Research & Practice*, 2(1):84–96.

Gallus, J.
2016. Fostering public good contributions with symbolic awards: A large-scale natural field experiment at Wikipedia. *Management Science*, 63(12):3999–4015.

Gandomi, A. and M. Haider
2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137–144.

García-Gavilanes, R., Y. Mejova, and D. Quercia
2014. Twitter ain't without frontiers: Economic, social, and cultural boundaries in international communication. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, Pp. 1511–1522, New York, NY, USA. ACM.

Geertz, C.
1973. *The interpretation of cultures: Selected essays*, volume 5019. New York: Basic books.

Geertz, C.
2008. Thick description: Toward an interpretive theory of culture. In *The Cultural Geography Reader*, Pp. 41–51. Routledge.

Gelfand, M. J., J. L. Raver, L. Nishii, L. M. Leslie, J. Lun, B. C. Lim, L. Duan, A. Almaliach, S. Ang, J. Arnadottir, et al.
2011. Differences between tight and loose cultures: A 33-nation study. *Science*, 332(6033):1100–1104.

Generous, N., G. Fairchild, A. Deshpande, S. Y. Del Valle, and R. Priedhorsky
2014. Global disease monitoring and forecasting with Wikipedia. *PLoS computational biology*, 10(11):e1003892.

Geyer, M. and C. Bright
1995. World history in a global age. *The American Historical Review*, 100(4):1034–1060.

Gieck, R., H.-M. Kinnunen, Y. Li, M. Moghaddam, F. Pradel, P. A. Gloor, M. Paasivaara, and M. P. Zylka
2016. Cultural differences in the understanding of history on Wikipedia. In *Designing Networks for Innovation and Improvisation*, Pp. 3–12. Springer.

Giles, J.
2005. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901.

Gloor, P., P. De Boer, W. Lo, S. Wagner, K. Nemoto, and H. Fuehres
2015. Cultural anthropology through the lens of Wikipedia - a comparison of historical leadership networks in the English, Chinese, and Japanese Wikipedia. In *Proceedings of the 5th International Conference on Collaborative Innovation Networks COINs15*, Tokyo, Japan.

Gluckman, P.
2016. The science–policy interface. *Science*, 353:969. http://science.sciencemag.org/content/353/6303/969. Accessed: 1 September 2018.

Goodenough, W. H.
1981. Culture, language, and society.

Graham, M., B. Hogan, R. K. Straumann, and A. Medhat
2014. Uneven geographies of user-generated information: Patterns of increasing informational poverty. *Annals of the Association of American Geographers*, 104(4):746–764.

Graham, M. and M. Zook
2013. Augmented realities and uneven geographies: Exploring the geolinguistic contours of the Web. *Environment and Planning A*, 45(1):77–99.

Gregory, D. and D. Ley
1988. Culture's geographies.

Gunning, R.
1952. The technique of clear writing.

Gupta, A. and J. Ferguson
2007. Beyond "culture": Space, identity, and the politics of difference. *Ethnographic fieldwork: an anthropological reader*, Pp. 337–346.

Gupta, V., P. J. Hanges, and P. Dorfman
2002. Cultural clusters: Methodology and findings. *Journal of world business*, 37(1):11–15.

Halavais, A. and D. Lackaff
2008. An analysis of topical coverage of Wikipedia. *Journal of computer-mediated communication*, 13(2):429–440.

Hale, S. A.
2014a. Global connectivity and multilinguals in the Twitter network. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, Pp. 833–842, New York, NY, USA. ACM.

Hale, S. A.
2014b. Multilinguals and Wikipedia editing. In *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14, Pp. 99–108, New York, NY, USA. ACM.

Hall, S.
1993. Culture, community, nation. *Cultural studies*, 7(3):349–363.

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

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

Hardy, D.
2013. The geographic nature of Wikipedia authorship. In *Crowdsourcing geographic knowledge*, Pp. 175–200. Springer.

Hardy, D., J. Frew, and M. F. Goodchild
2012. Volunteered geographic information production as a spatial process. *International Journal of Geographical Information Science*, 26(7):1191–1212.

Hartelius, E. J.
2008. *The rhetoric of expertise*. The University of Texas at Austin.

Head, A. and M. Eisenberg
2010. How today's college students use Wikipedia for course-related research. *First Monday*, 15(3).

Hecht, B. and D. Gergle
2009. Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the Fourth International Conference on Communities and Technologies*, Pp. 11–20, New York, NY, USA. ACM.

Hecht, B. and D. Gergle
2010a. 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*, CHI '10, Pp. 291–300, New York, NY, USA. ACM.

Hecht, B. J.
2013. *The mining and application of diverse cultural perspectives in user-generated content*. PhD thesis, Northwestern University.

Hecht, B. J. and D. Gergle
2010b. On the "localness" of user-generated content. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, Pp. 229–232, New York, NY, USA. ACM.

Heilman, J. M. and A. G. West
2015. Wikipedia and medicine: Quantifying readership, editors, and the significance of natural language. *Journal of medical Internet research*, 17(3).

Hennemann, S., D. Rybski, and I. Liefner
2012. The myth of global science collaboration—collaboration patterns in epistemic communities. *Journal of Informetrics*, 6(2):217–225.

Hensel, P. R.
2009. ICOW colonial history data set, version 0.4. *University of North Texas*. http://www.paulhensel.org/icowcol.html. Accessed: 2015-06-01.

Herring, S. C., J. C. Paolillo, I. Ramos-Vielba, I. Kouper, E. Wright, S. Stoerger, L. A. Scheidt, and B. Clark
2007. Language networks on LiveJournal. In *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on*, Pp. 79–79. IEEE.

Hickmann, K. S., G. Fairchild, R. Priedhorsky, N. Generous, J. M. Hyman, A. Deshpande, and S. Y. Del Valle
2015. Forecasting the 2013–2014 influenza season using Wikipedia. *PLoS computational biology*, 11(5):e1004239.

Hill BM, S. A.
2013. The Wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation. *PLoS ONE*, 8(6)(e65782).

Hinnosaar, M., T. Hinnosaar, M. E. Kummer, and O. Slivko
2017. Does Wikipedia matter? The effect of Wikipedia on tourist choices. *ZEW Discussion Papers*, 15-089.

Hofstede, G.
1980. *Culture's consequences: International differences in work-related values*. London and Beverly Hils: Sage Publications.

Hofstede, G.
1984. *Culture's consequences: International differences in work-related values*, volume 5. Sage.

Hofstede, G.
2001. *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications.

Hoijer, H.
1948. Linguistic and cultural change. *Language*, 24:335–45.

Holman Rector, L.
2008. Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. *Reference services review*, 36(1):7–22.

House, R. J., P. J. Hanges, M. Javidan, P. W. Dorfman, and V. Gupta
2004. *Culture, leadership, and organizations: The GLOBE study of 62 societies*. Sage publications.

Hubert, L. and J. Schultz
1976. Quadratic assignment as a general data analysis strategy. *Br J Math Stat Psychol*, 29(2):190–241.

Huntington, S. P.
1993. The clash of civilizations? *Foreign Affairs*, Pp. 22–49.

Hutchins, E.
1995. *Cognition in the Wild*. MIT press.

Inglehart, R. and W. E. Baker
2000. Modernization, cultural change, and the persistence of traditional values. *American sociological review*, Pp. 19–51.

Inglehart, R. F., M. Basanez, M. Basanez, A. Moreno, et al.
1998. *Human values and beliefs: A cross-cultural sourcebook*. University of Michigan Press.

Jackson, C.
2016. Using social network analysis to reveal unseen relationships in Medieval Scotland. *Digital Scholarship in the Humanities*, P. fqv070.

Jackson, D. A. and K. M. Somers
1989. Are probability estimates from the permutation model of mantel's test stable? *Canadian Journal of Zoology*, 67(3):766–769.

Jackson, P.
2012. *Maps of meaning*. Routledge.

Jang, S., M. Megawati, J. Choi, and M. Y. Yi
2015. Semi-automatic quality assessment of linked data without requiring ontology. In *NLP-DBPEDIA@ ISWC*, Pp. 45–55.

Jarvis, J.
2009. What would Google do.

Jatowt, A. and K. Tanaka
2012. Is Wikipedia too difficult?: Comparative analysis of readability of Wikipedia, Simple Wikipedia and Britannica. In *CIKM'12*, Pp. 2607–2610. ACM.

Jemielniak, D. and E. Aibar
2016. Bridging the gap between Wikipedia and academia. *Journal of the Association for Information Science and Technology*, 67(7):1773–1776.

Jensen, R.
2012. Military history on the electronic frontier: Wikipedia fights the War of 1812. *The Journal of Military History*, 76(4):523–556.

Johnson, I. L., Y. Lin, T. J.-J. Li, A. Hall, A. Halfaker, J. Schöning, and B. Hecht
2016. Not at home on the range: Peer production and the urban/rural divide. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, Pp. 13–25. ACM.

Judd, T. and G. Kennedy
2010. A five-year study of on-campus Internet use by undergraduate biomedical students. *Computers & Education*, 55(4):1564–1571.

Jung, C. G.
1951. Fundamental questions of psychotherapy. *Collected works*, 16:116.

Kaluza, P., A. Kölzsch, M. T. Gastner, and B. Blasius
2010. The complex network of global cargo ship movements. *J. R. Soc. Interface.*, 7(48):1093–1103.

Kämpf, M., E. Tessenow, D. Y. Kenett, and J. W. Kantelhardt
2015. The detection of emerging trends using Wikipedia traffic data and context networks. *PloS one*, 10(12):e0141892.

Karimi, F., L. Bohlin, A. Samoilenko, M. Rosvall, and A. Lancichinetti
2015. Mapping bilateral information interests using the activity of Wikipedia editors. *Palgrave Communications*, 1.

Kasneci, G., F. Suchanek, and G. Weikum
2006. Yago – a core of semantic knowledge.

Keegan, B., D. Gergle, and N. Contractor
2011. Hot off the Wiki: Dynamics, practices, and structures in Wikipedia's coverage of the Tōhoku Catastrophes. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, Pp. 105–113, New York, NY, USA. ACM.

Keegan, B., D. Gergle, and N. Contractor
2012. Do editors or articles drive collaboration? Multilevel statistical network analysis of Wikipedia coauthorship. In *Proc. ACM Conf. on Computer Supported Cooperative Work*, Pp. 427–436. ACM.

Keegan, B. C.
2013. A history of newswork on Wikipedia. In *Proceedings of the 9th International Symposium on Open Collaboration*, WikiSym '13, Pp. 7:1–7:10, New York, NY, USA. ACM.

Keesing, R. M.
1990. Theories of culture revisited. *Canberra Anthropology*, 13(2):46–60.

Kelly, G.
1955. *The Psychology of Personal Constructs: Vol. 2; Clinical Diagnosis and Psychotherapy*. WW Norton.

Kennedy, R., E. Forbush, B. Keegan, and D. Lazer
2015. Turning introductory comparative politics and elections courses into social science research communities using Wikipedia: Improving both teaching and research. *PS: Political Science & Politics*, 48(2):378–384.

Kensinger, S. A.
2015. Alexa and Barbie may be Siri's new rivals: They may seem creepy to some, but new voice-recognition products could change speech-language treatment. *The ASHA Leader*, 20(7):online–only.

Khadivi, P. and N. Ramakrishnan
2016. Wikipedia in the tourism industry: Forecasting demand and modeling usage behavior. In *AAAI*, Pp. 4016–4021.

Kim, S., S. Park, S. A. Hale, S. Kim, J. Byun, and A. H. Oh
2016. Understanding editing behaviors in multilingual Wikipedia. *PloS one*, 11(5):e0155305.

Kimmons, R. M.
2011. Understanding collaboration in Wikipedia. *First Monday*, 16(12).

Kiser, E. and M. Hechter
1991. The role of general theory in comparative-historical sociology. *American Journal of Sociology*, Pp. 1–30.

Kittur, A., E. H. Chi, and B. Suh
2009. What's in Wikipedia?: Mapping topics and conflict using socially annotated category structure. In *Proceedings of the SIGCHI conference on human factors in computing systems*, Pp. 1509–1512. ACM.

Kittur, A., B. Suh, B. A. Pendleton, and E. H. Chi
2007. He says, she says: conflict and coordination in Wikipedia. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, Pp. 453–462. ACM.

Klare, G. R.
1976. A second look at the validity of readability formulas. *Journal of reading behavior*, 8(2):129–152.

Kluckhohn, F. R.
1949. Dominant and substitute profiles of cultural orientations: Their significance for the analysis of social stratification. *Soc. F.*, 28:376.

Kluckhohn, F. R. and F. L. Strodtbeck
1961. Variations in value orientations.

Koltko-Rivera, M. E.
2000. *The Worldview Assessment Instrument (WAI): The development and preliminary validation of an instrument to assess world view components relevant to counseling and psychotherapy.* PhD thesis, ProQuest Information & Learning.

Koltko-Rivera, M. E.
2004. The psychology of worldviews. *Review of General Psychology*, 8(1):3.

Kornai, A.
2013. Digital language death. *PLoS ONE*, 8:52–64.

Kose, M. A. and E. O. Ozturk
2014. A world of change. *The Future Glob. Econ.*, P. 7.

Kottler, J. A. and R. J. Hazler
2001. The therapist as a model of humane. *The handbook of humanistic psychology: Leading edges in theory, research, and practice*, 355.

Krackardt, D.
1987. QAP partialling as a test of spuriousness. *Social networks*, 9(2):171–186.

Krackhardt, D.
1988. Predicting with networks: Nonparametric multiple regression analysis of dyadic data. *Social networks*, 10(4):359–381.

Kramsch, C.
1998. *Language and culture*. Oxford: Oxford University Press.

Krings, G., F. Calabrese, C. Ratti, and V. D. Blondel
2009. Urban gravity: A model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(07):L07003.

Kristoufek, L.
2013. Bitcoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. *Scientific reports*, 3:3415.

Kroeber, A. L. and C. Kluckhohn
1952. *Culture: a critical review of concepts and definitions*, volume XLVII(1). Massachusetts, USA: Peabody Museum of American Archeology and ethnology.

Krumm, J., N. Davies, and C. Narayanaswami
2008. User-generated content. *IEEE Pervasive Computing*, 7(4):10–11.

Kubiszewski, I., T. Noordewier, and R. Costanza
2011. Perceived credibility of Internet encyclopedias. *Computers & Education*, 56(3):659–667.

Kuznetsov, S.
2006. Motivations of contributors to Wikipedia. *ACM SIGCAS computers and society*, 36(2):1.

Lally, A. M. and C. E. Dunford
2007. Using Wikipedia to extend digital collections. *D-Lib magazine*, 13(5/6).

Lam, S. K. and J. Riedl
2011. The past, present, and future of Wikipedia. *Computer*, 44(3):87–90.

Lambiotte, R., V. D. Blondel, C. de Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Van Dooren
2008. Geographical dispersal of mobile communication networks. *Phys. A*, 387(21):5317–5325.

Lamont, M. and L. Thévenot, eds.
2000. *Rethinking Comparative Cultural Sociology. Repertoires of Evaluation in France and the United States*. Cambridge, UK: Cambridge University Press.

Lancichinetti, A., M. I. Sirer, J. X. Wang, D. Acuna, K. Körding, and L. A. N. Amaral
2015. High-reproducibility and high-accuracy method for automated topic classification. *Phys. Rev. X*, 5:011007.

Lardinois, F.
2016. Apple updates Siri with Twitter, Wikipedia, Bing integration, new commands and male and female voices. https://techcrunch.com/2013/06/10/apple-updates-siri-with-twitter-{W}ikipedia-bing-integration-new-commands-and-male-and-female-voice/. Accessed: 1 Sept 2018.

Latour, B.
1987. *Science in action: How to follow scientists and engineers through society*. Harvard university press.

Laufer, P., C. Wagner, F. Flöck, and M. Strohmaier
2014. Mining cross-cultural relations from Wikipedia – a study of 31 European food cultures. *arXiv preprint arXiv:1411.4484*.

Lee, D.
2012. SOPA and PIPA protests not over, says Wikipedia. *BBC News*, 19.

Lehmann, J. and L. Bühmann
2010. Ore – a tool for repairing and enriching knowledge bases. In *International Semantic Web Conference*, Pp. 177–193. Springer.

Lehmann, J., D. Gerber, M. Morsey, and A.-C. N. Ngomo
2012. Defacto-deep fact validation. In *International Semantic Web Conference*, Pp. 312–327. Springer.

Lehmann, J., R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer, et al.
2015. DBpedia – a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195.

Lévi-Strauss, C.
1990. *Mythologiques*, volume 4. University of Chicago Press.

Liao, H.-T.
2009. Conflict and consensus in the chinese version of Wikipedia. *IEEE Technology and Society Magazine*, 28(2).

Lieberman, M. D. and J. Lin
2009. You are where you edit: Locating Wikipedia contributors through edit histories. In *ICWSM*.

Lin, J.
1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Liu, J. H., R. Goldstein-Hawes, D. Hilton, L.-L. Huang, C. Gastardo-Conaco, E. Dresler-Hawke, F. Pittolo, Y.-Y. Hong, C. Ward, and S. Abraham
2005. Social representations of events and people in world history across 12 cultures. *Journal of Cross-Cultural Psychology*, 36(2):171–191.

Liu, R., A. Agrawal, W.-k. Liao, and A. Choudhary
2014. Enhancing financial decision-making using social behavior modeling. In *Proceedings of the 8th Workshop on Social Network Mining and Analysis*, P. 13. ACM.

Lucassen, T., R. Dijkstra, and J. M. Schraagen
2012. Readability of Wikipedia. *First Monday*.

Lucassen, T. and J. M. Schraagen
2010. Trust in Wikipedia: How users trust information from an unknown source. In *Proceedings of the 4th workshop on Information credibility*, Pp. 19–26. ACM.

Luyt, B.
2011. The nature of historical representation on Wikipedia: Dominant or alterative historiography? *Journal of the American Society for Information Science and Technology*, 62(6):1058–1065.

Luyt, B. and D. Tan
2010. Improving Wikipedia's credibility: References and citations in a sample of history articles. *J. of the American Society for Information Science and Technology*, 61(4):715–722.

Magnus, P.
2006. Epistemology and the Wikipedia.

Malkov, A. S.
2014. The Silk Roads: A Mathematical Model. *Cliodynamics: The Journal of Quantitative History and Cultural Evolution*, 5(1).

Martin, J. L.
2010. Life's a beach but you're an ant, and other unwelcome news for the Sociology of culture. *Poetics*, 38(2):229–244.

Maslow, A. H., R. Frager, J. Fadiman, C. McReynolds, and R. Cox
1970. Motivation and personality (vol. 2).

Massa, P. and F. Scrinzi
2011. Exploring linguistic points of view of Wikipedia. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, Pp. 213–214, New York, NY, USA. ACM.

Massa, P. and F. Scrinzi
2012. Manypedia: Comparing language points of view of Wikipedia communities. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, WikiSym '12, Pp. 21:1–21:9, New York, NY, USA. ACM.

Masukume, G., L. Kipersztok, D. Das, T. M. Shafee, M. R. Laurent, and J. M. Heilman
2016. Medical journals and Wikipedia: A global health matter. *The Lancet Global Health*, 4(11):e791.

Mc Laughlin, G. H.
1969. Smog grading – a new readability formula. *Journal of reading*, 12(8):639–646.

McIver, D. J. and J. S. Brownstein
2014. Wikipedia usage estimates prevalence of influenza-like illness in the united states in near real-time. *PLoS computational biology*, 10(4):e1003581.

McLean, P.
2016. *Culture in networks*. John Wiley & Sons.

McMahon, C., I. L. Johnson, and B. J. Hecht
2017. The substantial interdependence of Wikipedia and Google: A case study on the relationship between peer production communities and information technologies. In *Proceedings of the Eleventh International AAAI Conference on Web an Social Media (ICWSM 2017)*, Pp. 142–151.

Mesgari, M., C. Okoli, M. Mehdi, F. Å. Nielsen, and A. Lanamäki
2014. "The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia. *J. Assoc. Inf. Sci. Technol.*

Messer, S. B.
1992. A critical examination of belief structures in integrative and eclectic psychotherapy.

Messner, M. and M. W. DiStaso
2013. Wikipedia versus Encyclopedia Britannica: A longitudinal analysis to identify the impact of social media on the standards of knowledge. *Mass Communication and Society*, 16(4):465–486.

Mestyán, M., T. Yasseri, and J. Kertész
2013. Early prediction of movie box office success based on Wikipedia activity big data. *PloS one*, 8(8):e71226.

Michel, J.-B., Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden
2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Miquel-Ribé, M. and D. Laniado
2018. Wikipedia culture gap: Quantifying content imbalances across 40 language editions. *Frontiers in Digital Humanities*, 5:12.

Mitchell, D.
1995. There's no such thing as culture: Towards a reconceptualization of the idea of culture in Geography. *Transactions of the Institute of British Geographers*, Pp. 102–116.

Moat, H. S., C. Curme, A. Avakian, D. Y. Kenett, H. E. Stanley, and T. Preis
2013. Quantifying Wikipedia usage patterns before stock market moves. *Scientific reports*, 3:1801.

Moat, H. S., C. Curme, H. E. Stanley, and T. Preis
2014. Anticipating stock market movements with Google and Wikipedia. In *Nonlinear phenomena in complex systems: From nano to macro scale*, Pp. 47–59. Springer.

Mocanu, D., A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani
2013. The Twitter of Babel: Mapping world languages through microblogging platforms. *PloS ONE*.

Müller, C. and I. Gurevych
2008. Using Wikipedia and Wiktionary in domain-specific information retrieval. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, Pp. 219–226. Springer.

Müllner, D.
2011. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*.

Murray, H.
2018. More than 2 billion pairs of eyeballs: Why aren't you sharing medical knowledge on Wikipedia? *BMJ evidence-based medicine*, Pp. bmjebm–2018.

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

Newman, M.
2012. Interactive: Mapping the world's friendships. *See http://www. facebookstories. com/stories/1574/(accessed 17 March 2014)*.

Nguyen, D., A. Overwijk, C. Hauff, D. R. Trieschnigg, D. Hiemstra, and F. De Jong
2008. Wikitranslate: Query translation for cross-lingual information retrieval using only Wikipedia. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, Pp. 58–65. Springer.

Nickel, M., K. Murphy, V. Tresp, and E. Gabrilovich
2016. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33.

Nisbett, R.
2004. *The geography of thought: How Asians and Westerners think differently... and why*. Simon and Schuster.

Nye Jr, J. S.
2004. *Power in the global information age: From realism to globalization*. Routledge.

Okazaki, S. and S. Sue
1995. Methodological issues in assessment research with ethnic minorities. *Psychological Assessment*, 7(3):367.

Okoli, C., M. Mehdi, M. Mesgari, F. Nielsen, and A. Lanamäki
2012. The people's encyclopedia under the gaze of the sages: A systematic review of scholarly research on Wikipedia.

Okoli, C., M. Mehdi, M. Mesgari, F. Å. Nielsen, and A. Lanamäki
2014. Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership. *Journal of the Association for Information Science and Technology*, 65(12):2381–2403.

Okoli, C. and K. Schabram
2009. Protocol for a systematic literature review of research on the Wikipedia. In *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, P. 73. ACM.

O'Reilly, Tim
2005. What is web 2.0. http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html. Accessed: 1 May 2018.

Overman, H. G., S. Redding, and A. Venables
2003. *The Economic Geography Of Trade, Production And Income: A Survey Of Empirics*. Blackwell Publishing.

Oz, A.
2012. Legitimacy and efficacy: The blackout of Wikipedia. *First Monday*, 17(12).

Padgett, J. F. and C. K. Ansell
1993. Robust action and the rise of the Medici, 1400-1434. *American Journal of Sociology*, Pp. 1259–1319.

Palotti, J. R. d. M., G. Zuccon, and A. Hanbury
2015. The influence of pre-processing on the estimation of readability of Web documents. In *CIKM'15*, Pp. 1763–1766. ACM.

Pan, R. K., K. Kaski, and S. Fortunato
2012. World citation and collaboration networks: Uncovering the role of Geography in Science. *Sci. Rep.*, 2.

Paul, G.
1987. There ain't no black in the union jack: The cultural politics of race and nation.

Paulheim, H.
2014. Identifying wrong links between datasets by multi-dimensional outlier detection. In *WoDOOM*, Pp. 27–38.

Paulheim, H.
2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, 8(3):489–508.

Paulheim, H. and C. Bizer
2014. Improving the quality of linked data using statistical distributions. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 10(2):63–86.

Penn
2017. The University of Pennsylvania (Penn) Treebank tag-set . http://www.comp.leeds.ac.uk/amalgam/tagsets/upenn.html Accessed: 16 May 2017.

Pentzold, C.
2009. Fixing the floating gap: The online encyclopaedia Wikipedia as a global memory place. *Memory Studies*, 2(2):255–272.

Pentzold, C., E. Weltevrede, M. Mauri, D. Laniado, A. Kaltenbrunner, and E. Borra
2017. Digging Wikipedia: The online encyclopedia as a digital cultural heritage gateway and site. *Journal on Computing and Cultural Heritage (JOCCH)*, 10(1):5.

Pepper, S. C.
1942. *World hypotheses: A study in evidence*. Univ of California Press.

Péron, Y., F. Raimbault, G. Ménier, and P.-F. Marteau
2011. On the detection of inconsistencies in RDF data sets and their correction at ontological level.

Perrin, J. M., H. Winkler, K. Daniel, S. Barba, and L. Yang
2017. Know your crowd: A case study in digital collection marketing. *The Reference Librarian*, 58(3):190–201.

Peterson, M. F. and P. B. Smith
1997. Does national culture or ambient temperature explain cross-national differences in role stress? No sweat! *Academy of Management Journal*, 40(4):930–946.

Pew Research Center
2010. Religious diversity index scores by country. http://www.pewforum.org/2014/04/04/religious-diversity-index-scores-by-country/. Accessed: 24 Sept 2015.

Pfeil, U., P. Zaphiris, and C. S. Ang
2006. Cultural differences in collaborative authoring of Wikipedia. *Journal of Computer-Mediated Communication*, 12(1):88–113.

Pfister, D. S.
2011. Networked expertise in the era of many-to-many communication: On Wikipedia and invention. *Social Epistemology*, 25(3):217–231.

Pinker, S.
1994. The language instinct: How the mind creates language (New York: William Morrow).

Platt, E., R. Bhargava, and E. Zuckerman
2015. The international affiliation network of YouTube trends.

Ponza, M., P. Ferragina, and F. Piccinno
2018. Swat: A system for detecting salient Wikipedia entities in texts. *arXiv preprint arXiv:1804.03580*.

Potter, N.
2012. Wikipedia blackout: Websites Wikipedia, Reddit, others go dark Wednesday to protest SOPA, PIPA. *ABC news*.

Potthast, M., B. Stein, and R. Gerling
2008. Automatic vandalism detection in Wikipedia. In *European Conference on Information Retrieval*, Pp. 663–668. Springer.

Powell, A.
2012. Assessing the influence of online activism on internet policy-making: The case of SOPA/PIPA and ACTA. *Unreviewed preprint article on Social Science Research Network (SSRN) online archive/database*.

Preoţiuc-Pietro, D., S. Volkova, V. Lampos, Y. Bachrach, and N. Aletras
2015. Studying user income through language, behaviour and affect in social media. *PloS one*, 10(9):e0138717.

Priedhorsky, R., D. Osthus, A. R. Daughton, K. R. Moran, N. Generous, G. Fairchild, A. Deshpande, and S. Y. Del Valle
2017. Measuring global disease with Wikipedia: Success, failure, and a research agenda. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, Pp. 1812–1834. ACM.

Radinsky, K., E. Agichtein, E. Gabrilovich, and S. Markovitch
2011. A word at a time:Computing word relatedness using temporal semantic analysis. In *Proc. Internat. Conf. on World Wide Web*, Pp. 337–346. ACM.

Ramakrishnan, R. and A. Tomkins
2007. Toward a peopleweb. *Computer*, (8):63–72.

Rasberry, L.
2014. Wikipedia: What it is and why it matters for healthcare. *BMJ: British Medical Journal (Online)*, 348.

Rassool, N.
1998. Postmodernity, cultural pluralism and the nation-state: Problems of language rights, human rights, identity and power. *Language Sciences*, 20(1):89–99.

Reavley, N. J., A. J. Mackinnon, A. J. Morgan, M. Alvarez-Jimenez, S. E. Hetrick, E. Killackey, B. Nelson, R. Purcell, M. B. Yap, and A. F. Jorm
2012. Quality of information sources about mental disorders: A comparison of Wikipedia with centrally controlled Web and printed sources. *Psychological medicine*, 42(08):1753–1762.

Ribé, M. M. and H. Rodríguez
2011. Cultural configuration of Wikipedia: measuring autoreferentiality in different languages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, Pp. 316–322.

Rinser, D., D. Lange, and F. Naumann
2013. Cross-lingual entity matching and infobox alignment in Wikipedia. *Information Systems*, 38(6):887–907.

Risse, T.
2001. A European identity? Europeanization and the evolution of nation-state identities. In *Transforming Europe: Europeanization and Domestic Change*, M. G. Cowles, J. A. Caporaso, and T. Risse-Kappen, eds. Cornell University Press Ithaca, NY.

Rogers, R., E. Sendijarevic, et al.
2012. Neutral or national point of view? A comparison of Srebrenica articles across Wikipedia's language versions. *Proc. Wikipedia Academy*.

Ronen, S., B. Gonçalves, K. Z. Hu, A. Vespignani, S. Pinker, and C. A. Hidalgo
2014. Links that speak: The global language network and its association with global fame. *PNAS*.

Rosenzweig, R.
2006. Can history be open source? Wikipedia and the future of the past. *The Journal of American History*, 93(1):117–146.

Rosenzweig, R. and D. P. Thelen
1998. *The presence of the past: Popular uses of history in American life*, volume 2. Columbia University Press.

Rosvall, M., D. Axelsson, and C. T. Bergstrom
2010. The map equation. *The European Physical Journal Special Topics*, 178(1):13–23.

Rosvall, M. and C. T. Bergstrom
2008. Maps of random walks on complex networks reveal community structure. *Proc. Natl Acad. Sci. USA*, 105(4):1118–1123.

Rovira, D. P., J.-C. Deschamps, and J. W. Pennebaker
2006. The social psychology of history: Defining the most important events of the last 10, 100, and 1000 years. *Psicología Política*, (32):15–32.

Rüsen, J.
1996. Some theoretical approaches to intercultural comparative historiography. *History and Theory*, Pp. 5–22.

Samoilenko, A.
2017. Multilingual historical narratives on Wikipedia. Version 1.

Samoilenko, A., F. Karimi, D. Edler, J. Kunegis, and M. Strohmaier
2016. Linguistic neighbourhoods: Explaining cultural borders on Wikipedia through multilingual co-editing activity. *EPJ Data Science*, 5:9.

Samoilenko, A., F. Lemmerich, K. Weller, M. Zens, and M. Strohmaier
2017. Analysing timelines of national histories across Wikipedia editions: A comparative computational approach. In *Proceedings of the Eleventh International AAAI Conference on Web an Social Media (ICWSM 2017)*, Pp. 210–219.

Samoilenko, A., F. Lemmerich, M. Zens, M. Jadidi, M. Génois, and M. Strohmaier
2018. (Don't) mention the war: A comparison of Wikipedia and Britannica articles on national histories. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*, P. 843–852.

Samoilenko, A. and T. Yasseri
2014. The distorted mirror of Wikipedia: A quantitative analysis of Wikipedia coverage of academics. *EPJ Data Science*, 3(1):1–11.

Sapir, E.
1921. Language.

Schaffer, B. S. and C. M. Riordan
2003. A review of cross-cultural methodologies for organizational research: A best-practices approach. *Organizational research methods*, 6(2):169–215.

Schich, M., C. Song, Y.-Y. Ahn, A. Mirsky, M. Martino, A.-L. Barabási, and D. Helbing
2014. A network framework of cultural history. *Science*, 345(6196):558–562.

Schneider, D. M. and D. M. Schneider
1980. *American kinship: A cultural account*. University of Chicago Press.

Schwartz, H. A., J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al.
2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Schwartz, S. H.
1994. Beyond individualism/collectivism: New cultural dimensions of values.

Senter, R. and E. A. Smith
1967. Automated readability index. Technical report, DTIC Document.

Serrano, M. Á., M. Boguñá, and A. Vespignani
2007. Patterns of dominant flows in the world trade web. *J. Econ. Interac. Coor.*, 2(2):111–124.

Serrano, M. Á., M. Boguñá, and A. Vespignani
2009. Extracting the multiscale backbone of complex weighted networks. *Proc. Natl Acad. Sci. USA*, 106(16):6483–6488.

Shafee, T., D. Mietchen, and A. I. Su
2017. Academics can help shape Wikipedia. *Science*, 357(6351):557–558.

Shao, G.
2009. Understanding the appeal of user-generated media: a uses and gratification perspective. *Internet Research*, 19(1):7–25.

Sharpe, J. D., R. S. Hopkins, R. L. Cook, and C. W. Striley
2016. Evaluating Google, Twitter, and Wikipedia as tools for influenza surveillance using Bayesian change point analysis: A comparative analysis. *JMIR public health and surveillance*, 2(2).

Silverwood-Cope, S.
2012. Wikipedia: Page one of Google UK for 99% of searches. Intelligent positioning. http://www.intelligentpositioning.com/blog/2012/02/{W}ikipedia-page-one-of-google-uk-for-99-of-searches/. Accessed 16 May 2013.

Silvia-Fuenzalida, I.
1949. Ethnolinguistics and the study of culture. *American Anthropologist*, 51(3):446–56.

Simini, F., M. C. Gonzalez, A. Maritan, and A.-L. Barabasi
2012. A universal model for mobility and migration patterns. *Nature*, 484:96–100.

Sindbæk, S. M.
2007. The small world of the Vikings: networks in early medieval communication and exchange. *Norwegian Archaeological Review*, 40(1):59–74.

Singelis, T. M.
1994. The measurement of independent and interdependent self-construals. *Personality and social psychology bulletin*, 20(5):580–591.

Singer, P., D. Helic, A. Hotho, and M. Strohmaier
2015. Hyptrails: A Bayesian approach for comparing hypotheses about human trails. In *24th International World Wide Web Conference (WWW2015)*, Firenze, Italy. ACM.

Singhal, A.
2012. Introducing the Knowledge Graph: Things, not strings. https://www.blog.google/products/search/introducing-knowledge-graph-things-not/. Accessed: 1 Sept 2018.

Slattery, S. P.
2009. Edit this page: the socio-technological infrastructure of a Wikipedia article. In *Proceedings of the 27th ACM international conference on Design of communication*, Pp. 289–296. ACM.

Sliger Krause, R., J. Rosenzweig, and P. Victor Jr
2017. Out of the vault: Developing a Wikipedia edit-a-thon to enhance public programming for university archives and special collections. *Journal of Western Archives*, 8(1):3.

Smith, P. B., S. Dugan, and F. Trompenaars
1996. National culture and the values of organizational employees: A dimensional analysis across 43 nations. *Journal of cross-cultural psychology*, 27(2):231–264.

Sorg, P. and P. Cimiano
2012. Exploiting Wikipedia for cross-lingual and multilingual information retrieval. *Data & Knowledge Engineering*, 74:26–45.

Spoerri, A.
2007. What is popular on Wikipedia and why? *First Monday*, 12(4).

State, B., P. Park, I. Weber, and M. Macy
2015. The mesh of civilizations in the global network of digital communication. *PLoS ONE*.

Street, B.
1993. Culture is a verb: Anthropological aspects of language and cultural process. *Language and culture*, Pp. 23–43.

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

Subramanian, A. and S.-J. Wei
2007. The WTO promotes trade, strongly but unevenly. *Journal of international Economics*, 72(1):151–175.

Sumi, R., T. Yasseri, A. Rung, A. Kornai, and J. Kertész
2011. Edit wars in Wikipedia. *2011 IEEE Third International Conference on Social Computing (SocialCom)*, Pp. 724–727.

Sun, E. and V. Iyer
2013. Under the hood: The entities graph. https://www.facebook.com/notes/facebook-engineering/under-the-hood-the-entities-graph/10151490531588920. Accessed: 1 Sept 2018.

Sunstein, C. R.
2006. *Infotopia: How many minds produce knowledge*. Oxford University Press.

Swidler, A.
1986. Culture in action: Symbols and strategies. *American sociological review*, Pp. 273–286.

Szajewski, M.
2013. Using Wikipedia to enhance the visibility of digitized archival assets. *D-Lib Magazine*, 19(3/4).

Tägil, S.
1995. *Ethnicity And Nation Building In The Nordic World*. SIU Press.

Taras, V., J. Rowney, and P. Steel
2009. Half a century of measuring culture: Review of approaches, challenges, and limitations based on the analysis of 121 instruments for quantifying culture. *Journal of International Management*, 15(4):357–373.

Taras, V. and P. Steel
2005. Cross-cultural differences and dynamics of cultures over time: A meta-analysis of hofstede's taxonomy. In *Academy of Management Meeting, August*, Pp. 5–10.

Taras, V., P. Steel, and B. L. Kirkman
2016. Does country equate with culture? Beyond Geography in the search for cultural boundaries. *Management International Review*, 56(4):455–487.

Teplitskiy, M., G. Lu, and E. Duede
2017. Amplifying the impact of open access: Wikipedia and the diffusion of science. *Journal of the Association for Information Science and Technology*, 68(9):2116–2127.

Thakker, D., S. Karanasios, E. Blanchard, L. Lau, and V. Dimitrova
2017. Ontology for cultural variations in interpersonal communication: Building on theoretical models and crowdsourced knowledge. *Journal of the Association for Information Science and Technology*, 68(6):1411–1428.

The Wikimedia Incubator
2015. Requests for new languages. https://meta.wikimedia.org/wiki/Requests_for_new_languages. Accessed: 24 Sept 2015.

Thompson, N. and D. Hanley
2018. Science is shaped by Wikipedia: evidence from a randomized control trial. *MIT Sloan School of Management Working Paper 5238*.

Thornton, R. J.
1987. *Culture: A contemporary definition*.

Töpper, G., M. Knuth, and H. Sack
2012. DBpedia ontology enrichment for inconsistency detection. In *Proceedings of the 8th International Conference on Semantic Systems*, Pp. 33–40. ACM.

Török, J., G. Iñiguez, T. Yasseri, M. San Miguel, K. Kaski, and J. Kertész
2013. Opinions, conflicts, and consensus: modeling social dynamics in a collaborative environment. *Phys. Rev. Lett.*, 110(8):088701.

Tumminello, M., S. Miccichè, F. Lillo, J. Piilo, and R. N. Mantegna
2011. Statistically validated networks in bipartite complex systems. *PLoS ONE*, 6(3):e17994.

Turchin, P.
2011. Toward cliodynamics–an analytical, predictive science of history. *Cliodynamics*, 2(1).

Turchin, P., T. E. Currie, E. A. Turner, and S. Gavrilets
2013. War, space, and the evolution of Old World complex societies. *Proceedings of the National Academy of Sciences*, 110(41):16384–16389.

Tylor, S. E. B.
1871. Primitive culture: Researches into the development of mythology, philosophy, religion, language, art and custom. 2 volumes. *London: J. Murray*.

Usunier, J.-C. and J. Lee
2005. *Marketing Across Cultures*. Pearson Education.

Van Dijck, J.
2009. Users like you? Theorizing agency in user-generated content. *Media, culture & society*, 31(1):41–58.

Vickery, G. and S. Wunsch-Vincent
2007. *Participative Web and user-created content: Web 2.0 wikis and social networking*. Organization for Economic Cooperation and Development (OECD).

Vincent, N., I. Johnson, and B. Hecht
2018. Examining Wikipedia with a broader lens: Quantifying the value of Wikipedia's relationships with other large-scale online communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, P. 566. ACM.

Voegelin, C. F. and Z. S. Harris
1945. Linguistics in ethnology. *Southwestern Journal of Anthropology*, 1:455–65.

Vrandečić, D. and M. Krötzsch
2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Warncke-Wang, M., A. Uduwage, Z. Dong, and J. Riedl
2012. In search of the ur-Wikipedia: universality, similarity, and translation in the Wikipedia inter-language link network. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, P. 20. ACM.

Wei, P. and N. Wang
2016. Wikipedia and stock return: Wikipedia usage pattern helps to predict the individual stock movement. In *Proceedings of the 25th International Conference Companion on World Wide Web*, Pp. 591–594. International World Wide Web Conferences Steering Committee.

Weller, K., R. Dornstädter, R. Freimanis, R. N. Klein, and M. Perez
2010. Social software in academia : Three studies on users' acceptance of Web 2.0 services. In *Proceedings Web Science Conf*, Pp. 26–27.

Welser, H. T., D. Cosley, G. Kossinets, A. Lin, F. Dokshin, G. Gay, and M. Smith
2011. Finding social roles in Wikipedia. In *Proc. iConference*, Pp. 122–129. ACM.

Wertsch, J. V.
2002. *Voices of collective remembering: Test*. Cambridge University Press.

West, J. and J. L. Graham
2004. A linguistic-based measure of cultural distance and its relationship to managerial values. *MIR: Management International Review*, Pp. 239–260.

West, R., I. Weber, and C. Castillo
2012. Drawing a data-driven portrait of Wikipedia editors. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, P. 3. ACM.

Whorf, B. L.
1940. Science and linguistics. *Technology Review*, 42(6):229–231.

Wienand, D. and H. Paulheim
2014. Detecting incorrect numerical data in DBpedia. In *European Semantic Web Conference*, Pp. 504–518. Springer.

Wikimedia
2015. Wikimedia Tool Labs. https://wikitech.wikimedia.org/wiki/Main_Page. Accessed: 1 Sept 2016.

Wikipedia
2014. Wikipedia: Size of Wikipedia. http://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia. Accessed: 6 Jan 2015.

Wikipedia
2015. Wikipedia: Notability. http://en.wikipedia.org/wiki/Wikipedia:Notability. Accessed: 24 Sept 2015.

Wikipedia
2016. Wikipedia: List of Wikipedias. http://en.wikipedia.org/wiki/List_of_Wikipedias. Accessed: 1 Sept 2016.

WRD
2015. World Religion Database. http://www.worldreligiondatabase.org//. Accessed: 6 Jan 2015.

WTO
2015. World Trade Organisation - Trade and Tariff Data. https://www.wto.org/english/res_e/statis_e/statis_e.htm. Accessed: 6 Jan 2015.

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

Xu, G., Z. Wu, G. Li, and E. Chen
2015. Improving contextual advertising matching by using Wikipedia thesaurus knowledge. *Knowledge and Information Systems*, 43(3):599–631.

Xu, S. X. and X. M. Zhang
2013. Impact of Wikipedia on market information environment: Evidence on management disclosure and investor reaction. *Mis Quarterly*, 37(4):1043–1068.

Yang, H.-L. and C.-Y. Lai
2010. Motivations of Wikipedia content contributors. *Computers in human behavior*, 26(6):1377–1383.

Yao, L., Y. Zhang, B. Wei, L. Li, F. Wu, P. Zhang, and Y. Bian
2016. Concept over time: the combination of probabilistic topic model with Wikipedia knowledge. *Expert Systems with Applications*, 60:27–38.

Yasseri, T. and J. Bright
2014. Can electoral popularity be predicted using socially generated big data? *it-Information Technology*, 56(5):246–253.

Yasseri, T. and J. Bright
2016. Wikipedia traffic data and electoral prediction: Towards theoretically informed models. *EPJ Data Science*, 5(1):22.

Yasseri, T., A. Spoerri, M. Graham, and J. Kertész
2014. The most controversial topics in Wikipedia: A multilingual and geographical analysis. In *Global Wikipedia: International and Cross-Cultural Issues in Online Collaboration*, P. Fichman and N. Hara, eds., Pp. 25–48. Maryland: Rowman & Littlefield Publishers, Inc.

Yasseri, T., R. Sumi, and J. Kertész
2012a. Circadian patterns of Wikipedia editorial activity: A demographic analysis. *PloS one*, 7(1):e30091.

Yasseri, T., R. Sumi, A. Rung, A. Kornai, and J. Kertész
2012b. Dynamics of conflicts in Wikipedia. *PloS one*, 7(6):e38869.

Zhang, W., D. Wang, G.-R. Xue, and H. Zha
2012. Advertising keywords recommendation for short-text Web pages using Wikipedia. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(2):36.

Zhang, X., Y. Yu, H. Li, and Z. Lin
2016. Sentimental interplay between structured and unstructured user-generated contents: an empirical study on online hotel reviews. *Online Information Review*, 40(1):119–145.

Zhu, H., R. E. Kraut, and A. Kittur
2013a. Effectiveness of shared leadership in Wikipedia. *Human factors*, 55(6):1021–1043.

Zhu, H., A. Zhang, J. He, R. E. Kraut, and A. Kittur
2013b. Effects of peer feedback on contribution: a field experiment in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Pp. 2253–2262. ACM.

Zickuhr, K. and L. Rainie
2011. Wikipedia, past and present. http://pewinternet.org/Reports/2011/Wikipedia.aspx. Accessed: 8 Jul 2013.

Zlatić, V., M. Božičević, H. Štefančić, and M. Domazet
2006. Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physical Review E*, 74(1):016115.

Zooniverse.org
2018. All publications. https://www.zooniverse.org/about/publications. Accessed: 1 May 2018.

Zweig, K. A. and M. Kaufmann
2011. A systematic approach to the one-mode projection of bipartite graphs. *Social Network Analysis and Mining*, 1(3):187–218.