

NORTHWESTERN UNIVERSITY

The Mining and Application of Diverse Cultural Perspectives in User-Generated Content

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL  
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Electrical Engineering and Computer Science

By

Brent Jaron Hecht

EVANSTON, ILLINOIS

March 2013

© Copyright by Brent Hecht 2013

All Rights Reserved

## ABSTRACT

The Mining and Application of Diverse Cultural Perspectives in User-Generated Content

Brent Hecht

Wikipedia articles, tweets, and other forms of user-generated content (UGC) play an essential role in the experience of the average Web user. Outside the public eye, UGC has become equally indispensable as a source of world knowledge for systems and algorithms that help us make sense of big data. In this thesis, we demonstrate that UGC reflects the cultural diversity of its contributors to a previously unidentified extent, and that this diversity has important implications for Web users and existing UGC-based technologies. Focusing on Wikipedia, Flickr, and Twitter, we show how UGC diversity can be extracted and measured using techniques from artificial intelligence and geographic information science. Finally, through two novel applications – Omnipedia and Atlasify – we highlight the exciting potential for a new class of technologies enabled by the ability to harvest diverse perspectives from UGC.

## ACKNOWLEDGEMENTS

I use the term “we” instead of “I” in this thesis not only because this is the standard in my field but also because without the support and collaboration of dozens of my colleagues, the research below would not have been possible. The people who have contributed to the research in this thesis and my research that motivated it include:

Darren Gergle, Doug Downey, Patti Bao, Johannes Schöning, Ed Chi, Lichan Hong, Bongwon Suh, Samuel Carton, Mahmood Quaderi, Martin Raubal, Mike Horn, Emily Moxley, Antonio Krüger, Brad Weinberger, Jeffrey Geiger, Federico Paredes, Candace Brown, Liz Gerber, Kathleen Geraghty, and Kristina Rodriguez

I have also been fortunate enough to have benefited from valuable conversations with: Eytan Adar, Jaime Teevan, Merrie Morris, Krzysztof Janowicz, Nada Petrović, Lauren Scissors, Alan Clark, Brian Keegan, Brooke Foucault Welles

I would also like to specifically thank my committee – Darren Gergle, Doug Downey, Mike Horn, and Eytan Adar – who have been beyond helpful in the development of this thesis.

## DEDICATION

This thesis is dedicated to Stephanie Hille and my parents, Ellen and Barry Hecht. Steph, Mom, and Dad, without your support over the years, none of the pages below would have been written.

## Table of Contents

1 Introduction.....	12
2 Related Work.....	16
2.1 User-Generated Content.....	16
2.2 Motivation from the Social Sciences.....	22
2.2.1 Psycholinguistics.....	22
2.2.2 Linguistics.....	26
2.2.3 Geography.....	29
3 Cultural Contextualization in the Language Editions of Wikipedia.....	33
3.1 Related Work.....	37
3.1.1 Multilingual Wikipedia: Process.....	37
3.1.2 Multilingual Wikipedia: Content.....	40
3.1.3 Other Types of Cultures and Wikipedia.....	48
3.1.4 Language-defined Cultures and non-Wikipedia User-Generated Content.....	51
3.2 Parsing, Extraction, and Wikipedia Resources.....	52
3.2.1 Parsing and Extraction Process.....	53
3.2.2 Wikipedia Resources.....	56
3.2.3 Results of the Parsing and Extraction Process.....	72
3.3 Concept Alignment.....	78
3.3.1 The Conceptualign Algorithm.....	80
3.3.2 Exploring Parameters.....	81
3.4 Concept-level Diversity.....	83
3.5 Sub-concept-level Diversity.....	101

3.5.1 Sub-concept-level Methodology.....	103
3.5.1.1 Bag-of-Links Document Representation Model.....	103
3.5.1.2 Missing Links.....	105
3.5.1.3 Sub-article relationships.....	118
3.5.2 Study 1: Pairwise Sub-concept-level diversity.....	130
3.5.3 Study 2: Language-by-Language Sub-concept-level Diversity.....	134
3.5.4 Study 3: Percentage of Information in English.....	139
3.5.5 Study 4: Diversity When Controlling for Length.....	145
3.5.6 Discussion.....	147
3.6 Centrality Diversity.....	148
3.6.1 Centrality Methods.....	150
3.6.2 Centrality Diversity in the WAGs of Each Language Edition.....	151
3.6.3 Centrality and Concept-level Diversity.....	160
3.6.4 Centrality and Sub-concept-level Diversity.....	164
3.6.5 Discussion.....	172
3.7 Topic Diversity.....	175
3.7.1 Assigning Concepts to Topics.....	175
3.7.2 Concept-level Diversity by Topic.....	179
3.7.3 Sub-concept-level Diversity by Topic.....	188
3.7.4 Discussion.....	194
3.8 Diversity in the Consumption of Content.....	195
3.8.1 Content Consumption Diversity Methods.....	196
3.8.2 Basic Content Consumption Diversity.....	197

3.8.3 Content-level Diversity and Content Consumption.....	202
3.8.4 Sub-concept-level Diversity and Content Consumption.....	206
3.8.5 Discussion: Page Views vs. Centrality.....	209
3.9 Diversity over Time.....	211
3.9.1 Concept-level Diversity.....	212
3.9.2 Sub-concept-level Diversity.....	216
3.9.3 Discussion.....	218
3.10 Cultural Context and Multilingual Wikipedia Diversity.....	219
3.10.1 Methods: Mining Cultural Context.....	219
3.10.1.1 Theoretical Motivation.....	219
3.10.1.2 Connecting Wikipedia to Geography.....	221
3.10.1.3 Cultural Context Metrics .....	227
3.10.2 Results: Cultural Context in Multilingual Wikipedia.....	232
3.10.2.1 Visualizing Cultural Context in Multilingual Wikipedia.....	233
3.10.2.2 The Self-Focus Bias Ratio.....	253
3.10.3 Content vs. Consumption Self-Focus Bias.....	260
3.11 Discussion.....	273
3.11.1 Culture, Content Consumption, and the Future of Wikipedia.....	276
3.12 WikAPIdia.....	280
4 Geographic Localness Diversity in User-Generated Content.....	287
4.1 Background and Related Work.....	290
4.2 Data Preprocessing.....	291
4.2.1 Flickr.....	291

4.2.2 Wikipedia.....	292
4.2.3 The Problem of Scale.....	293
4.3 Study of Contributor Spatial Behavior.....	294
4.4 Discussion.....	297
5 Inferring Geographic Cultural Community Memberships from Tweets.....	299
5.1 Data Collection and Preprocessing.....	300
5.2 Classification model.....	301
5.3 Training and test sets.....	303
5.4 Experiments.....	304
5.5 Results.....	305
5.5.1 Country-prediction experiments.....	305
5.5.2 State-prediction experiments.....	306
5.6 Discussion.....	307
6 Implications for Existing Technologies: Semantic Relatedness Measures.....	310
6.1 Semantic Relatedness Measures.....	312
6.2 Experiment.....	314
6.2.1 Concept Sampling.....	315
6.2.2 Comparison Metrics.....	317
6.2.3 SR Measure Implementations.....	318
6.3 Results and Discussion.....	320
6.3.1 Basic Results.....	320
6.3.2 Cross-language SR versus Cross-time SR.....	323
6.3.3 Pairwise Comparisons.....	324

6.3.4 Concept Pair-by-Concept Pair Analysis.....	329
6.3.5 Other SR Distributions.....	331
6.4 Discussion.....	332
7 Omnipedia.....	334
7.1 Introduction.....	334
7.2 The Omnipedia System.....	340
7.3 Study.....	343
7.3.1 Results.....	344
7.3.2 Exploring Similarities and Differences.....	344
7.3.3 Discovering New Knowledge.....	346
7.3.4 Study Summary.....	347
7.4 Visualization Approach.....	348
7.5 Future Work and Conclusion.....	351
8 Atlasify.....	353
8.1 Related Work.....	361
8.2 Explicit Spatialization.....	363
8.2.1 Definition of Explicit Spatialization.....	363
8.2.2 Relationship to Traditional Spatialization.....	366
8.2.3 Spatiotagging.....	366
8.2.4 User-defined Reference Systems.....	367
8.2.5 Spatial Information Retrieval.....	368
8.3 Explanatory Semantic Relatedness Measures (SR+E).....	372
8.3.1 Adding Explanations to SR Measures.....	373

8.3.2 WikiRelate Explanations.....	374
8.3.3 MilneWitten, OutlinkOverlap, and WAGDirect Explanations.....	376
8.3.4 Explicit Semantic Analysis Explanations.....	377
8.3.5 AtlasifySR+E.....	377
8.4 Evaluation Experiments.....	378
8.4.1 SR Value Estimates.....	379
8.4.2 Explanation Ranking Experiments.....	383
8.4.2.1 Data Collection.....	383
8.4.2.2 Machine Learning.....	385
8.4.3 Quality of Mined Text.....	386
8.5 Atlasify and Cultural Context in User-Generated Content.....	388
9 Conclusion and Future Work.....	394
10 References.....	398
11 Appendices.....	414

# 1 Introduction

*“A Nazi and a non-Nazi version of the present war would have no resemblance to one another, and which of them finally gets into the history books will be decided not by evidential methods but on the battlefield.” - George Orwell ([149] via [127])*

Computing has experienced a user-generated content (UGC) revolution. Enormously popular websites such as Wikipedia and Twitter depend on their users to provide much or all of their value. These developments have been echoed in the enterprise, with businesses encouraging employees to share resources in central repositories of institutional knowledge. Moreover, as UGC has grown in importance with end users, UGC has become equally critical to areas of computer science like artificial intelligence (AI), information retrieval (IR), and natural language processing (NLP). Hundreds of products and research projects in these domains leverage UGC to provide previously unthinkable amounts of knowledge about the world and how it functions, allowing researchers to make incredible new advances (e.g. [13, 47, 185]). There is a reasonable argument to be made that UGC has become the “brains” of many computer systems.

However, despite its fundamental front-end and back-end role in computing, many questions about UGC remain unanswered. This thesis contributes to the literature on what we believe to be a particularly important set of these questions: those related to user-generated content and culture. More specifically, in this thesis, we make three overlapping contributions that shed light on the relationship between culture and UGC. First, through a series of experiments mining and measuring cultural diversity in various UGC databases, we demonstrate that user-generated content reflects the diverse cultural contexts of its contributors. Second, we show that this diversity can have important implications for existing technologies that utilize

user-generated content as a source of world knowledge. Namely, by adopting one culture's UGC as its "brains," a computer system can become biased towards that culture's viewpoint. Finally, we show that the cultural contextualization innate to user-generated content can also have positive implications for both practitioners and researchers. Specifically, we demonstrate through two novel UGC-based applications that by embracing the cultural information embedded in UGC instead of ignoring it, a whole new class of UGC-based technologies is made possible.

This thesis does not stand alone in its focus on the role of culture in user-generated content. While the literature about the final two contributions of this thesis – the effect of diverse cultural perspectives on UGC-based technologies and the development of new technologies around this diversity of perspectives – is quite sparse, there has been more work done on establishing that UGC is culturally contextualized. However, this thesis, which contains research published over the past four years, includes important contributions to this literature. Additionally, this thesis contains new work that pushes this literature forward. We are also the first to our knowledge to make a broad, holistic argument about the cultural contextualization present in UGC, tie the literature in this area together, and motivate it with relevant research in the social sciences.

There are many types of cultural communities, and UGC can be considered in the context of each of them. This thesis will primarily focus on two types of cultural communities: those defined by language and those defined by geography. While we briefly outline a project in our future work section that seeks to generalize our research to other varieties of cultural communities, language-defined and geographically-defined cultural communities comprise the bulk of the subject matter considered here.

This thesis consists of a number of research projects, some drawn from our seven peer-reviewed papers on cultural diversity and UGC [9, 78, 80–84] and some that represent new work

on the topic. These projects are presented in three sections organized according to the three contributions identified above. In section one, we present our work establishing that UGC reflects the diverse cultures of its contributors. In doing so, we introduce a series of algorithms that allow for the measurement of the extent and character of the diversity of cultural perspectives in UGC. These algorithms are then applied to three well-known UGC repositories: Wikipedia (Chapter 3), Flickr (Chapter 4), and Twitter (Chapter 5). As we detail in each chapter, our results indicate that the diversity in these repositories is extensive and, in some cases, is much greater than had been presumed in the literature.

In section two, we demonstrate that this diversity has important implications for existing technologies that rely on UGC as a source of world knowledge. Namely, we will show that UGC-based technologies can adopt the cultural viewpoints of the underlying UGC repositories, biasing their output towards the perspectives dominant in those repositories and marginalizing others. Our focus in this section is on a case study dedicated to the effect of cultural context on Wikipedia-based semantic relatedness measures, a family of algorithms used frequently in the natural language processing and artificial intelligence literatures (Chapter 6).

Finally, in the third major research section, we will show that while the cultural context in UGC presents risks for existing work, it also presents opportunities for a whole new class of technologies and research projects in which the diversity is embraced. Two applications will be presented – Omnipedia (Chapter 7) and Atlasify (Chapter 8) – both of which are demonstrations of the novel technologies that can be built by leveraging diverse perspectives mined from UGC.

The three major sections of this thesis are bookended by chapters that place our research in the context of existing work and set the stage for our research moving forward. With regard to the latter, we end this thesis with a forward-looking conclusion that summarizes our main

contributions and highlights our ongoing work in this area,(Chapter 9). With regard to the former, we begin the main portion of the thesis below with a discussion of literature related to our research from several different disciplines (Chapter 2).

## 2 Related Work

In this chapter, we present work from several disciplines that helps to place our research in the context of the existing literature. Our goal here is to describe research that is related to this thesis at a high-level. The work relevant exclusively to each specific experiment or project is considered in context in the chapters below. We begin this chapter with a discussion of the literature on user-generated content in general. Following that, we discuss work from the social sciences that motivates our research.

### 2.1 User-Generated Content

While there have been thousands of papers on user-generated content (UGC), there is no commonly agreed-upon definition of UGC [214]. The term is generally used to refer to large online repositories of information in which the individual users of the repository have contributed much or all of the content. UGC is often defined by example, with archetypal instances including Wikipedia, Twitter, Flickr, and OpenStreetMap, as well as YouTube, Amazon.com reviews, Yelp, and question-and-answer sites like Quora and Yahoo! Answers. Personal websites, e-mail and IM communication, and information that is collected from users in the background or without their knowledge (e.g. search logs, tracked locations) are typically not considered to be UGC. This loose definition is that which is implicitly adopted by most papers on user-generated content, and it is the one that is used in this thesis.

There have been some attempts to define user-generated content more specifically, although no definition has caught on widely. Krumm and colleagues [109] define UGC as content that “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.” Dhar and Chang [30] write that UGC is “the conjunction of blogs and social networking sites.” Finally, the Organization of Economic Cooperation and Development (OECD), an important multinational non-profit organization dedicated to economic development, defines UGC<sup>1</sup> as content that meets all of the following three loose requirements [214]:

- (1) It must be made available on publicly accessible webpages or limited-access sites like Facebook, and thereby excludes e-mail, instant messages, and content kept private.
- (2) It usually “reflects a certain amount of creative effort.” This includes new content and re-mixed content, examples of which are expressing ones thoughts on a blog, creating a new music video, or uploading ones photographs. The OECD notes that the amount of creative effort required “depends on context.”
- (3) It generally is created outside of professional routines and practices, with it sometimes being created by “non-professionals without the expectation of profit or renumeration.”

Alternative definitions for UGC exist in some disciplines. Most relevant to this thesis, within geography, user-generated content is often known as volunteered geographic information (VGI). Goodchild [57] defines VGI as,

“The widespread engagement of large numbers of private citizens, often with little in the way of formal qualifications, in the creation of geographic information, a function that for centuries has been reserved to official agencies.”

Goodchild is careful write that VGI is a “special case” of user-generated content, presumably that which is has been geographically-referenced to a location on or near the surface of the Earth [56]. The impact of VGI on geography has been similar to that of UGC on computer science, with Goodchild writing that VGI, “has the potential to be a significant source of

---

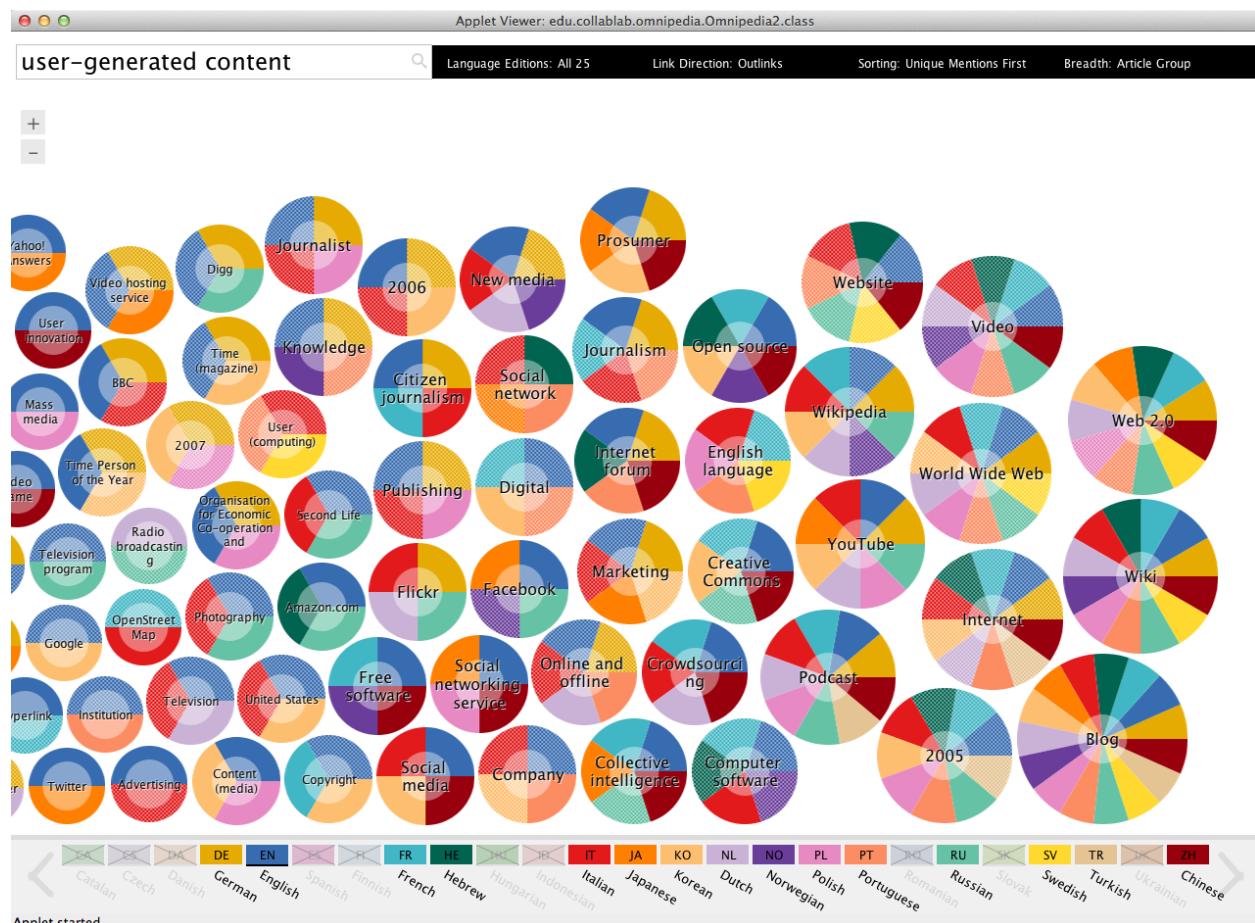
<sup>1</sup> They use the term “user-contributed content” (UCC) instead of UGC.

geographers' understanding of the surface of the Earth" [57], with that understanding being the key goal of the discipline of geography.

Finally, in a discussion of the definition of user-generated content, it would be remiss to not consider the consensus definition reached by English Wikipedia editors, which reads in part:

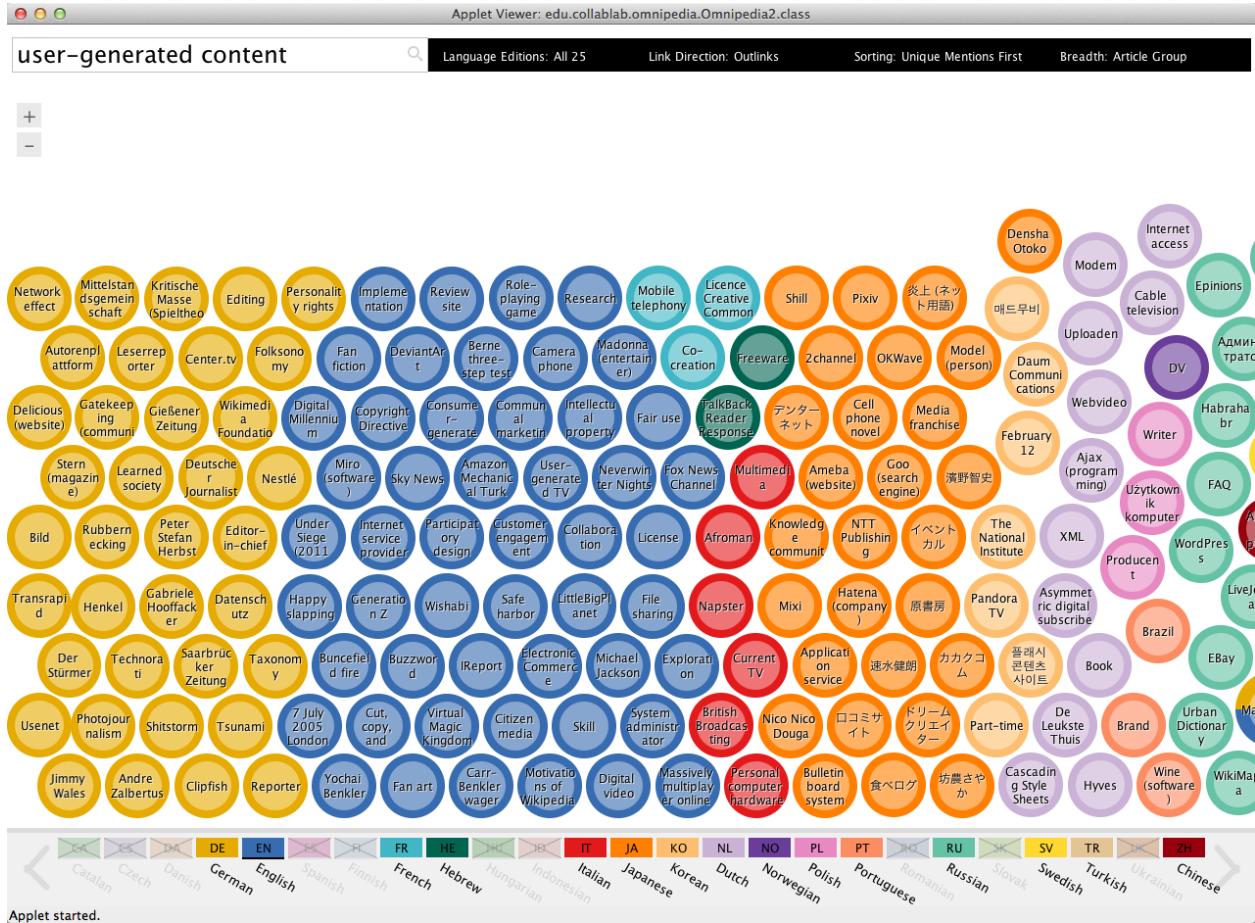
"User-generated content (UGC) covers a range of media content available in a range of modern communications technologies. It entered mainstream usage during 2005, having arisen in web publishing and new media content production circles. Its use for a wide range of applications, including problem processing, news, gossip and research, reflects the expansion of media production through new technologies that are accessible and affordable to the general public. All digital media technologies are included, such as question-answer databases, digital video, blogging, podcasting, forums, review-sites, social networking, mobile phone photography and wikis. In addition to these technologies, user-generated content may also employ a combination of open source, free software, and flexible licensing or related agreements to further reduce the barriers to collaboration, skill-building and discovery." [227]

However, as will be explained in great detail in Chapters 3, we must not only consider the English Wikipedia’s definition because each language edition of Wikipedia contains a great deal of unique information about many topics, often reflecting the corresponding language-defined culture’s perspective. Figures 2.1-a and 2.1-b are visualizations generated by Omnipedia, an application we built, of 25 different language editions’ definitions of user-generated content. Omnipedia will be explained in full in Chapter 7, but in Figures 2.1-a, one can see that blogs, wikis, Web 2.0, Wikipedia, and YouTube are discussed in many language editions’ definition of



*Figure 2.1-a: Omnipedia depicting the diverse content present in many Wikipedia language editions' coverage of the concept of user-generated content. These large and colorful dots indicate entities that are discussed in many language editions' definitions of user-generated content (e.g. blog, wiki, Wikipedia, YouTube). The smaller, less colorful dots in the figure below represent topics that are discussed in fewer language editions.*

user-generated content (the larger and more colorful the circle, the more language editions in which a given entity is discussed). In Figure 2.1-b, one can see aspects of the definition of user-generated content that occur in only a single language edition. For instance, only the Japanese Wikipedia discusses the Japanese search engine Goo in its definition of user-generated content. One can also see that the English Wikipedia's definition is incomplete with regard to certain topics, e.g. the idea of gatekeeping (German-only) and WordPress (Russian-only).



*Figure 2.1-b: Entities that are only mentioned in a single language edition of Wikipedia's definition of user-generated content, as visualized by Omnipedia. Note that many language editions mention news organizations specific to the corresponding language-defined culture. For instance, the English Wikipedia is the only one to mention Sky News and the Fox News Channel, while the German Wikipedia is the only one to mention Bild (a German tabloid) and Saarbrücker Zeitung (a regional newspaper). As we will show in Chapter 7, Omnipedia allows users to access the context of how an entity is discussed in a given language edition by clicking on the entity's circle. If we were to click on the news organizations' circles, we would find that nearly all of them are mentioned in a discussion of how traditional news institutions are adapting to the world of user-generated content.*

## 2.2 Motivation from the Social Sciences

Many disciplines in the social sciences can provide context and motivation for the findings and applications in this thesis. Each of these disciplines has at least one framework that is helpful in understanding the cultural contextualization of UGC, as well as a large body of research that predicts that it would exist. In this section, we provide a brief overview of relevant work from three such disciplines: Psycholinguistics, Linguistics, and Geography.

### 2.2.1 Psycholinguistics

Herbert Clark's well-known theory of language as joint action [24] provides motivation for the hypothesis that UGC will be culturally contextualized. Namely, when describing his central notion that common ground between communication partners is central to their patterns of communication, Clark writes “communal common ground defines cultural communities”. He continues by arguing that cultural communities are a “set of people with shared expertise [communal common ground] that other people lack,” with this shared expertise consisting of “facts, beliefs, procedures, norms, and assumptions.” “Cultural communities,” he writes, “are therefore identifiable by their [shared] expertise.”

Applied to user-generated content, Clark's theory suggests that members of each cultural community would generate different information reflective of their unique set of shared “facts, beliefs, procedures, norms, and assumptions,” resulting in a great deal of diversity in global repositories. Consider Twitter, for instance. In Chapter 5, we show that a Twitter user's geographic cultural community can be inferred just by looking at the tweets of that user. This finding can be expected from Clark's theory. Members of the Massachusetts geographic community, for example, share their extensive experience with (and love for) the Boston Red

Sox, and this fact helped our Naïve Bayes classifier understand that people who tweet about the Red Sox are likely residents of Massachusetts. If we assume that tweets are “traces” of a Twitter user’s expertise, we are effectively showing that Twitter users’ cultural communities are “identifiable by their [shared] expertise.”

Understanding UGC diversity across geographic cultural communities using Clark’s framework is quite straightforward. Applying his framework to diversity across language-defined communities is less so. Aside from some elements of the grammar and structure of a language, one cannot ascribe a great deal of shared expertise to speakers of the same language. While this is less the case for speakers of certain “nation-state” [40] languages like Slovak, Norwegian, Finnish, and so on, Spanish speakers in El Salvador, for instance, likely only share limited common ground with Spanish speakers in Spain. The same goes for English speakers in Texas and English speakers in Bangalore. How then, can we ascribe some of the diversity between, for instance, the Spanish and English Wikipedias, to culture, as it is defined by Clark?

The answer to this question lies in Clark’s discussion of the nested and correlated nature of cultural communities and their corresponding shared knowledge. For example, a member of the cultural community of American country music fans is likely also a member of the English-speaking community. As such, while the English-speaking community as a whole may share little more expertise than knowledge of syntax, phonology, and the like, the English-speaking community consists of a large number of other cultural communities, each with its own set of “facts, beliefs, procedures, norms, and assumptions.” The Spanish-speaking community also consists of its own group of “sub-communities,” which while not mutually exclusive from that of English speakers, is likely quite different. This difference is likely exacerbated by the fact that – as Clark notes – language is a vital “stratum” upon which cultural communities form. For

example, there are many fewer Spanish-speaking country music fans than English-speaking ones, and this is likely due to the fact that a relatively small proportion of Spanish speakers can understand the music's lyrics. This difference in the shared expertise of "sub-communities" is reflected in the corresponding Wikipedia language editions, with the English Wikipedia having much greater coverage of American country music than the Spanish one. The reverse is true of information that is likely to be relatively unique to the common ground of, say, Spanish-speaking geographic cultural communities (e.g. people who near the Sierra Tarahumara mountains of Mexico, which does not have an English-language Wikipedia page).

Clark's framework allows us to reasonably attribute our results that show that different cultural communities contribute different information to UGC repositories to the different shared expertise or commonly held knowledge bases of these communities. However, there are certain limitations to this attribution. For instance, there is no guarantee that a community will contribute information about certain shared expertise; German speakers may be less willing to contribute information about the Holocaust than speakers of other languages, even though they are more intimately familiar with that event than most other cultural communities<sup>2</sup>.

Another possible critique of the use of Clark's framework in this context would suggest that because certain knowledge is known to be known (the definition of a common ground) to all members of a community, this common knowledge would be considered obvious and perhaps *less* likely to appear in user-generated content. For instance, a Boston resident might tweet, "Go team!" or "Yeeeeeeeeaaaaahhhh!" in response to a Red Sox victory, rather than "Go Red Sox!" If all of this Bostonian's followers were members of his community, the "Red Sox" part of the tweet would be redundant; everyone had watched the game.

---

<sup>2</sup> However, in Chapter 7 we will see that the opposite is true in this specific case.

Why then, would we see such diversity in the information contributed from different cultural communities? Wouldn't everyone just tweet "Go team," no matter where they are, therefore destroying the predictive capability of our Naïve Bayes classifier (and nullifying the results of many of the studies in this thesis)? One explanation for why we saw, for example, people from Boston tweet terms like "Red Sox" (and South Carolinians tweet "Gamecocks" and Canadians tweet "Habs," and so on) derives from the diverse audience of the cultural communities considered in this thesis. In the Twitter context, this means that not all followers of a given Twitter user come from the exact same set of cultural communities as the Twitter user. Similarly, in the Wikipedia context, it is the job of a person from one cultural community – e.g. American country music fans – to explain a concept to speakers of the entire language-speaking community, made up all of sorts of "sub-communities."

More generally, Clark's theory suggests that people alter the information in "intercultural" communication in order to account for the differences in shared expertise between cultural communities. In other words, in order to clearly express their happiness about a Red Sox victory to all their Twitter followers, not just those of the Boston geographic cultural community (or the fan community, etc.), people use the term "Red Sox" and not "team." The same phenomenon can be said to occur in Wikipedia contributions, with the audience of these contributions likely even more diverse. This phenomenon has been experimentally verified outside the UGC context in many ways. For instance, researchers who asked for directions in Boston with Boston accents and rural Missouri accents received different sets of directions [46, 102].

To summarize, Clark argues that each cultural community has its own, unique set of shared experience, which suggests that each cultural community will contribute UGC about different topics. On the other hand, Clark's theory also intimates that when engaging in purely

intracultural communication, much of this shared experience would be difficult to detect with computational methods like those used in this paper due to lack of sufficient context. These methods are able to succeed, however, because contributors of UGC belonging to a certain cultural group know that their contributions will likely have a wide audience. They thus provide sufficient disambiguating information such that computational approaches are able to detect the shared experience of that cultural group in their contributed UGC.

### **2.2.2 Linguistics**

Linguistics provides some of the most powerful motivation for the hypothesis that UGC is culturally contextualized. According to Kramsch [108], many linguists argue that language, like that which appears in Wikipedia, Twitter, and other UGC repositories, “expresses cultural reality” [108]. Specifically, linguistics has established that language expresses the reality of two types of cultures: language-defined cultures and cultures that occur within (and sometimes across) language-defined cultures. The latter are sometimes called discourse communities while the former – people who speak the same language – are sometimes labeled speech communities [15, 108].

That culture is expressed by the language used in discourse communities is far less controversial. Kramsch [108] writes that a defining trait of discourse communities is that they talk about a unique set of topics (in addition to having a unique way to present information, a unique style of interaction, etc.). As such, these discourse communities are quite similar to Clark’s cultural communities, and are relevant to this thesis in the same fashion. For instance, in our use of tweets to geolocate users, we demonstrate in Chapter 5 that the geographic discourse community memberships of a Twitter user can be identified by the topics s/he discusses on

Twitter. Similarly, one could argue that the fact that each language edition of Wikipedia is written by members of different sets of discourse communities – defined by higher-level speech communities with which they are, in Clark’s words, “correlated” – is an important cause of the differences in content between the language editions.

The argument that additional cultural differences are introduced at the speech community level is more controversial. This claim roughly maps to the Sapir-Whorf hypothesis, or that “different people speak differently because their language offers them the different ways of expressing the world around them” [108]. While Sapir-Whorf is notoriously contentious, it is generally accepted in its weak form [108], and these cultural differences between speech communities (a.k.a. language-defined cultures) could play a role in the differences between Wikipedia language editions found in Chapter 3. For instance, the fact certain languages have words or terms for complex concepts that have no equivalent words or terms in other languages could increase the diversity of concepts covered in each language edition. That said, many of the examples of this phenomenon presented by Wierzbicka [207] are ameliorated in Wikipedia through the cross-language borrowing of the words or terms in question. For example, the Polish meat stew bigos, which has no direct translation in English, has an English Wikipedia article whose title is simply “Bigos”. The same goes for the Japanese matchmaking ritual of miai, and so on.

Perhaps more significantly, speech community diversity – through its known ability to affect semantic associations between concepts [108] – could cause some of the differences in *how* concepts are described in each language edition of Wikipedia (“sub-concept-level diversity” in Chapter 3). The fact that articles on bigos and miai exist in the English Wikipedia is no guarantee that they are linked to or discussed in the same way or at the same rate as in their

native languages. In fact, the article on *miai* in Japanese is linked to by over four times as many articles as the corresponding article in the English Wikipedia, despite the fact that the English Wikipedia has many more articles overall. Speech community differences in semantic associations could also explain some of the variation in the output of semantic relatedness algorithms (Chapter 6). However, disentangling speech community diversity’s effect from that of discourse communities is difficult and is the subject of future work rather than appearing in this thesis. To do this, qualitative techniques at a small scale (cf. [155]) may be more appropriate than the quantitative approach taken here.

In addition to motivating the UGC diversity hypothesis, linguistics also makes a strong case for utilizing corpora like large datasets of UGC for the study of cultures, essentially advocating for social scientific value of this thesis (although much future work will need to be done to fully develop this value). This case dates back at least to the work of Sapir, who wrote, “language is a symbolic guide to culture,” making it “of great assistance in the study of cultural phenomena” [175]. More recently, Wierzbicka [207] argued that the semantics of the language used by a culture can be a powerful window into the essence of the culture. She writes, “Cultural analysis can also gain important insights from linguistics, in particular from linguistic semantics, and the semantic perspective on culture is something that cultural analysis can ill afford to ignore” [207]. This thesis has a heavy focus on semantics, or “what [language] says or refers to” [108] (rather than the spelling, phonology, etc.) in all nine of its chapters.

Wierzbicka suggests several frameworks with which to analyze a culture through its use of language, some of which dovetail nicely or motivate the approaches taken here. For instance, Wierzbicka advocates for the use of term frequency as a way to understand the “cultural salience” of an underlying concept [207], an approach that is strongly echoed in our analysis of

centrality diversity (indegree) across language editions of Wikipedia (Section 3.6), our self-focus bias study (Section 3.10), and in our demonstration of the ability of tweets to identify the geographic cultural community of the corresponding Twitter user (Chapter 5). Similarly, Wierzbicka’s “principle of cultural elaboration” (e.g. the Hanunóo language of the Philippines having 90 words for rice) [207], likely plays a role in self-focus bias (Section 3.10), concept-level diversity (Section 3.4), and, in particular, “granularity” conflicts in the Wikipedia interlanguage link network (Section 3.3). Finally, Wierzbicka’s identification of “universal semantic primitives” [55, 207] has strong similarities with our identification of a “global encyclopedic core” (Section 3.4). This idea of universal semantic primitives also highlights an important latent theme in this thesis: that detecting diversity across cultures can also highlight the important similarities between cultures. In Wierzbicka’s words, “...Languages, and the ways of thinking reflected in them, exhibit both profound differences and profound similarities...the study of diversity can lead to the discovery of universals” [207].

### **2.2.3 Geography**

There are many geographic frameworks in which the results and applications found below can be understood. Some of these are specific to a single chapter and appear in that context. However, one geographic framework that is quite useful for interpreting the results at a higher level throughout this thesis is that of *mental maps*.

The key finding from the mental maps literature within geography relevant to this thesis is that mental maps of the same geographic spaces vary across culture groups. That is, members of one cultural group associate different attributes – e.g. dangerousness, likeability, stereotypes, and so on – to identical regions. For instance, analyzing hand-annotated “comfort level” maps of Los

Angeles from 215 residents of the city, Matei et al. [128] found that “people tend to perceive their own community as more secure while constantly projecting fear into the neighbor’s backyard, especially where people of another ethnicity live.” Ladd found a similar phenomenon in a smaller-scale, qualitative study of the mental maps of African-American children in Boston [110]. This study is summarized well in Gould and White’s seminal book on mental maps [62]:

“On Dave’s [one of the children in the study] map, the Mission Hill project is where the white children live, and he has drawn it as the largest, completely blank area on his map. From his taped conversation it is clear that he is physically afraid of the area and has never ventured near it. On his map the white residential area is literally *terra incognita*, while all the detail on the map is immediately around his home and school on the other side of Parker Street. Ernest [another child in the study] also puts in Parker Street dividing his area from the white Mission Hill project, and uses about a quarter of his sheet of paper to emphasize, quite unconsciously, the width of this psychological barrier.” [62]

Ladd and Matei et al.’s findings could easily have an effect on user-generated content. For instance, if a certain ethnicity with a higher likelihood to use Twitter (c.f. [71]) produces the lion’s share of tweets about a given neighborhood in which another ethnicity lives – as could be the case with commuters traveling through an inner-city district – Twitter content about that neighborhood could display a level of fear or a sense of dangerousness that may not reflect reality. Consider, then, the follow-on effect of this hypothetical Twitter bias for the many systems designed to analyze sentiment on Twitter, or for anyone who looks up that neighborhood in a Twitter application that allows for the browsing of tweets on a map. A similar phenomenon, though more latent, could appear in Wikipedia, as it is likely that certain ethnicities are extensively underrepresented in the Wikipedia editor community [230].

Another example of different cultural groups having different mental maps of the same

geographic locations comes from a study in which, utilizing a free association-based technique, Texans and Georgians were asked to describe the 48 contiguous American states using any word(s) that came to mind [62]. The researchers here found that the data from Texans and Georgians were vastly different. There were some similarities, however: in both cases, Wisconsin and Rhode Island were perceived quite negatively. An analogous study by Gould [61] found that, among Americans, one's local area is perceived quite favorably compared to the rest of the United States and that Southerners perceived the North very unfavorably (and vice versa). One can easily imagine that these differences in perspectives across various geographic cultures are reflected in hundreds of tweets on a daily basis.

The tendency to view one's own area more favorably than other areas becomes more concerning in the UGC context in light of another mental maps study by Orleans [147]. In the study, Orleans found that White (non-Latino), upper-class residents of Los Angeles had a much better understanding of the geography of Los Angeles than African-American and Latino residents. This suggests that, if these trends persist today, White (non-Latino), upper-class people are able to contribute information on more topics to UGC repositories like Wikipedia. Moreover, some have suggested<sup>3</sup> that the English Wikipedia is edited disproportionately by White Americans as opposed to people of other ethnicities and races in the United States [230]. Combining Whites' supposed greater ability to contribute to Wikipedia and their theorized greater presence in the editor community with the findings of Matei et al. and Gould, one could easily hypothesize, for instance, that Wikipedia articles about places in which Whites live will be written in a more positive light than those about places where other ethnicities live.

As one reads through this thesis, it is often easy to interpret the results as expressions of

---

<sup>3</sup> This has been suggested by the Wikipedia community, but there is, to our knowledge, no hard evidence yet.

differences in mental maps among different cultural groups, especially – although not exclusively – when we are examining geographic UGC. In Section 3.10, for instance, we show that each language edition of Wikipedia places very disproportionate emphasis on countries in which the corresponding language is predominant. If one can consider positive feelings about an area to be a predictor of the likelihood to produce UGC about that area, one can construe our research in Section 3.10 to be a modern-day expression of what these mental maps researchers found in an earlier, “smaller data” era.

### 3 Cultural Contextualization in the Language Editions of Wikipedia

As the Internet’s sixth most-popular website [4] and the top result for as many as 50 percent of queries on Google [59], Wikipedia is one of the most important repositories of user-generated content in the world. It is fitting then that we begin our investigation into the cultural contextualization of user-generated content with an in-depth study of the online encyclopedia “anyone can edit.” While most readers of this thesis will be quite familiar with the English Wikipedia, it is less commonly known that there are editions of the encyclopedia in over 280 languages [122]. The primary goal of this chapter is to determine if these language editions reflect the diverse cultural contexts of their corresponding language-defined communities.

In the computer science literature, the overwhelming assumption is that the language editions of Wikipedia are roughly identical when controlling for size, and thereby *do not* reflect the large cultural differences between their editor communities. In our work, we have called the belief that the language editions are mostly the same the *global consensus hypothesis* [82]. This hypothesis is most often manifest in its corollary, the *English-as-Superset hypothesis*, which assumes that because English is the largest language edition, it has basically all of the information contained in multilingual Wikipedia.

The English-as-Superset hypothesis is quite commonly adopted in Wikipedia-based research projects and systems. For example, DBpedia [13], a well-known Wikipedia-based ontology, for many years only considered entities that were covered in the English Wikipedia. Similarly, researchers in artificial intelligence and natural language processing who use Wikipedia information often presume that the output of their algorithms and models should be

the same, regardless of the language edition they consider (e.g. [74, 143]). Finally, efforts to translate large numbers of English Wikipedia articles into other language editions can be a somewhat extreme case of the global consensus hypothesis. The supposition behind many of these efforts (e.g. [35]) is that English is a “ground truth” language edition, and that the smaller language editions need to be fed information from English in order to “catch up”.

There is, of course, another hypothesis as to the character of the relationships between the language editions of Wikipedia. This hypothesis, which we have called the *global diversity hypothesis* in our work [82], is motivated by the social science theory in Chapter 2 and suggests that the language editions of Wikipedia *do* reflect the diverse cultural contexts of their contributors. In other words, the global diversity hypothesis posits that encyclopedic world knowledge differs extensively from language-defined culture to language-defined culture and predicts that this should cause a great deal of diversity in the represented world knowledge across multilingual Wikipedia.

In this chapter, we show again and again from many angles that there is indeed extensive diversity across the language editions of Wikipedia, and that a significant portion of this diversity is due to each language edition reflecting the cultural contexts of its contributors. Using a dataset of 25 different language editions, we demonstrate that this diversity – which has remained remarkably consistent over time – occurs both in terms of *which* concepts are covered in each language edition as well as *how* concepts are covered. We also establish that this diversity is uneven, occurring significantly more in content about some topics than others and having a different character in the core of each language edition than in the periphery.

Nearly all of the findings in this chapter represent evidence in support of the global diversity hypothesis and evidence against the global consensus and the English-as-Superset

hypotheses. An additional focus of this chapter is showing that where there *is* evidence in the content of multilingual Wikipedia for the global consensus and the English-as-Superset hypotheses, this evidence is not in alignment with the *consumption* of that content. That is, in almost every case, we see significantly more diversity in page views than in the information on pages. Below, we discuss the implications of this result for the aforementioned translation projects and for Wikidata<sup>4</sup>, a Wikimedia Foundation project that is attempting to remove some of the barriers between the language editions.

This chapter begins with an overview of work related to the cultural contextualization of the information in Wikipedia (Section 3.1). The subsequent two sections (Sections 3.2 and 3.3) are dedicated to methodology that is relevant throughout the chapter. In these sections, we highlight some of the methodological contributions we have made in order facilitate the execution of studies on multilingual Wikipedia and the building of multilingual Wikipedia-based systems. The fourth section (Section 3.4) discusses our findings that demonstrate that the set of concepts covered by each language edition varies extensively from one language edition to the next and Section 3.5 shows that even when two language editions do cover the same concept, they tend to cover that concept quite differently. Sections 3.6 and 3.7 examine the diversity found in the previous sections through the lens of article centrality and article topic, respectively.

The eighth section of this chapter (Section 3.8) shifts gears a bit and covers our findings related to the extensive variation in the popularity of subjects from language edition to language edition. In Section 3.9, we demonstrate that the amount of diversity between the language editions has barely shifted over time. Section 3.10, the last section in this chapter that reports research results, leverages theory from human geography to prove that the differences between

---

<sup>4</sup> <http://meta.wikimedia.org/wiki/Wikidata>

the language editions are at least partially due to each language edition reflecting the cultural contexts of its contributors rather than other causes. Section 3.11 is dedicated to high-level discussion of the findings in this chapter. Finally, the last section (Section 3.12) covers some details about WikAPIdia, the software library we developed to execute our multilingual Wikipedia research. We plan on releasing WikAPIdia following completion of this thesis and thus view it as an important contribution.

By the end of this chapter, the reader will have a detailed understanding of the cultural contextualization inherent to the information in multilingual Wikipedia. This is an understanding that we hope will (1) aid researchers in advancing the knowledge of this area, (2) help system builders be more aware of the potential cultural biases in their Wikipedia-based applications, and (3) remind end users of Wikipedia that the articles they read do not represent some perfect “ground truth” but rather must be understood in their cultural context.

#### *A Note on Terminology:*

Before beginning the main text of the chapter, it is important to first discuss three points related to terminology. First, in this thesis we typically refer to Wikipedia articles in the following fashion: “ARTICLE TITLE” (LANGUAGE EDITION). We only break from this standard where the language edition of an article is obvious, in which case we just use the quoted clause. Second, in the case that the meaning of a non-English title is not patently clear to an English speaker, we use footnotes or parenthetical statements that contain translations. Without exception, we do not refer to an article title in a non-English language edition with its English translation without explicitly noting in one of above ways that the title is translated.

Third, the *concept* that each article describes is not demarcated in any special fashion. For

instance, a valid sentence might be ““United States” (English) is about the United States’. The idea of concepts is a critical one to this entire chapter. While we discuss concepts in great detail in Section 3.3, it is important to note at this juncture that we define a concept in multilingual Wikipedia to be the subject of at least one article in one supported language edition. A concept can have an article describing it in anywhere from one to all 25 language editions considered here. For example, there are articles about the concept of breakfast in each of the 25 Wikipedias, e.g. “Breakfast” (English), “Frühstück” (German), “Desayuno” (Spanish), and so on.

## 3.1 Related Work

Research on multilingual Wikipedia exists on a spectrum from work that focuses on the contribution and collaboration patterns among editors (process) to work that focuses on the final product in the form of Wikipedia articles (content). While this chapter is dedicated to the analysis of content in multilingual Wikipedia (as well as consumption of that content), process is of course also relevant. As such, this section discusses both ends of this spectrum, necessarily bifurcating it into two discrete categories. We also cover in this section two additional domains relevant to the work in this chapter: (1) research on Wikipedia and cultures outside of those that are language-defined and (2) research on language-defined cultures and non-Wikipedia user-generated content.

### 3.1.1 Multilingual Wikipedia: Process

One of the first studies to examine Wikipedia from a multilingual perspective is the process-focused, small-scale, qualitative work of Pfeil and colleagues [155]. In this research, Pfeil et al. sought to understand the effect of culture on the contribution and collaboration patterns in four language editions of Wikipedia. Focusing on the English article “Game” and its corollaries in

German, French, and Japanese, they found that there were differences between language editions, and that some of these differences corresponded to Hofstede's dimensions of culture [93, 94].

Hara et al. [69] and Nemoto and Gloor [144] also used Hofstede's framework to interpret the actions of Wikipedia communities. Using English, Hebrew, Japanese, and Malay as a sample set, Hara et al. reported that large language editions and Eastern language editions had significantly more courtesy-related postings on talk pages than Western and small language editions, respectively. However, Hara et al. also found that all the language editions seemed to organize their discussions in the same way, with task-oriented messages occurring on articles' talk pages and relationship and community-related postings being made on users' talk pages. Nemoto and Gloor [144] found larger differences between the language editions, with their results suggesting that the collaboration behaviors of "egalitarian cultures" like the Scandinavian Wikipedia communities differed extensively from those of "more hierarchical" cultures like that made up by Japanese Wikipedians. For instance, while around 4% of editors are admins in the Swedish Wikipedia, only about 0.5% are in the Japanese Wikipedia.

Baxter [10] and van Dijk [31] adopt a somewhat opposite perspective from that of Hara et al., Nemoto and Gloor, and Pfeil et al., suggesting that the causal relationship between language edition and culture can flow both ways. Namely, they hypothesize that Wikipedias in languages with very small numbers of speakers could be actually be shaping the corresponding languages themselves. For instance, Breton and Lower Dutch Saxon do not have agreed-upon spellings for many words, and the Breton and Lower Dutch Saxon language editions have shown signs of being centralized resources through which these "language planning" tasks can be completed.

Ortega et al. [148] found that, at least as of 2008, a small number of contributors were

responsible for an outsized proportion of edits across all ten of the largest language editions of Wikipedia. Lih [120], on the other hand, describes several important anecdotes about various Wikipedia language editions that suggest differences in the corresponding editing communities. For instance, he writes that the Spanish Wikipedia was forked for about 1.5 years after a community controversy regarding ads and received essentially no edits during that time. In addition, he notes that the Japanese Wikipedia has the largest proportion of anonymous editors, something that may relate to Ortega et al.’s finding that Japanese had the least inequality of the ten examined language editions (but still was quite unequal). We have found in our work that the Japanese Wikipedia is quite unique in many additional ways, including the behavior of its readership and the level of cultural contextualization in its content.

Massa and Scrinzi [127] point out that the combination of (1) Wikipedia’s neutral point-of-view policy (NPOV) and (2) the fact that the socio-technical platform of Wikipedia does not afford consensus on NPOV across language editions may result in some of the differences between language editions’ content that we observe below. They suggest that each language edition may have its own “linguistic point-of-view” (LPOV), with neutrality being interpreted differently in each language-defined community.

Stvilia et al. [193] studied the “Featured Article” phenomenon across three language editions: Korean, Arabic, and English. Featured articles are, in the words of the English Wikipedia at least, selected by the Wikipedia community “to be the best articles Wikipedia has to offer” [229]. Stvilia and colleagues found, for instance, that in the Arabic Wikipedia, the (lack of) accuracy and (short) length of an article were significant predictors of the rejection of an article’s candidacy for featured status, while in the Korean Wikipedia, (lack of) accuracy and (lack of) writing quality were.

Finally, Yasseri et al. [215] examined the similarities and differences in the circadian patterns of editing across 34 language editions. They found that while all language editions followed a typical pattern of minimum editing around dawn and maximum editing in the late afternoon, there were interesting differences between the language editions. For instance, while English, German, Portuguese, Italian, and Simple English had maximum editing activity during weekdays, the opposite was true of the other language editions. In the Arabic and Persian Wikipedias, the weekend peak included Friday, which is a part of the weekend in most Muslim countries. This is an obvious reflection of the cultural context of these language editions.

### **3.1.2 Multilingual Wikipedia: Content**

This thesis focuses on Wikipedia content (and consumption of that content) as opposed to focusing directly on the peer production processes that resulted in that content. Several other researchers have viewed multilingual Wikipedia with a similar lens. For instance, Callahan and Herring [19] qualitatively examined 30 article pairs – all representing famous people – in the English and Polish Wikipedias. They found that while there are “systematic differences related to the different cultures, histories, and values of Poland and the United States,” a “U.S./English-language advantage is evident throughout.” This is an advantage we see in our work as well, although we also show that this advantage is nowhere near total. English is missing a substantial portion of the Polish content about a large number of topics, including biographies of people.

Callahan and Herring’s work is not alone in using some subset of the biographical domain as a representative sample of an entire language edition. The same approach is also taken in the work of Aragon et al. [6], which analyzed the betweenness centrality of biographical articles in the largest 15 language editions as of September 2011, finding more similarities than differences

in the five most-central articles in each of these language editions. For instance, George W. Bush, Hitler, and William Shakespeare appeared in the top five in many language editions. However, Aragon et al.’s results also highlight the cultural context of each language edition. For instance, Pope John Paul II is only in the top five for the Polish and Italian language editions. The same can be said of several Chinese leaders in the Chinese Wikipedia, and Castro and Che Guevara in the Spanish Wikipedia. Below, we show that the tendency for there to be more similarities than differences in the centrality of articles breaks down when considering other centrality measures (i.e. indegree centrality and PageRank centrality) and, more importantly, articles about all types of concepts rather than just biographies (Section 3.6).

The Aragon et al. study has an important additional limitation: they adopt the English-as-Superset corollary to the global consensus hypothesis by only looking at biographies that have an English version<sup>5</sup>. This resulted in, as they describe it, a serious “Anglo-Saxon bias,” with George W. Bush, Ronald Reagan, and Bill Clinton, for instance, being in the top five of many language editions. The English-as-Superset assumption likely caused this bias. Without it – as we will show below – the bias would likely go away, or at least be heavily tempered.

Filatova [41, 42] also focuses on biographies, using machine translation to analyze the similarities and differences across all Wikipedia language editions in the biographies of the 48 people considered in the DUC 2004 Task 5 biography summarization task [150]. She found, for instance, that many English articles about these 48 people are not the longest<sup>6</sup>, and, anecdotally, she noted that the shorter of two articles does not necessarily have all the content in the longer article. However, the work of Filatova also occasionally displays the global consensus hypothesis

---

<sup>5</sup> They used the DBpedia dataset prior to version 3.7, which was one of the most important cases of the Global Consensus Hypothesis.

<sup>6</sup> Although she did not control for “sub-articles” (see below).

in a strong form. Most notably, in an attempt to use the diversity across the language editions for document summarization, she writes that “the most trusted information [should be] repeated in the Wikipedia entry descriptions in different languages.” The word “trusted” here is highly problematic, and undermines her summarization approach. Information in only one language edition could be equally “trustworthy,” just not globally recognized to be relevant to the subject at hand.

By now it should be clear that a great deal of work in this space focuses specifically on biographies (e.g. [6, 19, 41, 42, 167]). In this chapter, we will problematize the notion that biographical articles are a representative sample of the content in multilingual Wikipedia and will show that whole-language edition and random sampling approaches are more appropriate.

There are a few papers in this space that have a wider scope than just biographies. Building off of our work [9, 80, 82], Warncke-Wang and colleagues, described the topic domains of the 823 articles with the most outgoing interlanguage links in all of multilingual Wikipedia. They found that the majority of these were about temporal concepts (e.g. 1932, September 22), while a solid minority described countries. Quite interestingly, they also found a few instances of spam or purposeful manipulation. For instance, the English article “True Jesus Church” had the most interlanguage links (254).

Also outside of the biographical domain, Massa and Scrinzi highlight several interesting examples of what we call sub-concept-level diversity when describing their Manypedia mashup [127], a tool with similar goals to Omnipedia ([9] and Chapter 7) but that relies on machine translation and focuses on only two language editions at once. For instance, Massa and Scrinzi note that the Arabic Wikipedia’s article about the Palestinian territories is called “Occupied Palestine” (at least according to Google Translate). Similarly, they note that the “List of

Controversial Topics” articles in the Chinese and Catalan Wikipedias are culturally contextualized, with the Catalan Wikipedia mentioning issues related to countries and nationality and the Chinese Wikipedia including content about China-related controversies.

In his well-known article on Wikipedia from the perspective of a historian [173], Rosenzweig highlights the Anglo-American bent of the English Wikipedia. He writes that he believes it to be more significant than any other bias in the language edition: “...The largest bias...favors Western culture (and English-speaking nations), rather than geek or popular culture” [173]. In this chapter, we directly demonstrate through a large-scale, quantitative study that the “English-speaking nation” bias is indeed quite substantial (Section 3.10) *in the English Wikipedia*. We also show that similar biases occur in all language editions we examine, with each language edition focusing on countries within the home cultural region of speakers of the corresponding language.

Halavais and Lackaff [67] disagree somewhat as to Rosenzweig’s conclusion about the most significant bias, suggesting that the bias against “expert” topics like law and medicine is the most significant, at least in terms of concept coverage. We build on these findings by showing that the coverage of each language edition can vary extensively by topic domain.

Using our WikAPIdia software, Ribé and Rodríguez [170] built on our *self-focus bias* work ([80] and Section 3.10) using many smaller language editions of Wikipedia and renaming self-focus bias “autoreferentiality.” They developed an index of autoreferentiality, with Icelandic, Swahili, and Japanese at the top (most self-focused), and Chinese, Dutch, and Catalan at the bottom (least self-focused).

Though a series of blog posts (e.g. [64, 65]) and prototype systems [8], Graham, Hogan, and their colleagues at the Oxford Internet Institute have been using a geographic lens to explore the

diversity in content across language editions. They have been particularly interested in Middle Eastern language editions. For instance, with their Mapping Wikipedia system [8], a user can see the self-focus biases in geographic article density in the Persian and Hebrew Wikipedias.

Looking at typical network metrics (e.g. average shortest path length, reciprocity, degree distribution) of each language edition’s article graph, Zlatić et al. [221] found that, by and large, there were far more similarities than differences, when controlling for maturity of the language edition. In their words, “...The similarities between Wikipedias in all the measured characteristics suggest that we have observed the same kind of a complex network in different stages of development.” Zlatić and colleagues did find some language-specific peculiarities, however. For instance, in the Polish Wikipedia, community decisions regarding pages about temporal topics (see below) caused the Polish article graph to differ from the others<sup>7</sup>. Zlatić et al. take this finding to mean that “the common growth process [they] observed is very sensitive to community-driven decisions.” In contrast to Zlatić et al., we find substantial differences between the language editions’ article graphs (Section 3.6), however our focus is on the content of the most-central articles rather than on a variety of properties of the graphs themselves.

In addition to their process-related work, Stvilia et al. [193] make several small contributions in the content space. They found anecdotal evidence that there was a tendency to promote “local” topics and priorities to featured status, quoting one contributor to the Arabic Wikipedia as writing, “We need encyclopedia articles that interest Arabic readers... I wish you (would) have made this effort to write a subject that benefits your people.” However, Stvilia et al. could not detect a significant relationship between “topic locality” and featured article vote outcomes in the Korean or Arabic Wikipedia. Stvilia and colleagues also relate anecdotal

---

<sup>7</sup> This is not something we found in our work with the Polish Wikipedia, however.

evidence that translation is an important task in featured articles, both to and from the English language edition. Interestingly, they report that one Korean or Arabic contributor wrote, “...It would be nice if the major parts could be ported to the English Wikipedia. From there it could find its way into the other Wikipedias as well.” This small piece of evidence backs the argument that English is becoming a global repository while the other language editions are focused more on parochial topics (see below), although our results show that at the scale of entire language editions, this argument has little support. Finally, Stvilia et al. also report that the concept coverage overlap between a random sample of 1,000 articles was 59% for Korean and English, which is quite similar to the result reported in our work (61%) [82] based on data from one year after the work of Stvilia et al. (2009) and the 65% reported below using 2012 data. Note that we show in Section 3.9.1 that concept coverage overlap overall has not increased substantially over this time period.

Some studies of the content of multilingual Wikipedia have also been done as a bi-product or result of more algorithmic research in the area. Most notably, demonstrating the application of their “polylingual topic modeling” approach, Minno et al. [138] studied a latent topic model analogue of what we call concept-level (Section 3.4) and sub-concept-level (Section 3.5) diversity across twelve language editions. Whereas we use an explicit topic model [47, 50] consisting of the number of concepts in our 25-language dataset, they use 400 latent topics as their frame of analysis.

Using their polylingual topic model, Minno et al. found extensive amounts of “topic diversity,” reporting that each of the twelve language editions they considered emphasizes different subsets of their 400 latent topics (as measured by the proportion of tokens from each language edition in each topic). All the examples of this phenomenon they discuss are instances

of cultural contextualization. For instance, according to Minmo et al., the Finnish Wikipedia discusses skiing frequently, while the same is not true of languages that are primarily spoken in regions with less snow, such as Hebrew, Farsi, and Turkish. Minmo et al. also note that they saw some similarities across the language editions at the topic level, with popular media-related tokens being mentioned at a more equal rate.

On the other hand, Minmo et al. found much less “sub-topic diversity,” an analogue to our sub-concept-level diversity. Examining the weights assigned to each of the latent topics of articles describing the same concept, they found the variation in weights on average to be roughly equal to that which they observed using documents that were direct translations of one another. There are many explanations for this seemingly “global consensus hypothesis” result, which conflicts with the findings of our work ([9, 82] and Section 3.5) and with that of others (e.g. [1, 19, 41, 42, 127]). First, the variation in the articles might be at a more detailed level than can be captured using 400 latent topics. For instance, the cultural contextualization of a given concept could be included in a single topic (e.g. the tokens representing brands of chocolate mentioned in each language edition’s articles about the chocolate could be placed in the same latent topic). Second, Minmo et al. generated these topics using only the first 1,000 *characters* of each article. It is likely that more diversity occurs past the first 1,000 characters, although investigating this question in a formal fashion is the subject of future work. The recent upgrades to our WikAPIdia software make this a trivial extension to the sub-concept-level work in Section 3.5. Third, and perhaps most importantly, like Aragon et al. [6], Minmo et al. only use articles that have an English equivalent, which could result in unforeseen biases. Fourth, it is possible that given that Minmo et al. trained their latent topic model on the same dataset they studied, the sub-topic diversity could have been explicitly “trained away” as their algorithm sought to obtain

the most cohesive concepts.

Also in the algorithmic space is the work of Adafre and de Rijke [1], who sought to use the Dutch and English Wikipedias as a means for generating parallel corpora for machine translation between these two languages. Adafre and de Rijke found that in a small sample of 30 concept pairs, articles about people had a high number of similar sentences across the language editions, while articles about general concepts (e.g. classicism and tennis) had fewer, a further rejection of the hypothesis that biographical articles form a representative sample of full language editions.

Adafre and de Rijke's work is also relevant to this chapter in that it is the first research to our knowledge to take the “bag of links” approach to representing content in Wikipedia articles, a key technique in our analyses of sub-concept diversity. They found that the bag of links approach was more accurate than a machine translation approach for their similar sentences task. In the multilingual Wikipedia space, researchers often make a choice between a bag of links approach (e.g. [9, 82, 136, 137]) and machine translation (e.g. [41, 42, 127]). The bag of links approach has numerous inherent advantages – e.g. it does not require access to unlimited machine translation resources and can be more human-readable – but this is the only study to our knowledge that has compared the performance of these two methods directly. This comparison was made possible because of the small number of concepts considered. Doing something equivalent across entire language editions would be impossible due to machine translation rate limits.

In general, as opposed to existing work on the content of multilingual Wikipedia, our research in this chapter has all of the following properties:

- We formally establish that cultural contextualization is the cause of some of the content diversity in multilingual Wikipedia (Section 3.10), something that is done only through

informal supposition in existing work.

- We consider entire language editions rather than small subsets of articles.
- We include far more language editions.
- Our work is domain-neutral and does not focus on specific topic areas. In fact, we compare across many topic areas in Section 3.7.
- We use a non-trivial concept alignment algorithm (Section 3.3).
- We use robust analyses to draw conclusions about the character of the diversity in multilingual Wikipedia in addition to anecdotal examples. With the small-scale exception of the work of Warncke-Wang et al. [203], every study above relies exclusively on anecdotal examples. Our robust analyses are the subjects of Section 3.6 through 3.10.
- Our work puts other language editions in the context of the English Wikipedia, allowing us to investigate the English-as-Superset hypothesis.

In general, this chapter represents the first work to comprehensively study and characterize the diversity across the entirety of a large number of language editions. It is also the first to definitively attribute this diversity to the cultural contexts of Wikipedia editors, analyze this diversity across time, study variation of this diversity across important dimensions like network centrality and topic, and evaluate the relationship between the diversity in content versus the diversity in the content’s consumption by Wikipedia readers. Our past published work in this area was some of the first to establish the widespread differences across the language editions of Wikipedia. Our goal in this chapter is to ask and address a series of new research questions with the goal of pushing the research in this area forward. The complete list of our contributions to the literature can be found in Section 3.11.

### **3.1.3 Other Types of Cultures and Wikipedia**

Researchers have also examined Wikipedia in the context of cultures other than those defined by language. Investigating the effect of the large gender gap in Wikipedia editors [25,

54], Lam et al. [111] used data from the MovieLens site [132] to find movies in which women were more likely to be interested than men. They then found that English Wikipedia articles about these movies tended to be shorter than articles about movies in which men were more interested. However, they did not find the same effect when they looked at articles about Nobel Prize winners from each gender.

Reagle and Rhue [167] took a slightly different approach to investigating gender-related patterns in Wikipedia content. Using a set of several thousand notable people, they found that Wikipedia had more of the corresponding biographies of people of both genders than Encyclopedia Britannica. However, they also found that where Wikipedia's coverage was lacking, it tended to be missing a biography of a woman. Similar to Lam et al., Reagle and Rhue found that there was no significant effect for gender in terms of biography length. That said, Reagle and Rhue's contention that "gender bias may not be a strong factor for article length" may be an overstatement given that, as noted above, there is a body of evidence suggesting that biographies are not a good representation of Wikipedia as a whole.

Political cultures and Wikipedia have also been the subject of interest in the research community. For instance, Greenstein and Zhu [66] found that while the English Wikipedia began with a Democratic bias, the encyclopedia has become centrist over time. However, this move towards the center has not been driven by changes at the sub-concept-level (Section 3.5), i.e. articles being re-written from a less liberal perspective. Rather, this move has largely been caused by an increase in the number of articles about more Republican-related topics (concept-level changes). In other words, once an article has a given "slant," that slant persists, but other articles can be written with the opposite slant. Greenstein and Zhu were able to perform this study using a LIWC-like [153] database of Democratic and Republican phrases developed by

Gentzkow and Shapiro [52]. It is important to note that the Greenstein and Zhu's work has been criticized for being United States-centric [195]. That is, a liberal bias in the United States could be a conservative bias elsewhere in the English speaking world.

Also in the political domain, Massa and Scrinzi [127] point out that due to Wikipedia's neutral point of view (NPOV) policy, several more extreme political cultures have left the Wikipedia community all together and started their own wiki-based encyclopedias. The most well known of these is Conservapedia<sup>8</sup>, which is developed by members of the far-right political community in the United States. Massa and Scrinzi, however, also highlight several other fascinating examples. For instance, Ecured<sup>9</sup> is an encyclopedia developed by the Cuban government explicitly written from a "decolonizer's point of view." Massa and Scrini write that,

"...The entry on the United States [on Ecured] describes it as the 'empire of our time, which has historically taken by force territory and natural resources from other nations, to put at the service of its businesses and monopolies' and that 'it consumes 25% of the energy produced on the planet and in spite of its wealth, more than a third of its population does not have assured medical attention'". [127, 223]

Lastly, Massa and Scrinzi also point out Anarchopedia<sup>10</sup>, which, interestingly, is available in numerous languages. The application of methods similar to those in this chapter to these politically-defined encyclopedias represents fertile ground for future work.

Given the strong correlation between nation-defined and language-define cultures, much of the work in this chapter is relevant in the nationality domain, especially in the case of nation-states [32] like Norway, Sweden, and so on. The work on nationality in Wikipedia largely relies on the same correlation (e.g. [144, 155]). The one major exception is the work of Liao [117],

---

<sup>8</sup> [http://www.conservapedia.com/Main\\_Page](http://www.conservapedia.com/Main_Page)

<sup>9</sup> [http://www.ecured.cu/index.php/EcuRed:Enciclopedia\\_cubana](http://www.ecured.cu/index.php/EcuRed:Enciclopedia_cubana)

<sup>10</sup> [http://www.anarchopedia.org/Main\\_Page](http://www.anarchopedia.org/Main_Page)

which describes how peoples from many nations have come together to build the Chinese Wikipedia, even though these nations are often in direct conflict. He writes, for instance, about the Chinese Wikipedia’s “Anti-Regionalism Policy,” which he describes as an antidote to Chinese (PRC) cyber-nationalism:

“...The policy mandates that China-centric, Han-centric, and Chinese-centric statements should be avoided. Thus, it avoids two premises underlying Chinese cyber-nationalism: that China is always a “Middle” or “Center” Kingdom, and that a strong “central” government is essential for China” [95].

### **3.1.4 Language-defined Cultures and non-Wikipedia User-Generated Content**

We now briefly turn our attention to the language-defined cultural contextualization of UGC other than that in Wikipedia. There has been some interest in the expression of language-defined cultures on Twitter. Hong and colleagues showed that English tweets make up only 51% of all tweets [80], results that were echoed by Semiocast [182]. More importantly, Hong et al. also demonstrated that the adoption of different Twitter conventions varies among the top 10 languages on Twitter. For instance, German-language tweets tended to contain more URLs and hashtags, while Korean-language tweets were more likely to be conversational in nature.

Hale [68] and Herring et al. [19] have examined the role of language-defined communities in the blogosphere. In a study of blog posts in English, Spanish, and Japanese on the 2010 Haitian earthquake, Hale found that English was the only language to have more incoming cross-language links than outgoing ones and that most cross-language links signaled reference rather than direct translation, a finding in line with our results showing that articles about the same concept in different language editions are far from translations of one another ([9, 82] and Section 3.5). Interestingly, the largest single destination of cross-language links in Hale’s dataset was a photoblog by a Denver Post photographer, suggesting that if we executed some of the

below studies on images instead of text content, we would find more consensus across the language editions.

Herring et al. [91] performed an early analysis of LiveJournal use among English, Russian, Portuguese, Finnish, and Japanese bloggers. They found that English dominated globally, but languages like Russian also had a significant network of journals. In terms of cross-language information transfer, they reported that “young, multilingual, geographically mobile bloggers link to, and are linked by, journals in different language groups, creating de facto bridges across cultures.” Finally, their work agrees with Hale’s with regard to the importance of non-verbal communication; videoblogs and other visual content were another important means through which the language barrier was overcome.

### **3.2 Parsing, Extraction, and Wikipedia Resources**

Now that we have laid out the related literature, we can begin to discuss our research showing that the language editions of Wikipedia reflect the cultural contexts of their contributors. In this section, we describe the low-level methodology underlying all of this work. First, we briefly discuss our methods for parsing and extracting information from many different language editions of Wikipedia. Here, we show that while the language-neutral approaches that are typically used in the literature to process multilingual Wikipedia are successful for some types of data, language-specific approaches are absolutely necessary for others. This discussion is framed in the context of *WikAPIdia*, the Wikipedia software library we built that we use for nearly all the research in this chapter. WikAPIdia, an important contribution of this thesis that will be released under a LGPL license, is described in much greater detail in Section 3.12.

In the second part of this section, we describe the numerous resources in Wikipedia we

utilize throughout this chapter such as article metadata, links between articles, anchor texts, and redirects. While many of these resources are considered widely in the literature, we discuss how we were the first to identify and address important nuances in these resources that can have significant implications for Wikipedia-based research and technologies.

Finally, we close this section with a discussion of the results of our parsing and extraction process. Here, we highlight the extent to which working with multilingual Wikipedia as opposed to just the English Wikipedia can increase dataset sizes by an order of magnitude or more.

### **3.2.1 Parsing and Extraction Process**

The parsing functionality of WikAPIdia takes as input static *database dumps* of each supported language edition of Wikipedia, as well as several other related datasets described in more detail as the chapter progresses (e.g. page views, topics, geographic references). The static database dumps are made available by the Wikimedia Foundation<sup>11</sup>. The most significant component of these dumps is the XML file that contains the content of all the pages in a given language edition at the time the database was exported. Not surprisingly, these files can be quite large. The English XML file that was used for much of this thesis is 43GB uncompressed.

Our approach to parsing and extraction – and that built into WikAPIdia – is to rely on language-neutral approaches whenever possible, but to back-off to language-specific adaptations when necessary. Language-neutral approaches are enabled by the fact that editors in all language editions of Wikipedia use the *Wiki markup* language [90] when writing and editing articles. The XML database dump of each language edition includes the content of all pages in native Wiki markup form. This means that a parser of the XML file can always interpret certain Wiki markup constructs in the same way, no matter the language edition. For example, text enclosed by two

---

11 <http://dumps.wikimedia.org/backup-index.html>

brackets indicates a link in all language editions.

However, just because all language editions use Wiki markup does not prevent each language edition-specific editor community from developing its own set of unique syntax and constructs. For instance, each language edition has its own set of words used to distinguish categories from articles (e.g. “Category” in English, “Kategorie” in German, “Categoría” in Spanish), its own category with which to flag *disambiguation pages* (e.g. “Disambiguation pages” in English, “Flertydig” in Danish, “동음이의어\_문서” in Korean), and so on.

Categories, disambiguation pages, and the other structures that can only be identified using language-specific mechanisms are essential to many aspects of the research in this chapter and this thesis more generally. As such, we needed to be sure WikAPIdia could handle parsing and extraction of these resources as well as it did the language-neutral ones. Our approach here was to semi-manually build a large dictionary of the necessary terms and regular expressions. The automated portion of our approach involved extracting information from the source code of MediaWiki, the open-source software upon which all language editions operate. This was useful in identifying signifiers for categories and that of several other structures. However, this approach could not be adopted in several important cases. Where this occurred, we manually accessed dozens of pages in each of the language editions included in our studies to identify all variants of the identifiers for a given Wikipedia structure. For instance, to identify the category to which disambiguation pages belong in each language edition, we used strategies such as looking up terms that tended to be ambiguous in many languages (e.g. “MS”, “PC”, “US”). The most significant instance of this manual investigation approach occurred when identifying signifiers for *sub-articles*, a process that is discussed in detail in Section 3.5.1.3.

There are also higher-level ways that the language editions differ that are important to

consider when doing multilingual Wikipedia research. Diversity in the use of namespaces is a particularly significant issue. While most language editions (English included) consider articles like “Northwestern University” (English) and “List of Northwestern University alumni” (English) to be equivalent in the type of information they describe, this is not true of the Portuguese or the Spanish Wikipedias. In these Wikipedias, lists are segregated into their own namespace – “Anexo:” in both cases – just as categories have their own namespace in all language editions. To ignore this structural variation would be to ignore over 10,000 articles in each of the Spanish and Portuguese Wikipedias (as well as all their inlinks, outlinks, and other properties). In addition, doing so would add language-specific error to the diversity analyses in these chapters, a particularly dangerous type of error for our work. As such, we carefully have written WikAPIdia such that it has dedicated support for variation in namespaces across language editions.

For researchers and system builders in the multilingual Wikipedia space, there is an important high-level takeaway from WikAPIdia’s approach to parsing and extraction. It is tempting to assume that since all language editions are based on the Wiki markup language, one does not need to be concerned with language community-specific adaptions of how that language is used. However, in developing WikAPIdia, we have found that ignoring language community-specific properties can lead a larger number of errors (e.g. the “Anexo” namespace) and missed opportunities to access useful resources (e.g. disambiguation pages, sub-articles). In other words, fully leveraging multilingual Wikipedia requires “going into the weeds” and seeing how Wiki markup is actually used by each language community.

At a lower level, the most basic of the latest version of WikAPIdia’s parsing and extraction capabilities are built on the Java Wikipedia Library’s (JWPL) [220] parsing package. Previous

versions of WikAPIdia did not rely on an external package for parsing and extraction. However, for reasons of code maintenance, reliability, and repeatability, it was determined that the use of JWPL would be optimal moving forward. That said, WikAPIdia includes a large number of customizations on top of JWPL, particularly to support language-edition specific resources, which JWPL does not consider. There are also numerous additional parsing and extraction steps involving link properties (parseable v. unparseable), topical information, page view data, spatial references, and many other structures that are not made available directly in an XML language edition database dump. As JWPL (or any other Wikipedia API) does not consider these types of data, parsing and extraction in these cases is handled natively in WikAPIdia.

### 3.2.2 Wikipedia Resources

*Wikipedia article metadata* (Figure 3.2-a) is a resource that plays an important role throughout the thesis. This metadata describes all articles in all supported language editions and includes properties such as the article's title, the article's MediaWiki-given ID number, the number of characters in the article's content (calculated during the parsing process), the date of last edit, and so on. Article metadata is one of the most important ingredients to nearly all Wikipedia-related projects in this thesis. For instance, it is used to connect text to concepts in semantic relatedness algorithms (Chapter 6), it allows us to display language-specific labels on Omnipedia circles (Chapter 7), and it provides structure for the *Conceptualign* algorithm that groups articles about the same concept in different language editions (Section 3.3).

McCarthy, Alaska – Wikipedia, the free encyclopedia

W McCarthy, Alaska – Wikipedia, t...

en.wikipedia.org/wiki/McCarthy,\_AK

Article Title (Metadata)

McCarthy, Alaska

From Wikipedia, the free encyclopedia  
(Redirected from McCarthy, AK)

**Redirect**

McCarthy is a census-designated place (CDP) in Valdez-Cordova Census Area, Alaska, United States. The population was 28 at the 2010 census.

Contents [hide]

- 1 Geography and location
- 2 Demographics
- 3 History
- 3.1 Shooting
- 4 References
- 5 External links

**Intralanguage Links**

Geography and location [edit]

McCarthy is 193 km (120 mi) northeast of Cordova at the foot of the Wrangell Mountains. According to the United States Census Bureau, the CDP of McCarthy has a total area of 148.3 square miles (384 km<sup>2</sup>). None of the area is covered with water. It is connected to the outside world via the McCarthy Road spur of the Edgerton Highway to Chitina, and must be passed through to reach Kennecott, a destination of tourists seeking access to Wrangell-St. Elias National Park and Preserve.

Demographics [edit]

As of the census [2] of 2000, there were 42 people, 26 households, and 6 families residing in the CDP. The population density was 0.3 people per square mile (0.1/km<sup>2</sup>).

**Anchor Text**  
*link target is "Kennecott, Alaska"*

The old McCarthy hardware store

Location of McCarthy, Alaska

Coordinates: 61°25'58"N 142°54'39"W<sup>[1]</sup>

Country	United States
State	Alaska
Census Area	Valdez-Cordova
Area	• Total 148.3 sq mi (384.0 km <sup>2</sup> )

**Interlanguage Links**

Categories: Census-designated places in Alaska

Populated places in Valdez–Cordova Census Area, Alaska

**Category Memberships**

This page was last modified on 1 November 2012 at 00:37.

Figure 3.2-a: The article “McCarthy, Alaska” (English) labeled with examples of the various Wikipedia resources discussed in this section. This page is shown as rendered by the MediaWiki software on which all language editions operate. The source of each piece of information labeled above is information encoded in specific Wiki markup syntax, sometimes modified for each language edition (e.g. categories and redirects).

*Intralanguage links* are another critical Wikipedia resource extracted by WikAPIdia. Intralanguage links, typically referred to simply as “links,” are the familiar blue hyperlinks that are peppered throughout most articles in all language editions. Each link connects an article in language edition  $l$  to another article in language edition  $l$  and represents some (unlabeled) semantic relationship between these two articles. For instance, in Figure 3.2-a, the links labeled “United States Census Bureau” and “Edgerton Highway” are intralanguage links that point to the corresponding articles in the English Wikipedia.

The intralanguage links of a given language edition comprise the edges of that language edition's *Wikipedia Article Graph* (WAG) [77, 85], with the articles in the language edition comprising the vertices. WAGs and the links they contain are essential to much of the research in this chapter and that in Chapters 6 and 7. For instance, links are utilized as structured proxies for the content of an article in Section 3.5, centrality in each language edition's WAG is the entire focus of Section 3.6, and inlinks and outlinks provide the world knowledge for several of the semantic relatedness measures in Chapter 6.

The work in this thesis is the first (even relative to our published research) to distinguish between *parseable* intralanguage links and *unparseable* intralanguage links. Parseable links are those that can be extracted by parsing the Wiki markup of any given Wikipedia page. This means that parseable links are *generally* inserted by a human Wikipedia editor contributing to the content on a single specific Wikipedia page, although there are exceptions to this rule-of-thumb (see Section 3.10.2.1). Unparseable links, on the other hand, are “hidden” behind *templates* and are not directly accessible via the Wiki markup of a page. A template is:

“...a Wikipedia page created to be included in other pages. Templates usually contain repetitive material that might need to show up on any number of articles or pages. They are commonly used for boilerplate

messages, standard warnings or notices, infoboxes, navigational boxes and similar purposes.” [226]

In other words, templates generate many of the formulaic parts of Wikipedia pages. For instance, many articles about geographic entities include templates that insert dozens to hundreds of links to articles about nearby geographic entities at the bottom of the articles (i.e. “navigation boxes”). Figure 3.2-b shows a subset of these unparseable links on the “Ann Arbor, Michigan” (English) article. The figure also shows that these links are the result of just three templates in the Wiki markup of the article.

Unparseable links appear on a page via almost entirely automated means; the only manual part of the process is the original inclusion of the template on the page. Moreover, the content on a specific page that comes from a given template cannot easily be changed without changing the content on all pages that reference that template. While some parseable links are the output of automated Wikipedia bots – most famously some of the links on English Wikipedia pages about places in the United States [120] – they can still be manually manipulated on an article-by-article basis by Wikipedia editors, and this has occurred to an enormous extent.

Ann Arbor, Michigan – Wikipedia, the free encyclopedia

W Ann Arbor, Michigan – Wikipedia...

en.wikipedia.org/wiki/Ann\_Arbor,\_Michigan

Google

<b>Topics</b>	Lieutenant Governors · Lighthouses · Museums · National Historic Landmarks · National Register of Historic Places listings · People · State Historic Sites · State parks · Supreme Court · Tallest buildings · Timeline · Topics · Visitor attractions	
<b>Regions</b>	<b>Upper Peninsula:</b> Copper Country · Keweenaw Peninsula · Gogebic Range <b>Lower Peninsula:</b> Central Michigan · Southern Michigan · Flint/Tri-Cities · The Thumb · The Greater Tri Cities · Michiana · Northern Michigan · Southeast Michigan · Metro Detroit · West Michigan	
<b>Largest Municipalities</b>	Alpena · <b>Ann Arbor</b> · Battle Creek · Bay City · Bloomfield Township · Canton Township · Chesterfield Township · Clinton Township · Commerce Township · Dearborn · Dearborn Heights · Detroit · East Lansing · Eastpointe · Farmington Hills · Flint · Flint Township · Georgetown Township · Grand Rapids · Holland · Jackson · Kalamazoo · Kentwood · Lansing · Lincoln Park · Livonia · Macomb Township · Meridian Township · Midland · Muskegon · Novi · Pontiac · Portage · Redford · Rochester Hills · Roseville · Royal Oak · Saginaw · Saginaw Township · St. Clair Shores · Shelby Township · Southfield · Sterling Heights · Taylor · Troy · Warren · Waterford Township · West Bloomfield · Westland · Wyoming · Ypsilanti Township	
<b>Counties</b>	Alcona · Alger · Allegan · Alpena · Antrim · Arenac · Baraga · Barry · Bay · Benzie · Berrien · Branch · Calhoun · Cass · Charlevoix · Cheboygan · Chippewa · Clare · Clinton · Crawford · Delta · Dickinson · Eaton · Emmet · Genesee · Gladwin · Gogebic · Grand Traverse · Gratiot · Hillsdale · Houghton · Huron · Ingham · Ionia · Iosco · Iron · Isabella · Jackson · Kalamazoo · Kalkaska · Kent · Keweenaw · Lake · Lapeer · Leelanau · Lenawee · Livingston · Luce · Mackinac · Macomb · Manistee · Marquette · Mason · Mecosta · Menominee · Midland · Missaukee · Monroe · Montcalm · Montmorency · Muskegon · Newaygo · Oakland · Oceana · Ogemaw · Ontonagon · Osceola · Oscoda · Otsego · Ottawa · Presque Isle · Roscommon · Saginaw · Sanilac · Schoolcraft · Shiawassee · St. Clair · St. Joseph · Tuscola · Van Buren · Washtenaw · Wayne · Wexford	
V · T · E	<b>Municipalities and communities of Washtenaw County, Michigan, United States</b>	[show]
V · T · E	<b>Metro Detroit</b>	[hide]
<b>Topics</b>	Architecture · Culture · Detroit River · Economy · Freeways · History · Historic places · International Riverfront · Lake St. Clair · Media · Music · Parks and beaches · People · Skyscrapers · Sports · Theatre · Tourism · Transportation	
<b>Major city</b>	Detroit	
<b>Municipalities over 80,000</b>	Canton Township · Clinton Township · Dearborn · Livonia · Sterling Heights · Troy · Warren · Westland	
	Bloomfield Township · Chesterfield Township · Dearborn Heights ·	

---

 **WIKIPEDIA**  
The Free Encyclopedia

Main page  
Contents  
Featured content  
Current events  
Random article  
Donate to Wikipedia

## Editing Ann Arbor, Michigan

Content that violates any copyrights will be deleted. Encyclopedic content must be verifiable. Work submitted to Wikipedia can be edited, used, and redistributed—by anyone—subject to certain terms and conditions.

**B I**  Advanced Special characters Help Cite

```
{{Michigan}}
{{Washtenaw County, Michigan}}
{{Metro Detroit}}
```

Figure 3.2-b: The top part of the figure shows a selection of the unparseable links present in the “Ann Arbor, Michigan” (English) article. The bottom of the figure shows the just three templates in Wiki markup that are responsible for the hundreds of links in the top of the figure.

In the human-computer interaction (HCI), artificial intelligence (AI), and natural language processing (NLP) communities (among others), there are dozens of research projects that use WAGs. However, few of them report whether they consider parseable links or all links, and when they do, they do so implicitly. The issue has, in fact, never been considered in the literature to our knowledge. The primary reason for the confusion is that there are two sources of link information available from the Wikimedia Foundation. The first is the XML dump file discussed above, which allows a Wikipedia XML parser such as WikAPIdia to extract link information from Wiki markup. However, the Wikimedia Foundation also makes available a SQL-formatted link dump file that, rather than containing raw Wiki markup, is copied directly from its database of links between articles. The XML file only contains parseable links, whereas the SQL file contains all links, including unparseable ones.

<b>Parseable WAG</b>	<b>Unparseable WAG</b>	<b>All Links WAG</b>
France	Kingdom (biology)	United States
United States	Phylum	Area
Spain	Class (biology)	France
Municipality	Genus	Population
Barcelona	Sea level	Sovereign state
2007	Biological classification	Sea level
Italy	Altitude	Popularity density
Catalonia	Order (biology)	Altitude
Sovereign state	Superfamília	Spain
Germany	Forma specialias	Geographic coordinate system
Median	Area	Catalonia
2009	Population	Kingdom (biology)
English language	Straight (Biology)	Barcelona
Europe	Grup (biologia)	Species
French language	Legió (biologia)	Class (biology)

Table 3.2-a: The Catalan articles with the top PageRank scores according to the parseable, unparseable, and all links Catalan WAGs.

The source of link information is an important factor to consider in Wikipedia-based research and applications for two reasons. First and foremost, unparseable links make up a substantial proportion of links. We show in the section on the results of our parsing and extraction process (Section 3.2.3) that this proportion can be as high as over 80 percent in some language editions. This represents a four-fold increase in the content of each WAG.

Second, the generative process for parseable and unparseable links is different and, as such, the interpretation of research involving Wikipedia links needs to be understood in the context of the links' parseability. For instance, consider Table 3.2-a, which shows the articles in the Catalan Wikipedia with the 15 highest PageRank scores. PageRank is discussed in detail in Section 3.6, but for now it is enough to know that it is a means of judging the “importance” of a given article in a WAG. Note that in Table 3.2-a the parseable and unparseable lists have *zero overlap* with one another; not a single concept in the parseable WAG’s top 15 is in the unparseable WAG’s top 15. More importantly, the unparseable WAG is dominated by concepts related to biological taxa, an artifact of a prolific template that automatically generates links to these concepts from articles about all flora and fauna. When considering the WAG made of all links, the enormous outliers that are the taxa-related articles in the unparseable WAG have a substantial effect. Barcelona, for instance, moves down 10 ranks to below the concept about biological kingdoms.

The dangers of ignoring the differences between parseable and unparseable links are clear in Table 3.2-a. If a researcher assumes that the large majority of the links s/he got from the SQL database dump are created in the traditional process of a Wikipedia editor contributing Wiki markup-formatted text, there is a serious risk of drawing false conclusions. While there are times where unparseable links can be useful, their nature and the process by which they are created must be considered.

Put together, what seems like an innocuous choice of file format can make an enormous difference in WAG-based algorithms and studies. Moreover, generally the only way to determine if a paper in the literature used all links or just parseable ones is if the authors happen to report the origin of their WAG information (i.e. XML dump or SQL dump). With WikAPIdia and with this thesis, we sought to improve upon this situation. As we will discuss in more detail in Section 3.12, the latest version of WikAPIdia supports both types of links and distinguishes parseable from unparseable links. It does so by calculating the intersection and XOR of links extracted from the XML dump and those extracted from the SQL file. WikAPIdia is thus able to facilitate easy access to the parseable, unparseable, and combined WAG of each supported language edition, a capability we make use of this in chapter. Given that we will release WikAPIdia upon completion of this thesis, this feature will also make it possible for other researchers and practitioners to consider multiple WAG versions in their work. Our goal here is to increase the number of research projects and systems that carefully consider the issue of link parseability.

Our research also addresses one more property of links not widely considered in the Wikipedia literature: the location of (parseable) links in an article. In all language editions considered here, the first paragraph and first section of longer articles generally contain a summary of the article, or, in the language of the NLP community, a *gloss* of the article. As such, a few papers in the Wikipedia-related NLP literature (e.g. [219]) – inspired by the nature of glosses in WordNet [134] – have hypothesized that content in these summaries might be more representative of the topic of the article than content further down in the article. In order to support more research in this space, we have constructed WikAPIdia such that it flags any link that occurs in the first paragraph or the first section of an article and facilitates access to the (parseable) WAG that only contains these links. As of this writing, we only use this feature to

support the “article summary” mode in Omnipedia. In this mode, users can explore the diversity present in the summaries of articles about the same concept in different language editions, rather than exploring the articles as a whole. However, an important area of future work is investigating the WAG-related results below in the context of link location.

While a regular, intralanguage link connects two articles in the same language edition, an *interlanguage link* (ILL) connects two articles in *different* language editions. The ILL structure of multilingual Wikipedia is a vital resource upon which the research in this chapter is based. An ILL indicates that the two articles it connects are about the same concept, even though they are not written in the same language. For instance, in Figure 3.2-a, the blue links labeled with names of languages are interlanguage links that point to articles about McCarthy, Alaska in the corresponding languages. The link labeled “Português” in the figure points to the Portuguese Wikipedia article “McCarthy (Alasca),” the link labeled “Català” points to the Catalan Wikipedia article “McCarthy (Alaska),” and so on.

In the *interlanguage link graph* (ILL graph), the vertices are all articles in *all* language editions and the edges are the interlanguage links that connect them. While the interlanguage link graph is used only in a single section of this thesis (Section 3.3), that section is fundamental to nearly all of our multilingual Wikipedia work. Specifically, the ILL graph is the most important input to the *Conceptualign* algorithm, which allows for language-neutral, concept-based access to all Wikipedia resources described in this section and the more specialized resources discussed later in this chapter.

McCarthy – Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/McCarthy

Brenthecht Talk Sandbox Preferences Watchlist Contributions Log out

Article Talk Read Edit View history Search

# McCarthy

## Link to McCarthy, Alaska

From Wikipedia, the free encyclopedia

**McCarthy** may refer to:

- McCarthy (surname) (article includes many notable persons with this surname)
- McCarthy, Alaska, United States
- McCarthy (band), an indie pop band
- Château MacCarthy, a Bordeaux wine
- McCarthy Tétrault, a Canadian law firm
- McCarthy evaluation, programming-language semantics also called short-circuit evaluation, named after John McCarthy (computer scientist)

"Candidate Senses"

**See also**

**Disambiguation page category**

This disambiguation page lists articles associated with the same title.

If an internal link led you here, you may wish to change the link to point directly to the intended article.

Categories (+): Disambiguation pages | (+)

This page was last modified on 29 August 2012 at 09:44.

Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. See Terms of Use for details.

Wikimedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.

Contact us

Privacy policy About Wikipedia Disclaimers Mobile view

WIKIMEDIA project Powered By MediaWiki zotero

Figure 3.2-c: The English-language page “McCarthy”, an example of a disambiguation page. Note that the “McCarthy, Alaska” page is indicated to be one of the possible meanings of the term “McCarthy”. In the terminology of the semantic relatedness literature, these possible meanings are known as “candidate senses”. Disambiguation pages belong to the disambiguation page category for their language edition, which in English is “Disambiguation pages”.

Although it is a community-defined structure rather than a native feature of Wiki markup, all of the editor communities behind the language editions considered in this thesis utilize what are known in English as *disambiguation pages*. The disambiguation page resource, a subset of the article metadata resource, is another important input to *Conceptualign*. Disambiguation pages are “non-article page[s] which (sic) lists the various meanings of [a given ambiguous term] and links to the articles which (sic) cover them” [228]. For instance, Figure 3.2-c shows the “McCarthy” (English) disambiguation page. This page is intended to direct a Wikipedia reader who searches for the term “McCarthy” to the article that describes the concept about which the reader was seeking information. Each link in the “may refer to” section of the page refers to one such concept, including McCarthy, Alaska.

In the NLP community, disambiguation pages are used as a source of information about the *candidate senses* of a given term. In other words, disambiguation pages help NLP algorithms connect *terms* (e.g. “McCarthy”) with *concepts* as described by Wikipedia articles (e.g. “McCarthy, Alaska”). This approach is taken, for instance, in our version of the *MilneWitten* semantic relatedness algorithm [136, 137], which we use in the concept alignment process in Section 3.3 and is featured prominently in Chapter 6.

In *Conceptualign*, on the other hand, disambiguation pages are useful because of the special role they play in the ILL graph. Specifically, as described in Section 3.3, due to their nature as primarily language-specific entities rather than language-neutral concepts, they are “stopping points” in *Conceptualign*’s ILL graph breadth-first search.

Disambiguation pages belong to disambiguation categories, which are a part of the Wikipedia category structure. Each Wikipedia article in all language editions may be “tagged” with one or more language edition-specific *category memberships*. For example, “McCarthy,

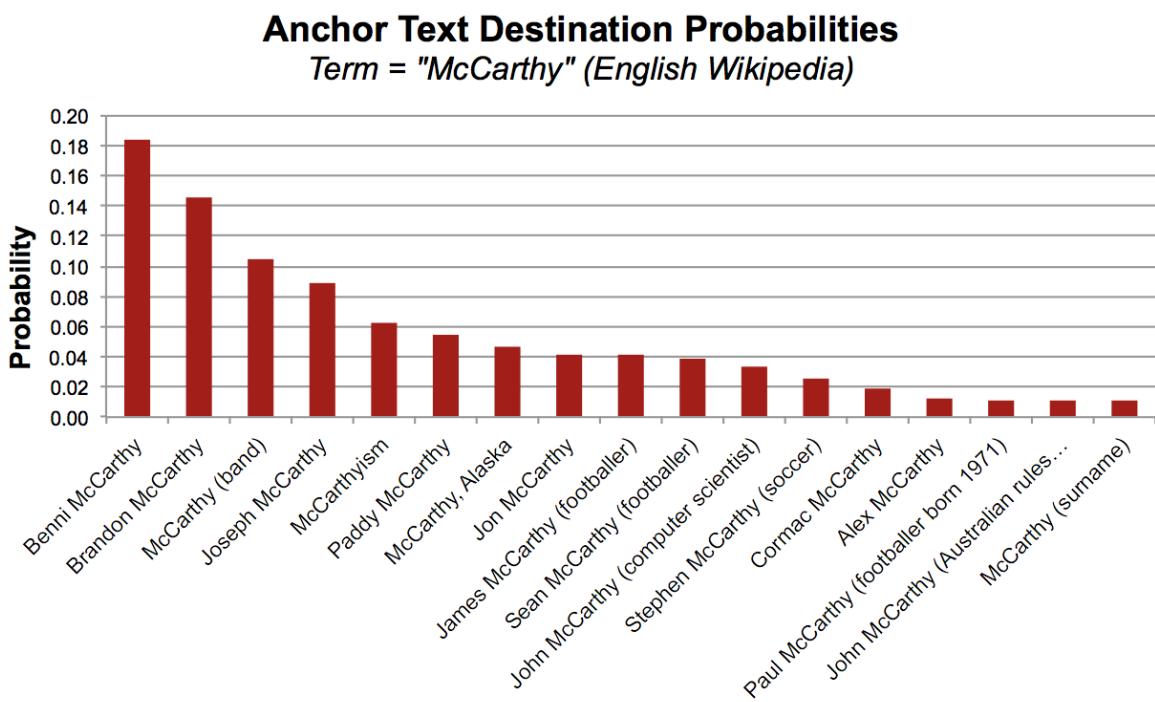
Alaska” (English) is indicated to be a member of “Category: Census-designated places in Alaska” (English) and “Category: Populated places in Valdez–Cordova Census Area, Alaska” (English) categories (Figure 3.2-a). The “McCarthy” (English) article belongs to “Category: Disambiguation pages” (English), the *category* designated by the English Wikipedia community to contain disambiguation pages. These category memberships are aggregated on language-specific category pages, which themselves can be tagged with category memberships. This nested structure forms each language edition’s *Wikipedia Category Graph* (WCG). Because it is a bottom-up classification schema defined by the editor communities of each language edition and contains cycles, the WCG in each language edition is a folksonomy [199] rather than a true taxonomy. The vertices of each WCG are the category pages in the corresponding language edition and the edges are the category memberships that connect these pages. The “leaves” of each WCG are the articles that belong to each category. WCGs are not used frequently in this thesis but are easily accessed in a structured fashion in WikAPIdia, primarily in order to support the implementation of *WikiRelate* [156, 192], one of the semantic relatedness algorithms in Chapter 6.

As category pages resemble regular articles in Wiki markup syntax, WikAPIdia has been designed such that any algorithm that can be run on the WAG of each language edition can also be run on the corresponding WCG, potentially opening up the WCGs to novel experiments and applications by other research groups. Along the same lines, *Conceptualign* can be executed on the category ILL graph just as it can be on the article ILL graph, and we have done just this. However, because the quality of the category ILL graph has not yet been verified, we do not report WCG statistics in this chapter.

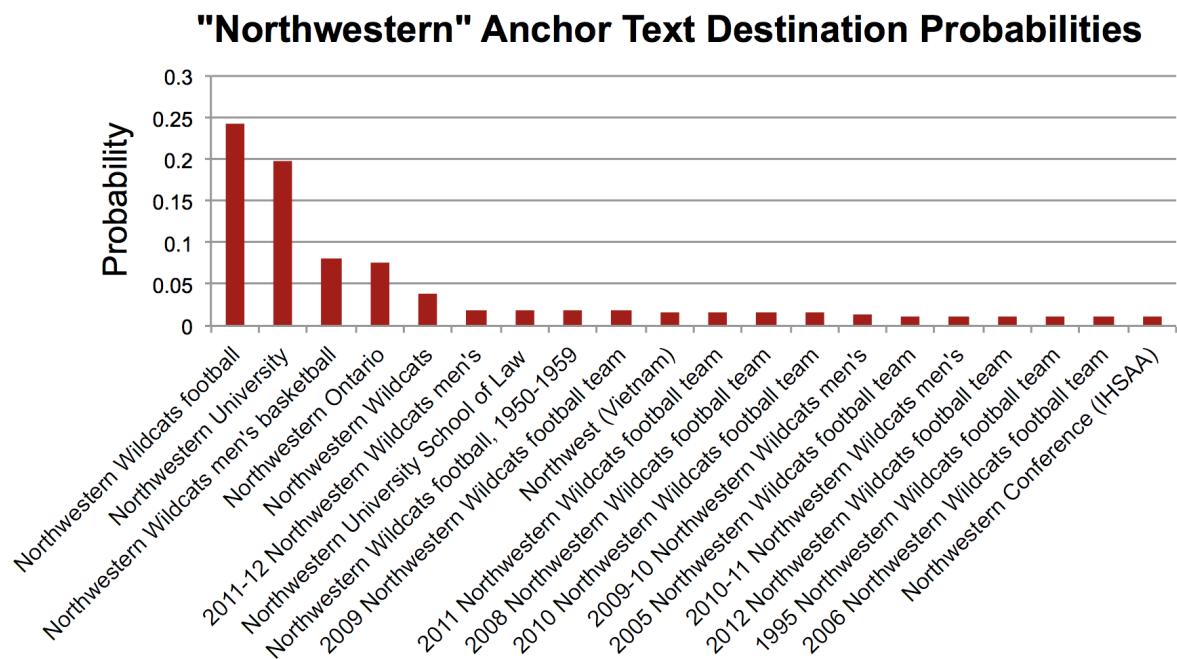
The *redirects* resource is a simple but important source of information for our work. In all

language editions considered here, Wikipedia editors have encoded numerous alternative titles for articles such that when a Wikipedia reader enters one of these alternative titles in the Wikipedia search bar (or searches for that title in a search engine), they will be directed to the corresponding article. These alternative titles are known as redirects, and effectively act as article aliases. For example, in Figure 3.2-a, a flag at the top of the page shows that the user had searched for the term “McCarthy, AK” and was redirected to the “McCarthy, Alaska” article, indicating that the English Wikipedia community has constructed a redirect from “McCarthy, AK” to “McCarthy, Alaska.”

Whereas the disambiguation page resource can be understood as a repository of homonyms, redirects can be used as a source of synonymy information. Redirects are thus effective at connecting terms to Wikipedia articles in an *unambiguous* fashion. For instance, Omnipedia uses redirects so that users are not required to enter the exact title of an article in order to access the corresponding “multilingual article.” In other words, a user who types in “McCarthy, AK” will receive the same result as a user who typed “McCarthy, Alaska.” Atlasify (Chapter 8) takes a similar approach.



*Figure 3.2-d: The probability that a link with the anchor text “McCarthy” in the English Wikipedia will have a given article as its destination. Only articles with  $p > 0.01$  are included. The probability that “McCarthy, Alaska” will be the destination is around 4.5%, while it is around 18% for the article “Benni McCarthy,” which is about a South African soccer player.*



*Figure 3.2-e: The probability that a link with the anchor text “Northwestern” in the English Wikipedia will have a given article as its destination. Only articles with  $p > 0.01$  are included. Interestingly, “Northwestern Wildcats football” has a higher probability than “Northwestern University,” perhaps signaling either a sports bias in the English Wikipedia or a tendency for Wikipedia editors to use alternative terminology (e.g. the full name of the university) when referring to the university.*

Another resource that can be used to connect terms to Wikipedia articles is the *anchor text* structure. Anchor texts in Wikipedia are roughly the same as anchor texts in the wider web, i.e. they are the “visible, clickable text in a hyperlink” [222]. For instance, in Figure 3.2-a, the blue intralanguage link with the label “Kennicott” does not link to the article “Kennicott” (English), which happens to be a disambiguation page, but rather to the article “Kennicott, Alaska” (English). Wikipedia editors are able to choose the visible text that describes the destination of a link using simple Wiki markup syntax. Anchor texts, when aggregated, have the useful property of being a *probabilistic* way of connecting terms to Wikipedia articles. For instance, Figure 3.2-d and Figure 3.2-e show the link target probability distributions of the anchor texts “McCarthy” and “Northwestern” in the English Wikipedia. In Figure 3.2-d, the probability of “McCarthy, Alaska” being the destination of a link with the anchor text “McCarthy” is only approximately 0.045. Interestingly, in Figure 3.2-e it can be seen that the article about the Northwestern University football team is the most likely destination of the anchor text “Northwestern,” with the actual university being the second most likely by a decent margin.

The actual text on the page of each article in each language edition makes up the *Wikipedia article text* or *Wikitext* [220] resource. WikAPIdia stores this information in a text index that is optimized for search as well as for document-by-document access and is implemented using the Lucene Java-based indexing and search library<sup>12</sup>. WikAPIdia provides access to the Wikitext resource in Wiki markup format as well as in plaintext, with all Wiki markup syntax removed. While it is utilized less often than article metadata, Wikitext plays a critical role in several important components of this thesis, including the missing link-finding algorithm and corresponding experiments discussed below and in our implementation of the well-known

---

<sup>12</sup> <http://lucene.apache.org/>

Explicit Semantic Analysis semantic relatedness algorithm [47] (Chapter 6).

One challenge we faced with the Wikitext resource was incorporating into WikAPIdia versions of all the standard indexing methods like stemming and stop word removal for all 25 language editions considered in this thesis. To do this, we leveraged and adapted the large number of language-specific tokenizers built into various extensions of Lucene. However, for two supported languages – Hebrew and Slovak – no such analyzer was available. In these cases, we used a tokenizer based on a language-neutral Unicode standard for word splitting. As we will see below, this resulted in a small but not-trivial hit to accuracy for the Hebrew and Slovak language editions of Wikipedia.

### **3.2.3 Results of the Parsing and Extraction Process**

In this section, we report the results of the parsing and extraction process in terms of the size and character of the resulting Wikipedia structures. Except where noted, all of the Wikipedia results in this thesis use data exported by the Wikimedia Foundation between October 25, 2012 (Polish language edition) and November 8, 2012 (Spanish and Dutch language editions). These were the latest database dumps available as of November 8, 2012.

As noted above, all of our multilingual Wikipedia work considers 25 different language editions of Wikipedia. These 25 language editions are listed in Table 3.2-b, along with the number of articles, categories, disambiguation pages, and redirects that were identified in each during the parsing process.

There are two important takeaways in Table 3.2-b. The first is the absolute scale of multilingual Wikipedia. Even just considering the 25 language editions we do here, there are 17.8 million articles, 3.6 million categories, 838K disambiguation pages, and 15.5 million

redirects. The second takeaway is the scale of multilingual Wikipedia relative to that of the English Wikipedia. Even though the English Wikipedia is the oldest, largest, and most well-known of the language editions, multilingual Wikipedia is a much more extensive informational resource. As can be seen in the bottom row of Table 3.2-b, our 25-language dataset has 4.32 times more articles than the English Wikipedia alone, 3.64 times more categories, 6.00 times more disambiguation pages, and 2.77 times more redirects. Although it is early in the reporting of our results, this is a preliminary hint that the English-as-Superset hypothesis may be flawed.

<b>Language</b>	<b>Articles</b>	<b>Categories</b>	<b>Disam. Pages</b>	<b>Redirects</b>
Catalan	387,652	43,430	8,771	258,786
Chinese	593,297	130,609	21,639	401,837
Czech	245,925	62,909	7,786	154,764
Danish	170,695	35,153	7,178	95,188
Dutch	1,128,396	72,362	54,819	471,184
English	4,136,587	922,603	143,605	5,609,176
Finnish	309,978	46,683	11,500	180,208
French	1,310,730	220,820	77,784	1,221,238
German	1,482,653	159,987	172,589	1,024,890
Hebrew	141,782	30,642	8,077	123,618
Hungarian	228,592	40,179	10,120	134,575
Indonesian	206,440	51,331	2,082	189,797
Italian	981,586	178,487	47,241	468,030
Japanese	835,805	102,607	35,907	492,375
Korean	221,266	94,959	13,679	174,598
Norwegian	347,057	78,145	10,474	191,038
Polish	928,183	104,541	50,504	275,483
Portuguese	759,326	148,613	27,150	577,255
Romanian	219,693	73,488	6,879	185,931
Russian	927,782	220,709	44,662	982,021
Slovak	178,934	42,896	2,775	49,138
Spanish	960,048	190,712	37,871	1,358,787
Swedish	560,943	136,549	19,035	435,816
Turkish	199,102	95,838	5,959	203,806
Ukrainian	414,481	69,498	11,643	279,610
<b>TOTAL</b>	<b>17,876,933</b>	<b>3,353,750</b>	<b>839,729</b>	<b>15,539,149</b>
<b> 25-language  /  English  ratio</b>	<b>4.32</b>	<b>3.64</b>	<b>6.00</b>	<b>2.77</b>

Table 3.2-b: The size of the Wikipedia article, category, disambiguation page, and redirect resources, using data from late October / early November 2012.

Similar patterns are revealed when examining other multilingual Wikipedia resources. For example, consider Table 3.2-c, which contains a number of statistics describing the WAG of each language edition. Here again we see the absolute scale of multilingual Wikipedia. Our 25-language dataset has over *1.04 billion* links. Table 3.2-c also illustrates the extent to which multilingual Wikipedia can dwarf the English Wikipedia alone. Our multilingual dataset has 3.56 times more links than the English Wikipedia, including 3.98 times more parseable links and 3.21 more unparseable ones.

Another clear trend in Table 3.2-c is the variation in the share of parseable and unparseable links across the language editions. The language edition with the largest share of parseable links is Spanish, with 87.7% of its links being visible to Wikipedia editors and XML parsers. The smallest share of parseable links belongs to the Romanian Wikipedia, in which only 15.7% of the links are parseable. Even among the large and well-established language editions, there is a great deal of variation in this respect. 30.8% of German links are unparseable, while 43.7% of French ones are. In the English Wikipedia – the language edition whose WAG is most-often studied and applied – a full 44.2% percent of links are unparseable. As noted above, this raises the distinct possibility that studies and systems involving the English WAG will have substantially different results when considering all links or just parseable links.

Before closing our discussion of parsing and extraction results, it is important to note that we validated that the parsing and extraction process was executed successfully by comparing our final offline content to that in the live version of Wikipedia. We were able to do so using WikAPIdia's built-in ability to access information from live Wikipedia's resources just as easily as it can from these resources' offline versions (Section 3.12). Of course, retrieving the live data is orders of magnitude slower than retrieving the parsed and extracted information, but the live

data is useful for several purposes, including parsing and extraction validation.

Our validation process consisted of randomly selecting 250 articles in each language edition and comparing the resources (e.g. outlinks, category memberships) available in these articles on the live version of Wikipedia with those in the parsed and extracted version. We found that the average difference between the set of outlinks found in the offline and online data sources was only 1.39%. Moreover, despite the fact that thousands of links were considered in each language edition, for no language edition was this figure higher than 3.56%. Investigating the causes of these differences, we found that in the large majority of cases, they represented content changes rather than errors. For all language editions, the date of the XML dump and the validation date were approximately one month apart, which was more than enough time for content to change<sup>13</sup>. For instance, the differences between the online and offline versions of the Chinese Wikipedia (2.86% of links) were due in large part to a change in a commonly used template, causing the unparseable links on several live pages to differ from their offline counterparts. Similarly, other differences in several language editions were caused by links whose destination had been deleted by editors in the time between the dump date and the validation date.

The average rate of divergence between the offline and online resources for category memberships was 1.33%, that for redirect destinations was 1.1%, and that for links between categories was 2.1% (including unparseable links). All of these rates were consistent with the change-induced differences we saw with outlinks.

---

<sup>13</sup> Note that we compared our offline versions to online versions as they appeared on the date of the parse, i.e. we used the historical version of each article last edited prior to the dump date. Despite this, changes that occurred in templates and the deletions of articles caused the bulk of the changes between the parsed and online versions.

Language	Total Links	# Parseable	%	# Unparseable	%	# Pot. Sub-article Reln's	%	% in 1 <sup>st</sup> Para.	% in 1 <sup>st</sup> Sec.
Catalan	28,799,280	10,756,330	37.3%	18,042,950	62.7%	89,668	0.3%	29.9%	43.2%
Chinese	34,225,641	14,713,274	43.0%	19,512,367	57.0%	189,136	0.6%	24.6%	40.8%
Czech	11,345,396	7,429,582	65.5%	3,915,814	34.5%	68,615	0.6%	23.3%	41.2%
Danish	5,587,203	3,599,746	64.4%	1,987,457	35.6%	36,453	0.7%	28.5%	57.8%
Dutch	40,104,569	19,702,027	49.1%	20,402,542	50.9%	191,376	0.5%	34.3%	50.6%
English	293,801,886	129,927,629	44.2%	163,874,257	55.8%	2,639,445	0.9%	19.5%	35.9%
Finnish	9,595,685	6,679,314	69.6%	2,916,371	30.4%	62,642	0.7%	27.2%	50.8%
French	78,945,999	44,465,815	56.3%	34,480,184	43.7%	830,529	1.1%	15.6%	28.7%
German	60,570,156	41,924,055	69.2%	18,646,101	30.8%	382,857	0.6%	17.3%	32.0%
Hebrew	10,584,355	5,970,465	56.4%	4,613,890	43.6%	60,952	0.6%	21.1%	36.6%
Hungarian	14,744,356	7,050,304	47.8%	7,694,052	52.2%	70,635	0.5%	18.4%	33.3%
Indonesian	8,746,907	3,290,195	37.6%	5,456,712	62.4%	43,495	0.5%	36.4%	54.2%
Italian	80,613,635	32,148,938	39.9%	48,464,697	60.1%	594,714	0.7%	16.1%	32.0%
Japanese	58,194,346	38,778,903	66.6%	19,415,443	33.4%	1,061,609	1.8%	15.9%	24.0%
Korean	10,836,996	5,182,965	47.8%	5,654,031	52.2%	98,735	0.9%	21.7%	37.1%
Norwegian	12,571,060	7,373,904	58.7%	5,197,156	41.3%	71,566	0.6%	28.9%	52.2%
Polish	47,012,161	22,908,642	48.7%	24,103,519	51.3%	308,601	0.7%	30.5%	51.0%
Portuguese	38,299,487	18,611,148	48.6%	19,688,339	51.4%	93,909	0.2%	30.0%	49.2%
Romanian	21,998,468	3,446,009	15.7%	18,552,459	84.3%	69,557	0.3%	29.9%	54.3%
Russian	59,793,029	25,951,300	43.4%	33,841,729	56.6%	414,759	0.7%	19.6%	35.8%
Slovak	6,140,037	3,093,489	50.4%	3,046,548	49.6%	74,497	1.2%	34.3%	53.2%
Spanish	37,679,922	33,027,435	87.7%	4,652,487	12.3%	544,652	1.4%	22.4%	41.0%
Swedish	25,879,880	19,212,736	74.2%	6,667,144	25.8%	183,937	0.7%	17.7%	74.2%
Turkish	10,256,172	4,042,678	39.4%	6,213,494	60.6%	36,365	0.4%	22.0%	51.0%
Ukrainian	38,500,445	8,348,658	21.7%	30,151,787	78.3%	224,800	0.6%	25.4%	50.3%
<b>TOTAL</b>	<b>1,044,827,071</b>	<b>517,635,541</b>	-	<b>527,191,530</b>	-	<b>8,443,504</b>	-	-	-
25-language  /  English  ratio	3.56	3.98		3.21		3.20			

Table 3.2-c: Statistics describing the size of the WAG-related structures in our dataset.

### 3.3 Concept Alignment

*Note: This work originally appeared in our paper in the Proceedings of the 30th ACM Conference on Human Factors in Computing Systems (CHI 2012) [9]. While much of the text here is original to this thesis, portions have been adapted from the original publication, of which my colleague Patti Bao and I were primary co-authors. Also, it is important to note that this work is based on database dumps from August 2011 rather than October/November 2012.*

Wikipedia articles are fundamentally language-specific entities. However, in order to investigate the cultural contextualization in multilingual Wikipedia, it is necessary group these language-specific structures into language-neutral representations. In other words, a fundamental prerequisite to the work in this chapter is knowing that the articles “Schokolade” (German) and “Chocolate” (Spanish) both describe chocolate, that the articles “United States” (English) and “Estados Unidos” (Spanish) both describe the United States, that the articles “American literature” (English) and “Literatura de Estados Unidos” (Spanish) both describe American literature, and so on. This section describes our approach to solving this key challenge in the analysis of the similarities and differences in multilingual Wikipedia.

In our work, we use a construct we call the *concept* as our language-neutral representation into which we convert language-specific articles. We define a concept to be the subject of at least one article in at least one language edition. Concepts can have articles in up to all 25 language editions. For example, the concept described by “Schokolade” (German) has articles in every language edition considered in this thesis, e.g. “Chocolate” (Spanish), “Chocolate” (English), “Chocolat” (French), and so on.

We make use of Wikipedia’s interlanguage link graph to identify and group articles about the same concept in different language editions. As noted above, interlanguage links (ILLs) are connections between articles in different language editions entered by humans and propagated by

bots. They are supposed to indicate near conceptual equivalence between pages in different languages. For instance, the article “Schokolade” (German) contains an ILL to “Chocolate” (English).

ILLs are typically viewed as pairwise dictionary-like entities, e.g. as done by Erdmann et al. [38] and Sorg and Cimiano [188]. Obviously, this approach is not compatible with the work here as it only supports two language editions. When more than two language editions are considered in the literature, the general assumption is that any two articles connected by a path in the ILL graph belong to the same concept [2, 82]. In other words, this assumption interprets connected components of the ILL graph as having a 1:1 relationship with concepts. Here again we found the literature insufficient for most of our research. The primary problem in this case is that the 1:1 relationship assumption ignores ambiguities in the ILL graph. Ambiguities occur when multiple articles in the same language edition are connected via articles in other language editions, meaning that a corresponding concept would have more than one article per language edition. While only 1.0% of connected components are initially ambiguous, many of them describe concepts that are of general and global interest because the potential for ambiguity increases as more language editions cover a given concept. This presents an important challenge for the majority of studies and applications in this thesis.

One major source of ambiguities in the ILL graph is conceptual drift across language editions. Conceptual drift stems from the well-known finding in cognitive science that the boundaries of concepts vary across language-defined communities [51, 207]. For instance, the English articles “High school” and “Secondary school” are grouped into a single connected component. While placing these two articles in the same concept may be reasonable given their overlapping definitions around the world, excessive conceptual drift can result in a semantic

equivalent of what happens in the children’s game known as “telephone.” For instance, chains of conceptual drift expand the aforementioned connected concept to include the English articles “Primary school,” “Etiquette,” “Manners,” and even “Protocol (diplomacy).” It would be incorrect in our studies to group together “Protocol (diplomacy)” (English) and “Primary school” (English) into the same concept. A similar situation occurs in the large connected component that spans the semantic range from “River” (English) to “Canal” (English) to “Trench warfare” (English), and in another that contains “Woman” (English) and “Marriage” (English) (although, interestingly, not “Man” (English)).

### **3.3.1 The Conceptalign Algorithm**

In order for our studies and applications to correctly handle this vital 1.0% of concepts, we needed an algorithm to split concepts that were subject to runaway conceptual drift. However, at the same time, the algorithm needed to respect the fact that different languages may define a concept more widely or narrowly than other languages.

Our approach draws on the conceptual spaces framework from cognitive science [51], in which a concept is a region in a multi-dimensional semantic space. Generally speaking, the higher the average semantic similarity between pairs of concept instances, the smaller the area of the concept. The goal of our approach is thus to split ambiguous concepts by dividing them into regions with higher average semantic similarity. One method would be to attempt to match the average semantic similarity of the 99% of concepts that are not ambiguous. Alternatively, a multilingual Wikipedia researcher or application designer may want to allow for slightly more conceptual drift (e.g. to include “High School” (English) and “Secondary School” (English) in the same concept), while at the same time eliminating cases like “Woman” (English) and

“Marriage” (English).

In order to enable this approach in this thesis and in practice, we developed an algorithm we call *Conceptualign* that allows us (and other multilingual Wikipedia researchers and application designers) to adjust the amount of allowable conceptual drift to suit the needs of a particular study or system. Our algorithm strategically removes ILL edges from ambiguous concepts, splitting connected components of the ILL graph into more coherent groups. Edges are removed along two dimensions: (1) limiting the number of edges from a given language that can point to the same article in another language (*MaxEdges*), and (2) using a voting scheme that requires a certain percentage of language editions to agree on an edge before it can remain (*MinLangs*). Finally, to measure the semantic similarity of multilingual articles generated by our algorithm, we developed a version of the *MilneWitten* semantic relatedness<sup>14</sup> measure that allows for cross-language semantic relatedness calculation (Chapter 6). This measure can be used to calculate the semantic similarity between pairs of articles that make up the newly generated concepts, regardless of the languages in which those articles are written.

### 3.3.2 Exploring Parameters

To better understand the ability of our algorithm to generate cohesive concepts as well as its ability to allow some flexibility in that cohesiveness, we randomly selected 2,000 ambiguous concepts from our dataset and performed a grid search on the parameters. To establish a reasonable upper-bound, we also randomly selected 2,000 unambiguous concepts with articles in two or more languages. For both groups of articles, we calculated the pairwise in-concept semantic similarity for each possible article pair. As a baseline, we did the same for the default

---

<sup>14</sup> Although semantic relatedness and semantic similarity are distinct ideas, semantic relatedness measures are often used to approximate semantic similarity and vice versa (see Section 6).

state of the ILL graph. For the default state of the ILL graph and the output of our algorithm, we also report the mean “out-concept” similarity, which is the average similarity of articles not in the same concept. Setting *MaxEdges* to any value other than one significantly reduced the average in-concept semantic similarity in all cases, so we only report data where *MaxEdges* = 1. Finally, in order to provide an additional perspective on our algorithm’s performance, we evaluated our results against the comparable portions of de Melo and Weikum’s bilingual German/English dataset [130].

As shown in Table 3.3-a, using our algorithm it is possible to match and even exceed the semantic cohesiveness of non-ambiguous concepts, at least with our 25-language dataset. Moreover, for the parameters that result in these high average similarities, performance on the de Melo and Weikum dataset matches and exceeds that of de Melo and Weikum’s algorithm. This is true even though our algorithm is far simpler than their complex linear program solution, although their work is focused on graph theory aspects of the problem. Table 3.3-a also shows that our algorithm provides significant leeway in allowing for more conceptual drift, meeting the

<b>MaxEdge</b>	<b>MinLang</b>	<b>In-Concept Similarity</b>	<b>Out-Concept Similarity</b>	<b>de Melo Accuracy</b>
1	0%	0.65	0.29	73.7
1	20%	0.67	0.29	77
<b>1</b>	<b>50%</b>	<b>0.73</b>	<b>0.3</b>	<b>81.2</b>
1	70%	0.78	0.31	87.5
1	90%	0.81	0.33	91.4
1	100%	0.82	0.41	87.5
ILL Graph		0.41	0.26	51.2
Unambiguous Articles		0.78	n/a	n/a
de Melo Algorithm		n/a	n/a	89.7

Table 3.3-a: Ambiguity levels of the concepts output by our concept alignment algorithm. Bold indicates the parameters used in this thesis.

second goal for the algorithm.

The question then becomes, which parameters should we choose for this thesis? We initially used  $\text{MaxEdges} = 1$  and  $\text{MinLangs} = 70\%$ , matching the semantic similarity of unambiguous concepts. This effectively normalized ambiguity across our entire dataset. However, after examining hundreds of concepts split by our algorithm set to these parameters, we determined that it was too strict to meet our thesis's goal of respecting diversity in concept definitions. For instance, "High school" (English) and "Secondary school" (English) were split into separate concepts, even though in many languages these concepts are one and the same.

By reducing  $\text{MinLang}$  to 50%, we found that we could still maintain a high in-concept similarity, while also including these two articles in the same multilingual concept. Moreover, the algorithm with these parameters had no trouble splitting runaway conceptual drift cases like "Woman" and "Marriage," "River" and "Trench warfare," etc.

After applying *Conceptualign* using  $\text{MinLang} = 50\%$  to the October / November 2012 dataset, we found that the 17.9 million articles in our 25 language editions formed exactly 8,669,484 concepts. It is this set of concepts that we use for all the studies of multilingual Wikipedia below.

### **3.4 Concept-level Diversity**

The application of *Conceptualign* to our 25-language dataset allows us to commence the process of measuring the extent of world knowledge diversity in multilingual Wikipedia. We begin by analyzing *concept-level diversity*, or the similarities and differences in the set of concepts for which each language edition has articles. The global consensus hypothesis – the widespread belief that encyclopedic world knowledge is roughly the same across cultures –

suggests that each language edition has articles about roughly the same set of concepts, when controlling for language edition size. On the other hand, the global diversity hypothesis – the much less-common assumption that encyclopedic world knowledge varies across cultural boundaries – suggests that the language editions cover their own, unique set of concepts, with even small language editions having information on concepts about which there are no articles in large language editions.

Our primary tool of analysis for understanding and reporting concept-level diversity is a straightforward metric we call *conceptual coverage*. A concept that is covered by an article in just a single language edition – a *single-language concept* – is defined to have a conceptual coverage of exactly one. A concept that is covered by articles in two language editions has a conceptual coverage of two, and so on. Concepts whose conceptual coverage is 25 are *global concepts*<sup>15</sup> and are covered by all language editions in this thesis.

Let us consider a situation in which we find that most concepts have very high conceptual coverage (when controlling for size of the language editions). This would suggest that the language-defined cultural communities behind each language edition are largely in agreement as to what concepts belong in encyclopedia world knowledge. In this case, we would say that the concept-level diversity across the language editions of Wikipedia is small and we would have evidence in support of the global consensus hypothesis. If, on the other hand, we find that most concepts have a very low conceptual coverage, this would suggest the opposite, or that each language-defined community includes its own, largely unique set of concepts in its repository of encyclopedia world knowledge. In other words, in this latter situation, the concept-level diversity

---

15 We use the term “global” for simplicity’s sake. These concepts may not be truly “global” in that they may not be covered in the language editions we do not consider here. As noted later in the thesis, exploring the many smaller language editions not included in our datasets is a direction of future work.

would be large and the global diversity hypothesis would be supported.

In order to determine for which hypothesis there was more support, we calculated the conceptual coverage for all 8.67 million concepts identified by *Conceptualign*. Table 3.4-a shows the number and percentage of single-language and global concepts in each language edition. From Table 3.4-a it is clear that there is substantial concept-level diversity in multilingual Wikipedia and that the encyclopedia of each language-defined community covers a set of concepts that differs extensively from that of the others. Most notably, every language edition contributes no less than 28,000 single-language concepts, and single-language concepts make up at least 15.9% of each language edition. This includes small language editions like Hebrew, Danish, and Romanian. Moreover, global concepts make up no more than 8.18% of any language edition.

<b>Language</b>	<b># Single</b>	<b># Global</b>	<b>% Single</b>	<b>% Global</b>
Catalan	81,589	10,853	21.66%	2.88%
Chinese	221,995	10,853	39.01%	1.91%
Czech	61,925	10,853	26.15%	4.58%
Danish	41,414	10,853	25.50%	6.68%
Dutch	459,876	10,853	42.98%	1.01%
English	2,141,677	10,853	53.76%	0.27%
Finnish	74,948	10,853	25.27%	3.66%
French	356,898	10,853	29.07%	0.88%
German	506,071	10,853	38.78%	0.83%
Hebrew	28,511	10,853	21.49%	8.18%
Hungarian	56,764	10,853	26.09%	4.99%
Indonesian	97,483	10,853	47.89%	5.33%
Italian	258,613	10,853	27.80%	1.17%
Japanese	419,545	10,853	52.65%	1.36%
Korean	59,583	10,853	28.90%	5.26%
Norwegian	92,245	10,853	27.55%	3.24%
Polish	247,141	10,853	28.28%	1.24%
Portuguese	209,429	10,853	28.74%	1.49%
Romanian	33,800	10,853	15.95%	5.12%
Russian	306,410	10,853	34.87%	1.23%
Slovak	32,338	10,853	18.42%	6.19%
Spanish	241,420	10,853	26.31%	1.18%
Swedish	179,092	10,853	33.22%	2.01%
Turkish	63,840	10,853	33.22%	5.65%
Ukrainian	97,364	10,853	24.30%	2.71%

Table 3.4-a: The number and percent of concepts that are single-language and global concepts in each language edition. Even small language editions like Hebrew contribute a large number of single-language concepts.

Examining the conceptual coverage distribution across all 8.67 million concepts provides another perspective on the extensive concept-level diversity in multilingual Wikipedia. Figure 3.4-a shows this distribution. The left-most data point in the figure indicates that single-language concepts make up *over 73 percent* (73.48%) of all concepts. In other words, 73.48 percent of concepts in multilingual Wikipedia are described by an article in only *one* of the 25 language editions. Figure 3.4-a also makes salient the fact that as conceptual coverage increases, the number of concepts at each level of conceptual coverage decreases rapidly. Only 7.95% of concepts appear in 5 or more language editions, only 3.03% appear in 10 or more, and only 0.47% appear in 20 or more. Most surprisingly, as indicated in Table 3.4-a, only 10,853 concepts (0.12%) are 25-language global concepts.

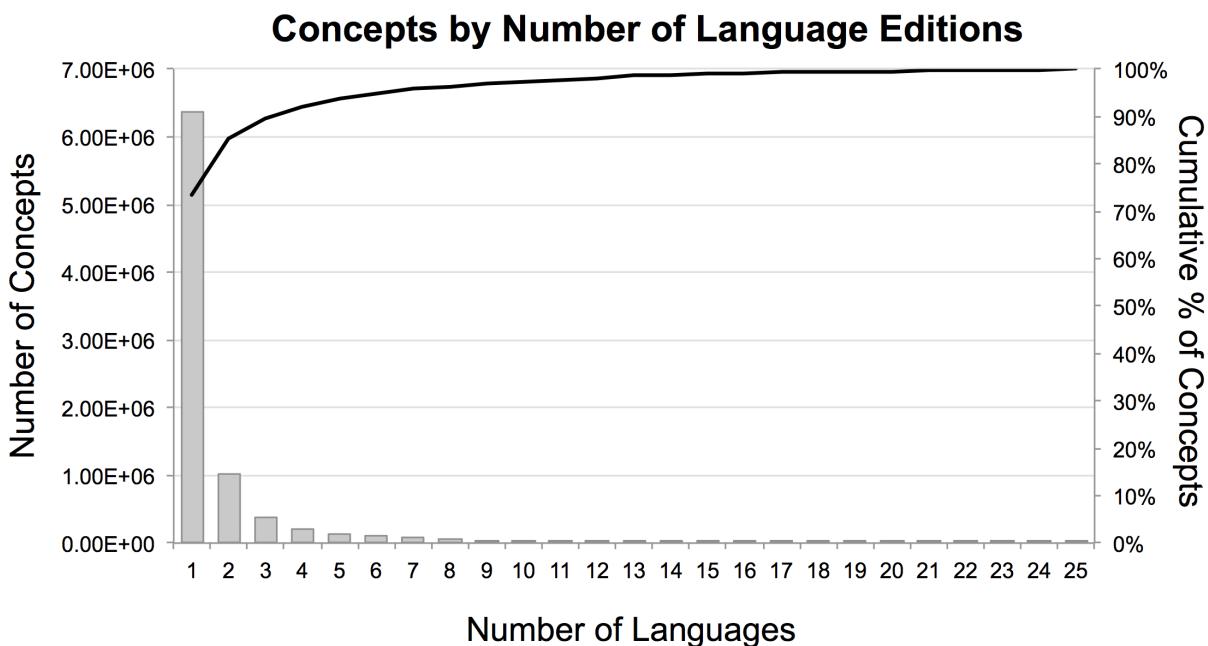


Figure 3.4-a: Concept-level diversity in multilingual Wikipedia. The number of languages in which a concept appears is on the x-axis. The y-axis indicates the percentage of concepts that appear in the corresponding number of languages. Over 73 percent of concepts appear in only a single language, while only 0.12 percent of concepts appear in all 25 languages considered here.

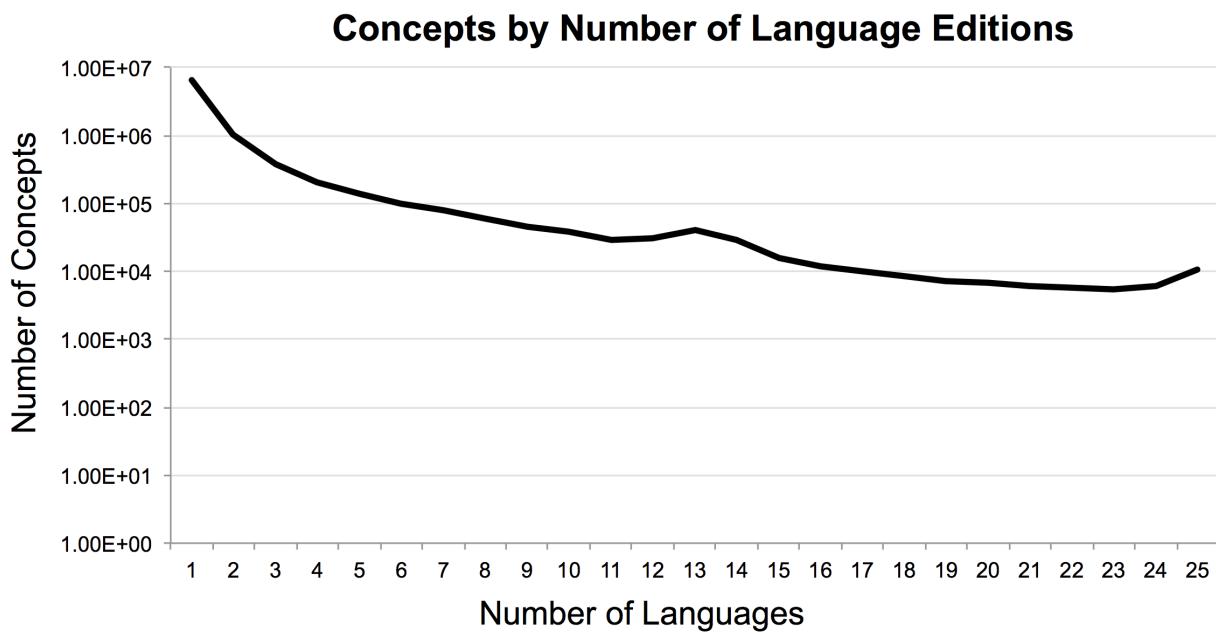


Figure 3.4-b: Here, the y-axis of the above figure has been converted to a log scale in order to highlight the variation in the number of concepts that appear in large numbers of language editions.

Converting the y-axis of Figure 3.4-a to a log scale as is done in Figure 3.4-b allows for a closer examination of variation in the distribution of concepts with a high conceptual coverage. Figure 3.4-b reveals two “peaks” of conceptual coverage, one at about 12-14 languages per concept and one that occurs for global concepts. With regard to the former, the likely causes are unclear and are the subject of future work. The global concepts peak, however, is more easily interpretable. This peak suggests that global concepts are not just “any old” group of concepts that have the same level of conceptual coverage. Rather, the fact that the conceptual coverage distribution has an outlier for global concepts hints at a *global encyclopedic core* with properties that differ from groups of concepts that are covered by smaller numbers of language editions. This core is made up of concepts that are in the cultural contexts of all language-defined cultures considered here.

In order to evaluate the robustness of these findings against variation in the language editions considered and against language edition size, we performed the same analysis on a number of different language edition collections. The results of these analyses can be found in Table 3.4-b. No matter the group of language editions and the size of these language editions,

<b>Language Set</b>	<b>Percent Single-language</b>	<b>Percent in All Lang. Editions</b>
All 25 lang. editions, sans English	75.13%	0.17%
Largest 10 language editions	73.89%	1.07%
Largest 5 language editions	76.12%	3.29%
Largest 3 language editions	78.79%	8.06%
Smallest 5 language editions	84.69%	2.13%
Smallest 3 language editions	86.29%	5.23%

*Table 3.4-b: The percentage of single-language and global concepts (where global = the number of language editions) in a number of different language edition collections. No matter the collection or the size of the language edition, single-language concepts far outnumber those that appear in all language editions considered.*

there is a great deal of concept-level diversity. In every case, single-language concepts form the large majority of concepts and far outnumber concepts that appear in all language editions considered.

In addition to the global consensus hypothesis generally, we can also examine its English-as-Superset corollary in the context of conceptual coverage distributions. Let us assume for a moment that the English-as-Superset hypothesis is true and that we can extend the hypothesis such that any language edition  $l_1$  covers all the concepts in any language edition  $l_2$  where the number of articles in  $l_1$  is greater than in  $l_2$ . In other words, the assumption is that English has all the concepts in German, which has all the concepts in French, which has all the concepts in Dutch, and so on. Figure 3.4-c depicts the conceptual coverage distribution that would result from this generalized English-as-Superset hypothesis and puts it in the context of the actual conceptual coverage distribution.

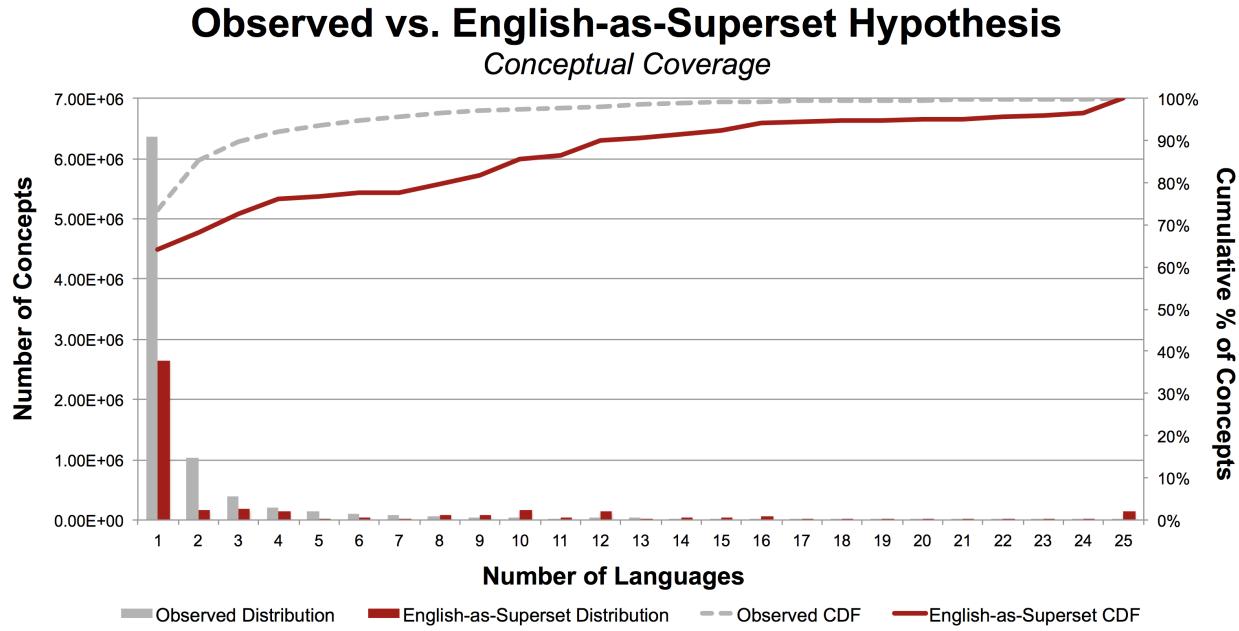


Figure 3.4-c: The observed conceptual coverage distribution compared to that which is predicted by the English-as-Superset hypothesis.

In Figure 3.4-c, we find further confirmation of that which is indirectly implied in Table 3.4-a: the English-as-Superset hypothesis is not in agreement with observations of real multilingual Wikipedia. While single-language concepts still make up a substantial percentage of concepts under the English-as-Superset hypothesis due to the size of the English Wikipedia relative to the next largest language edition (German), the number of *overall* single-language concepts, and the number of concepts generally, is much less under the English-as-Superset hypothesis than in our observations. In other words, if one restricts their system's or study's world knowledge to the scope of the English Wikipedia, one is excluding information about a very large number of concepts. In fact, comparing the two left-most bars in Figure 3.4-c, we found that only 42% of single-language concepts come from the English Wikipedia; the remaining 3.7 million single-language concepts come from the other language editions.

Another way to understand concept-level diversity is to examine the degree of overlap in

the concepts covered by any two language editions. For example, we might compare the number of concepts that are described in both the German and French Wikipedias to that of concepts described in only one of the two language editions. If concept-level diversity were low, we would expect that this overlap would be substantial. However, since we have found that concept-level diversity is extensive, we expect the overlap to be much smaller. Table 3.4-c, which depicts the overlap in the concepts covered in all 600 language edition pairs, shows that this is indeed the case. Each cell in Figure 3.4-c contains the fraction of the concepts from its row's language edition covered in its column's language edition. Looking at the (German, French) cell and the (French, German) cell we can see that these two language editions only cover 34-36% of the other's concepts. Put another way, there are no French articles on 66% of concepts that are described in the German Wikipedia and there are no German articles on 64% of the concepts in the French Wikipedia. This is particularly interesting because French and German are very comparable language editions: they are approximately the same size and are both old (relatively speaking), well-developed online encyclopedias with strong editor communities.

But what about English's coverage of French, German and the other language editions? As discussed above, the English-as-Superset hypothesis – a corollary to the global consensus hypothesis – suggests that English should cover all or nearly all of the concepts in the other language editions. Table 3.4-c adds to the strong evidence against this hypothesis. Consider the column that shows the percentage of the concepts in each language edition covered by the English Wikipedia. Here we see that English covers 76.5% of concepts in Hebrew, which is the highest concept overlap of all 600 pairs. Concepts in the Japanese Wikipedia are least covered by English, with only 42.2% of Japanese concepts having articles in English. This means that *there is no article in the English Wikipedia on anywhere from 23.5% to 57.8% of the concepts that are*

*covered in the other language editions.* The case of overlap between German and English – the two largest language editions – is quite illustrative. English is more than three times the size of German, but only covers slightly more than 50 percent of its concepts. In sum, any human or algorithmic consumer of Wikipedia information that only examines the English language edition is missing out on information about a significant number of concepts.

Another hypothesis that frequently arises when we present our concept-level diversity findings is what we call the “Well, it’s already in English...” hypothesis. This hypothesis suggests that editors of non-English language editions from language-defined communities in which English fluency is widespread will put the majority of their focus on concepts that are not covered in English. This would serve to exaggerate the concept-level diversity between the language editions and would complicate our ability to draw conclusions from multilingual Wikipedia about UGC reflecting the cultural contexts of its contributors. However, the fact that English covers 76.5% of Hebrew and only 42.2% of Japanese is good evidence that the hypothetical “Well, it’s already in English...” phenomenon is not a predominant driver of diversity between the language editions. English is very prominent as a second and even first language in Israel [189], meaning that most Hebrew speakers can access information in the English Wikipedia. The same cannot be said, however, about Japanese speakers; Japan has a level of English proficiency that is below average for all 34 OECD countries [224]. If “Well, it’s already in English...” were an outsized factor behind the concept-level diversity between the language editions, we would expect Hebrew to have much less in common with English. Certainly we would not expect Japanese to have the least in common with English of all the language editions, many of which are written by communities that, like Hebrew speakers, tend to have a very high level of English proficiency (e.g. Danish, Dutch, German, Norwegian) [224].

Covered Language Edition

Covering Language Edition

	Cata	Chin	Czec	Dani	Dutc	Eng	Fin	Fren	Ger	Heb	Hun	Indo	Ital	Japa	Kor	Nor	Pol	Port	Rom	Rus	Slvk	Spa	Swe	Turk	Ukr	Min	Max
<b>Catalan</b>		.331	.169	.131	.493	<b>.665</b>	.205	.529	.395	.129	.155	.123	.471	.241	.146	.180	.399	.448	.209	.349	.191	<b>.640</b>	.338	.134	.290	.123	<b>.665</b>
<b>Chinese</b>	.219		.122	.095	.319	.535	.141	.373	.310	.091	.111	.106	.344	.299	.165	.148	.308	.303	.162	.291	.135	.331	.226	.100	.218	.091	.535
<b>Czech</b>	.268	.293		.201	.425	.653	.306	.509	.535	.209	.237	.170	.468	.336	.213	.277	.469	.369	.183	.448	.248	.416	.352	.191	.264	.170	.653
<b>Danish</b>	.303	.332	.293		.481	.666	.374	.534	.563	.237	.248	.219	.483	.379	.257	<b>.429</b>	.472	.431	.233	.456	.182	<b>.461</b>	<b>.473</b>	.241	.275	.182	.666
<b>Dutch</b>	.174	.170	.094	.073		.463	.110	.332	.289	.061	.086	.064	.285	.138	.071	.119	.271	.252	.111	.224	.101	.287	.182	.073	.140	.061	.463
<b>English</b>	.063	.076	.039	.027	.124		.052	.191	.171	.025	.037	.026	.152	.084	.034	.056	.139	.120	.040	.116	.031	.145	.083	.029	.055	.025	.191
<b>Finnish</b>	.260	.271	.244	.205	.397	.698		.537	.531	.186	.213	.162	.468	.352	.205	<b>.305</b>	.453	.397	.164	.445	.137	.439	.423	.186	.232	.137	.698
<b>French</b>	.162	.173	.098	.071	.289	.621	.130		<b>.358</b>	.066	.091	.064	.345	.180	.081	.124	.276	.254	.100	.252	.086	.306	.197	.070	.138	.064	.621
<b>German</b>	.114	.135	.097	.070	.237	.521	.121	<b>.337</b>		.060	.079	.056	.270	.158	.072	.120	.247	.197	.076	.231	.062	.229	.170	.065	.104	.056	.521
<b>Hebrew</b>	.365	.388	.373	.291	.490	<b>.765</b>	.415	.608	.594		.293	.251	.543	.467	.320	.372	.530	.493	.253	.549	.207	.547	.458	.297	.342	.207	.765
<b>Hungarian</b>	.268	.290	.258	.185	.423	.671	.290	.514	.475	.178		.159	.500	.326	.197	.268	.463	.416	.208	.412	.204	.415	.342	.191	.309	.159	.671
<b>Indonesian</b>	.228	.297	.198	.174	.336	.506	.236	.383	.358	.164	.170		.355	.304	.218	.221	.329	.326	.183	.331	.135	.346	.278	.179	.220	.135	.506
<b>Italian</b>	.191	.211	.119	.084	.328	.651	.149	.455	.379	.077	.117	.078		.198	.093	.146	.343	.323	.137	.302	.118	.348	.227	.086	.186	.077	.651
<b>Japanese</b>	.114	.213	.100	.077	.185	<b>.422</b>	.131	.277	.258	.078	.089	.078	.231		.138	.123	.207	.193	.074	.214	.056	.218	.155	.076	.116	.056	.422
<b>Korean</b>	.266	<b>.455</b>	.245	.203	.369	<b>.650</b>	.295	.481	.453	.206	.208	.215	.418	<b>.534</b>		.294	.391	.377	.184	.422	.149	.419	.329	.206	.259	.149	.650
<b>Norwegian</b>	.203	.252	.196	.208	.381	.666	.270	.456	.468	.147	.174	.135	.404	.293	.181		.391	.340	.154	.377	.114	.361	.369	.147	.201	.114	.666
<b>Polish</b>	.172	.200	.127	.088	.332	.633	.154	.387	.369	.080	.115	.077	.365	.189	.092	.150		.300	.136	.311	.123	.304	.229	.093	.203	.077	.633
<b>Portuguese</b>	.231	.237	.120	.096	.369	.656	.162	.429	.353	.090	.124	.091	.412	.211	.107	.156	.360		.156	.323	.128	.436	.247	.109	.217	.090	.656
<b>Romanian</b>	.372	.434	.205	.179	.559	.757	.230	.582	.468	.159	.214	.176	.601	.277	.179	.243	.560	.536		.472	.284	.496	.411	.181	.459	.159	.757
<b>Russian</b>	.149	.188	.121	.084	<b>.1272</b>	.527	.150	.352	.344	.083	.102	.077	.320	.194	.099	.144	.310	.268	.114		.083	.289	.189	.099	.240	.077	.527
<b>Slovak</b>	.411	.437	.335	.168	<b>.1616</b>	.706	.231	.602	.460	.157	.253	.156	.623	.254	.175	.218	.611	.531	.343	.418		.498	.457	.198	.504	.156	.706
<b>Spanish</b>	.262	.205	.107	.082	.334	.632	.142	.410	.326	.079	.098	.077	.353	.190	.094	.132	.289	.346	.115	.277	.095		.220	.085	.167	.077	.632
<b>Swedish</b>	.236	.239	.155	.143	.362	.613	.233	.449	.412	.113	.138	.105	.392	.229	.126	.229	.371	.333	.161	.308	.149	.374		.113	.216	.105	.613
<b>Turkish</b>	.262	.296	.235	.203	.408	.600	.287	.446	.444	.205	.216	.189	.415	.316	.221	.256	.421	.414	.200	.455	.181	.405	.316		.286	.181	.600
<b>Ukrainian</b>	.273	.309	.156	.112	.373	<b>.547</b>	.172	.422	.340	.113	.168	.112	.433	.230	.133	.168	.444	.394	<b>.527</b>	.220	.384	.290	.137		.112	<b>.547</b>	
<b>Minimum</b>	.063	.076	.039	.027	.124	<b>.422</b>	.052	.191	.171	.025	.037	.026	.152	.084	.034	.056	.139	.120	.040	.116	.031	.145	.083	.029	.055	.025	.191
<b>Maximum</b>	.411	<b>.455</b>	.373	.291	.616	<b>.765</b>	.415	.608	.594	.237	.293	.251	.623	<b>.534</b>	.320	.429	.611	.536	.343	<b>.549</b>	.284	.640	.473	.297	.504	.237	<b>.765</b>

Example: English covers (has) 42%  
of the concepts in Japanese

Example: Russian covers (has) 53%  
of the concepts in Ukrainian

Table 3.4-c: Pairwise concept coverage overlap between all 600 language edition pairs. The examples demonstrate how to read the matrix. The red results are discussed in more detail in the text

The final pattern present in Table 3.4-c that we will discuss is the tendency of similar language-defined cultures to have Wikipedias that have relatively high concept overlap with one another. Table 3.4-c displays numerous examples of this phenomenon:

- The Chinese Wikipedia covers more of the Korean Wikipedia than any other language edition.
- The same is true of the Japanese Wikipedia with regard to the Korean Wikipedia.
- Spanish (1/4 the size of English) covers almost as much of the Catalan Wikipedia as English does.
- The same is true of the Russian and Ukrainian Wikipedias, with Russian also being approximately 25% the size of English.
- The Scandinavian language editions generally have relatively high coverage of one another. For instance, the Norwegian and Swedish Wikipedias cover more of Danish than any other language edition, with Norwegian doing so by a significant margin.

We will see throughout this chapter that concept overlap is not the only way these groups of language editions are similar, suggesting that similar language-defined cultures encode similar content in their representations of encyclopedic world knowledge.

Above, we have examined multilingual Wikipedia's extensive concept-level diversity from a conceptual coverage perspective and from a concept overlap perspective. We now turn our attention to understanding concept-level diversity at a smaller scale: that of individual concepts. Below we ask questions such as, "What types of concepts are non-English single language concepts?" and "What types of concepts are single language concepts generally?" While we address these and similar questions in a detailed, structured fashion along a variety of dimensions (e.g. topic, article centrality, content consumption) in the sections that follow, it is helpful to provide some initial representative examples for discussion purposes.

With regard to single-language concepts, the Austrian organic foods brand Ja! Natürlich is one of the German language edition's 506,671 German-only concepts. The same is true of Fachbereich Design der Fachhochschule Münster, one of the oldest design schools in Germany. German also has articles about 914 different people with the first name of *Ulrich* who are not covered in any other language edition, from Ulrich Adrian to Ulrich Zwetz. Among Spanish's 241,220 single-language concepts are the Spanish film, "*Manolito Gafotas ¡Mola ser jefe!*" and, not surprisingly, 1,145 people named José. Smaller language editions' articles about single-language concepts include "Aspargessauce" (Danish), which is the only page to cover this particular element of Danish cuisine, and "Customs de Girona" (Catalan), the exclusive article on a specific part of the legal history of the city of Girona, Spain.

The English Wikipedia has about 2.6 million single-language concepts. Like the above examples, many are stark reflections of the cultural contexts of the encyclopedia's editors. For instance, John Rich, one half of the modern country music duo Big & Rich, is an English-only single language concept. Similarly, "The Victors" (English) is the only article in multilingual Wikipedia dedicated to the University of Michigan fight song. "Landmarks in Omaha, Nebraska" (English) is the equivalent for notable locations in Omaha. Additional examples of single-language concepts can be found in Appendix A.

Moving rightward on the conceptual coverage spectrum we find concepts that exist in some language editions, but not all of them. Like is the case with single-language concepts, these concepts are often covered in language editions that seem implicitly appropriate from a language-defined culture perspective. Take, for instance, the Québec Capitales, a member of the "bush league" Canadian American Association of Professional Baseball (CAAPB) whose home stadium is in Québec City, Canada. The Capitales have a substantial article in English and

French, and a short one in Japanese. This is approximately the exact conceptual coverage one might expect given the cultural context of the Capitales, a team based in the heart of French Canada that plays a sport immensely popular in the United States and Japan. It is illustrative to compare the coverage of the Capitales to that of the Saint Paul Saints, a minor league team that plays in a much higher-level league. Despite their increased prominence, the Saints, based in the decidedly non-French-speaking state of Minnesota, are only covered by articles in English and Japanese.

The final set of concepts we will discuss in detail are the global concepts that make up the global encyclopedic core of multilingual Wikipedia. As will be discussed throughout this thesis, one advantage of mining cultural diversity from user-generated content is that doing so provides a clearer view of that which exists in the intercultural common ground. Although they are tiny in number (relatively speaking), the global concepts identified here exist in the intersection of the encyclopedic world knowledge encoded by a very large number of language-defined communities.

The 10,853 global concepts in our multilingual Wikipedia dataset are a diverse bunch. Nearly every country in the world has articles in all 25 language editions, as do other spatiotemporal concepts such as major cities and administrative districts, and many recent years and centuries. Pop stars also make up a sizable proportion of global concepts. Naturally, information about Justin Bieber is available to speakers of all 25 language editions considered here, and those who are not interested in Justin Bieber have equal access to articles about Katy Perry, Britney Spears, the Backstreet Boys, James Franco, One Direction, Beyonce, Nicki Minaj, Enrique Iglesias (although not his father), Pink, Jon Bon Jovi, and so on.

The encyclopedic core is – fortunately for some – replete with concepts of significantly

greater gravitas than the Backstreet Boys. Many key concepts in world history appear in the 25-language list (e.g. World War II, the Phoenician alphabet, Franz Joseph I of Austria, Benjamin Franklin), including topics that are generally perceived as highly controversial, especially in cross-cultural settings (e.g. the Holocaust, the Palestine Liberation Organization, Osama bin Laden, racism). Additionally, topics from science and technology, in particular those related to technology, are common in the global core. Alan Turing, Albert Einstein, all the periodic table elements, every major release of Windows since Windows 3.0, Steve Jobs, and Nikola Tesla all have articles in each of the 25 language editions considered here. An additional set of randomly selected global concepts (listed by their English language edition title) can be found in Appendix A.

In many cases, what is *not* in the global core can be equally informative as it can reveal that certain concepts that may be hypothesized to be well-known in many language-defined cultures in fact are not. For instance, even though Garth Brooks, another American country music star, has sold more records in the United States than any other artist besides the Beatles and Elvis [169], an article about him does not appear in Chinese, Czech, Hungarian and three others in our 25 language edition set. The same is true of country artists George Strait and Tim McGraw; they have far outsold many of the artists above in the United States [169], but they only appear in 14 and 15 languages respectively.

We have seen in the examples above that the cultural contexts of contributors to Wikipedia plays a role in the concepts that appear in the encyclopedic world knowledge defined by each language edition. However, it is important to note that conceptual coverage is not *entirely* defined by these easy-to-detect cultural signals. For instance, the fact that a given concept is single-language is many times not due to the knowledge of that concept being isolated to the

corresponding language-define culture, or at least not obviously so. Take for instance, the article “Adelaide River (Tasmanien)” (German), about the Adelaide River in Tasmania (as opposed to the one in the Northern Territories of Australia). While people of German heritage do make up a non-trivial proportion of the Tasmanian population [53, 197], from a purely language-defined cultural standpoint, this article should at least appear in English as well. Similar situations can be found along the entire conceptual coverage spectrum. One could argue, for instance, that the Capitales should have an article in the Spanish Wikipedia, as there is a large population of Spanish speakers who are at least as passionate about baseball as many English and Japanese speakers [72].

There are also methodological limitations to our concept-level diversity work. First, all of the above statistics must be considered in the context of our missing interlanguage link study [82]. This study found that as many as eight percent of same-concept article pairs are not recognized by *Conceptualign*<sup>16</sup> as being in the same concept due to missing interlanguage links. However, even if we assume that this eight percent figure persists across multilingual Wikipedia (and that it has not been reduced since the time the missing ILL study was run), a great deal of concept-level diversity would still exist. Consider the English-Italian cells in Table 3.4-b, for instance. The English-Italian language pair was the source of the eight percent missing ILL rate and even if we incorporate this rate into the table, we find that English would still only cover around 73% of Italian and Italian would only cover around 23% of English. The Japanese-English language pair is perhaps even more informative. In our Japanese-English missing ILL study, only 2% of concepts were affected by missing ILLs. If we incorporate this two percent

---

16 An older version of *Conceptualign* was used in this study. It is possible that a larger error rate could occur with the current version due to concept splitting, but given the relatively small number of concepts that are split, it is very unlikely that the splitting would result in a non-trivial change in the number of errors.

rate into Table 3.4-b, English would still only cover 44% of the Japanese Wikipedia, which would still be the lowest coverage of any language edition by the English Wikipedia.

The other methodological limitation that is important to consider is the ambiguity inherent in the results of *Conceptualign*. While we showed in Section 3.3 that *Conceptualign* is able to “split” cases of conceptual drift in a fashion that greatly increases the intra-concept semantic relatedness while at the same time barely affecting the inter-concept semantic relatedness, this does not mean that splitting issues do not exist. The *Conceptualign* parameters one chooses will also have an affect on the conceptual coverage of a number of concepts: more restrictive parameters will result in more, lower-coverage concepts while more cohesive parameters will result in fewer, higher-coverage concepts. However, all of this said, we noted in Section 3.3 that *Conceptualign* only gets applied on the significant minority of connected components in the ILL graph that are not complete (i.e. have interlanguage link conflicts), a number that is too small to significantly affect overall concept-level diversity. Moreover, when we consider diversity in multilingual Wikipedia over time in Section 3.9, we show that high concept-level diversity exists even when skipping the application of *Conceptualign*.

### 3.5 Sub-concept-level Diversity

We have seen that while most concepts exist in only one language edition, a substantial minority exist in two or more language editions. However, just because a concept is covered in two language editions, or 10 language editions, or 25 language editions does not mean the concept will be covered *in the same way* in these language editions. Each language edition could contextualize its discussion of the concept for its corresponding language-defined culture. Measuring this *sub-concept-level diversity*, or the diversity in the content of articles about the

same concept, is the subject of this section.

Narrowing things down a bit, what we are trying to accomplish here is to compare, for example, the articles “Schokolade” (German) and the “Chocolate” (Spanish), but do so across all 25 languages and at a massive scale. In this context, the global consensus hypothesis suggests that any two articles about the same concept should be roughly the same. In other words, according to the global consensus hypothesis, “Schokolade” (German) and “Chocolate” (Spanish) should describe chocolate in a very similar fashion.

The global diversity hypothesis, on the other hand, suggests that articles about the same concept are going to be different, with some of these differences arising from the cultural contextualization of the described concept. In other words, according to the global diversity hypothesis, “Schokolade” (German) and “Chocolate” (Spanish) should reflect the differences in the meaning of chocolate in each corresponding language-defined culture. For instance, the German article may discuss the history of Swiss chocolate while the Spanish article may discuss the history of cacao in Latin America. Alternatively, one of the two language editions could go into a great deal more detail about chocolate, at least partially because chocolate is more important to its editors’ cultural context.

The goal of the four studies in this section is to determine whether there is more support for the global consensus hypothesis or the global diversity hypothesis at the level of concept descriptions. In our first study, we take a multilingual Wikipedia-wide approach, asking questions related to the amount of shared content between same-concept articles, regardless of language edition. Next, we present two studies that examine sub-concept-level diversity at a language-by-language level. Here, we calculate the amount of content shared between same-concept articles from any pair of language editions and, specifically focusing on English,

investigate the level of support for the English-as-Superset hypothesis at the sub-concept level. Finally, we examine the extent to which diversity between articles exists even when controlling for the diversity and cultural contextualization inherent in article length.

However, prior to presenting our studies and their results, we first define our general sub-concept-level diversity methodology. We begin by showing how we use a “bag-of-links” document representation approach [99] to model the content of articles about the same concept in a fashion that allows for comparison across language editions. Although the bag-of-links approach has many advantages in the multilingual Wikipedia context, it also presents several challenges. The remaining two sub-sections in our methodology discussion are dedicated to our solutions to the two most serious of these challenges. First, we discuss the challenge of *missing links*, demonstrate its impact on the bag-of-links model in a Wikipedia context, and show how missing links can be “found” by adapting a monolingual approach known as *wikification* to a multilingual context. Second, we introduce the challenge of *sub-article relationships*, discuss its importance for both monolingual and multilingual work that directly involves concept descriptions, and demonstrate how we were able address this challenge by mining out sub-article relationships using machine learning techniques.

### **3.5.1 Sub-concept-level Methodology**

#### **3.5.1.1 Bag-of-Links Document Representation Model**

As noted above, in this section (and in several others that follow) we adopt what Joachim et al. [99] and others have called the “bag-of-links” (BOL) document representation model. At the core of this model is the assumption that the links on a page of hypertext can accurately represent the page’s overall content. Applied in the context of this section, a BOL representation of a

Wikipedia article is the set of its outlinks (or occasionally its inlinks) to other articles. Following the BOL assumption, the idea here is that the collection of other articles to which a given article links is a good summary of the content in the article.

Let us consider the “McCarthy, Alaska” (English) page in Figure 3.2-a. A BOL representation of this page would be a set of links including “Wrangell Mountains” (English), “Census-designed Place” (English), and “Chitina, Alaska” (English). Examining the actual text of the page, we see that the BOL assumption holds up in this case. That is, the outlinks in the top part of the page correspond to the two major themes of the top part of the page: what McCarthy is (a census-designed place) and where it is located (near Chinita and the Wrangells).

At least in Wikipedia, the BOL model has many advantages relative to the standard “bag-of-words” method of representing documents<sup>17</sup>. These advantages include (1) BOL models are highly structured and very human-readable, (2) the most important subjects in each article are usually the target of a link, and these links have been manually hand-annotated by Wikipedia editors in a very high-quality fashion (creating a less noisy dataset), (3) disambiguation issues have been resolved in a hand-curated way as well, and (4) we can gain large amounts of exogenous information about a given linked subject by accessing the various Wikipedia resources associated with the subject.

In the context of multilingual Wikipedia, however, the BOL model has one additional property that is essential to the research in this section: the links that are in Wikipedia (and are captured by the BOL) represent *semantic* relationships, abstracting away the difficult process of associating words in many different languages with the meaning of those words. The end result is that comparing a BOL representation of an article in one language edition – e.g. “McCarthy,

---

<sup>17</sup> The bag-of-words model also has its share of advantages (see Chapter 6).

Alaska" (English) – to that of an article about the same concept in another language edition – e.g. "McCarthy, Alasca" (Portuguese) – is as trivial as comparing two sets. Under the BOL assumption, this means that comparing the *content* of these two articles is as trivial as comparing two sets, as well. There are some additional important considerations here such as defining the high-dimension conceptual space in which the members of these sets exist (see Section 3.3). But after this preprocessing has been done, comparing encyclopedic world knowledge across many language editions of Wikipedia – and the study and application of multilingual Wikipedia in general – becomes much more straightforward using BOL representations.

The BOL model does, however, have two major drawbacks that, if not addressed, existentially challenges the assumption that links can accurately represent content. The first of these drawbacks related to missing links and the second involves sub-article relationships. The following two sub-sections address each of these issues in turn.

### 3.5.1.2 Missing Links

Missing links are a problem for a number of monolingual Wikipedia-based applications and research projects that use BOLs, but they are an issue of much greater magnitude in an

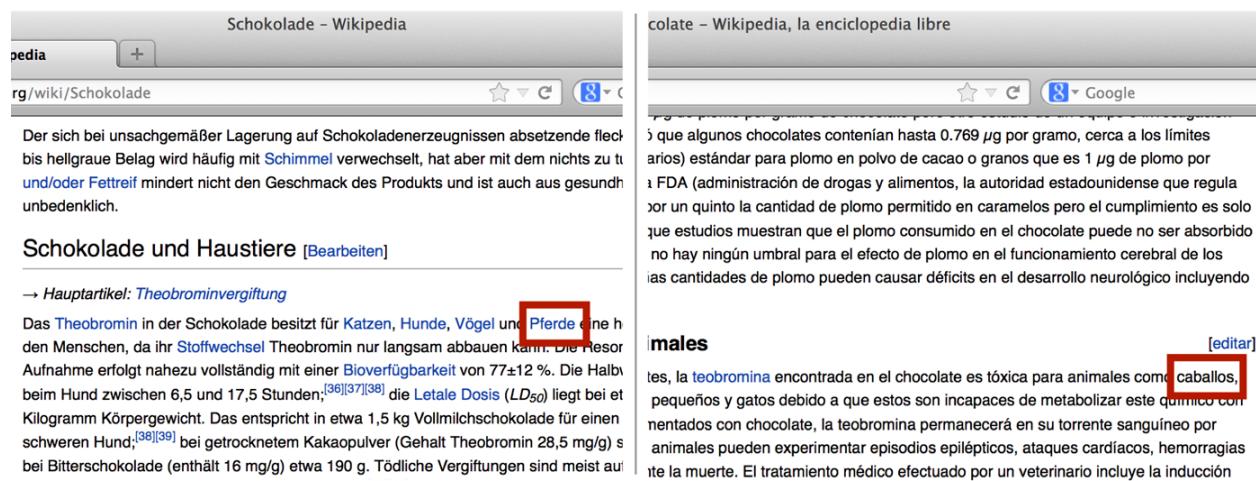


Figure 3.5-a: An example of a violation of the bag-of-links hypothesis.

integrative multilingual context. Consider, for example, the BOL representations of the articles “Schokolade” (German) and “Chocolate” (Spanish). According to the BOL assumption, we should be able to gain an accurate understanding of the similarities and differences in the content of these articles simply by comparing the links in the BOLs. However, in Figure 3.5-a we see that while the German article about Chocolate links to the German article about horses – “Pferde” (German) – the Spanish article *mentions* horses (“caballos”) *but there is no link*. In fact, no link to “Caballo” (Spanish) occurs anywhere in “Chocolate” (Spanish). This means that the Spanish BOL representation is not accurately representing the content of the article. Obviously, missing links like that to “Caballo” (Spanish) make it difficult to accurately compare the content about chocolate – or any other concept – across language editions.

The problem of identifying missing links in multilingual Wikipedia strongly resembles a task known as *wikification* [131, 137, 165]. The goal of wikification is to “identify and link expressions in [plain] text to their referent Wikipedia pages” [165]. This plain text can include non-linked content in Wikipedia itself, which presents the possibility of leveraging advances in wikification to help us address the BOL issues relevant to our work. This would amount effectively to “wikifying Wikipedia”<sup>18</sup>. Indeed, the existing literature on wikification could have likely provided the solution to our missing links problem, except for one show-stopping caveat: wikification has never to our knowledge been considered in a multilingual context, especially one involving so many large comparable corpora [159] like is the case with multilingual Wikipedia.

As such, in order to support our investigations into the sub-concept-level diversity in multilingual Wikipedia, we introduce several straightforward approaches to multilingual

---

<sup>18</sup> Phrase borrowed from the WebSAIL group at Northwestern University.

wikification in large sets of sizable comparable corpora. Our objective here is not to solve this problem, but rather to (1) meet the minimum needs of the studies in this chapter by establishing a means by which BOLs may be reliably compared and (2) initiate this line of research by establishing baselines comprised of relatively “low-hanging fruit” approaches. That said, wikification baselines like those we introduce here have been shown to be fairly robust and difficult to improve upon [165], so it is possible the performance reported below may approach that of more sophisticated methods.

### ***Straightforward Methods for Multilingual Wikification***

Basic, tokenized string matching<sup>19</sup> is the foundation of our technique for multilingual wikification in large sets of comparable corpora. While basic string matching suffers from significant disambiguation-related issues in many contexts, we hypothesized that the fact that we are dealing with comparable corpora would ameliorate some of these problems. Consider a situation in which we notice that the article “Northwestern University” (English) article links to “Chicago” (English), but that there is no such link in the article about the university in another language edition. Searching for the term “Chicago” in the other language edition’s article is much safer from a disambiguation standpoint than doing so in the wider web, as it is highly likely the referent of the term will be the city in Illinois, not the movie, musical, band, or Sufjan Stevens song with the same name.

The goal of our approach is to identify whether a link that occurs in one article’s BOL but not in that of another article about the same concept represents a true difference in content between the articles, or is simply a missing link. In other words, the intent here is to determine

---

<sup>19</sup> Text is processed and tokenized using the same best practices that are used to process the Wikitext resource (Section 3.2.2).

whether the “Chocolate” (Spanish) page really has no content about horses, or whether it does and just does not link to the Spanish article about horses. Moreover, we want to be able to do this for all articles in our 25-language dataset.

We conducted two experiments in which we evaluated the precision and recall of our basic technique with respect to this goal. Our first consideration was the overall experimental design. In the wikification literature, a commonly used and robust evaluation strategy involves measuring accuracy against manually labeled data (e.g. [137]). However, this method runs into a problem we often face in this chapter: finding several human labelers for each of many languages is a considerable challenge. Fortunately, we were able to ameliorate this issue by leveraging the large supply of labeled data we already have in the form of the links that *do* exist in each language edition. Following Mihalcea and Csomai [131], we strip each test article of all markup (indications of links) and then assess whether or not we can accurately rediscover the links in the article. If our basic algorithm can do this successfully, we can reasonably suggest that our algorithm can do the same in text that is missing links in the actual multilingual dataset. In other words, if we convert the article “Schokolade” (German) to plain text and successfully re-find the link to “Pferde” (German), we can assume that we could do the same with true missing links in the German Wikipedia. If we can do the same with different articles in Spanish, we can assume that the algorithm will also find the missing link to “Caballo” (Spanish).

Within this experimental framework, we designed one experiment to specifically focus on the overall recall of our wikification approach and another to focus specifically on precision<sup>20</sup>. The recall experiment simply evaluated the percentage of links that were “re-found” following

---

<sup>20</sup> It is common in the wikification literature to measure precision and recall (or proxies for these statistics) separately (e.g. [137]).

the stripping of the markup. The precision experiment was more complex. Prior to stripping all links from articles, we recorded the location of these links in the tokenized text. We then ran our wikification algorithm and recorded the location of the identified potential missing links. If the location of a potential missing link intersected with that of a true link, the targets of these links were compared. If they matched, the link was considered to be correct<sup>21</sup>.

It is important to note that, due to the multilingual nature of our work and the nature of its goals, we only attempted to “re-find” links that occurred in the target article *and* in at least one other article about the same concept. That is, for an article  $a$  about concept  $c$  in language edition  $l_m$ , we only attempted to re-find those links from  $a$  that occurred in  $a$  and at least one more article about  $c$  in  $l_n$  where  $n \neq m$ . This evaluation strategy perfectly matches the desired application of our wikification approach as it directly addresses the approach’s ability to ameliorate the effect of missing links in cross-language BOL comparisons.

For both experiments, we considered a variety of approaches for selecting the “queries” (i.e. candidates) for each potential missing link. The most basic query selection approach is to use only the title of the destination of the link (*WikipediaTitle*). In the experiments, this would involve simply searching for “pferde” in “Schokolade” (German), and would involve simply searching for “caballo” in “Chocolate” (Spanish) in the live application of the algorithm.

Another option is to include the redirects to the destination of the link as well as the title (*WikipediaTitle+Redirect*). Consider a situation in which we find that an article about concept  $c$  in Japanese links to an article about the United States, but that the article in Spanish about  $c$  does

---

21 This is a much stricter definition of precision than has been used in the wikification literature, which usually does not consider link position in its precision calculations ([131, 137]). We had to use this stricter definition due to the changes in the experiment design we made to accommodate the multilingual and comparable corpora properties of our application space.

not. Searching for the term “Estados Unidos” in the Spanish article, we do not find any matches. However, in this hypothetical situation, the negative result is due to the Spanish article using the abbreviated “EEUU” instead of the full “Estados Unidos.” Since there is a redirect from “EEUU” (Spanish) to “Estados Unidos” (Spanish), the *WikipediaTitle+Redirect* approach would find the missing link, whereas the *WikipediaTitle* approach would not.

A third technique involves incorporating the anchor text resource as well (*WikipediaTitle+Redirect+AnchorText*). For instance, in addition to searching for “Estados Unidos” and “EEUU,” this technique adds in all anchor texts for “Estados Unidos” (Spanish). This includes terms such as “estadounidense,”<sup>22</sup> “U.S.A.”, and “Washington”, but also terms such as “San Francisco”, “Florida”, and “Miss U.S.A”. Because this technique considers *all* expressions that are used to refer to a given concept in an entire language edition<sup>23</sup>, we hypothesized that the recall using this technique would be high, but that the precision would be low.

While powerful, all of the above methods of identifying queries for missing links have one important drawback. To understand this shared weakness, recall the German / Spanish example above. Although the *WikipediaTitle* method of generating queries would work in this case, it relies on there being an article in the Spanish Wikipedia that covers the same concept as “Pferde” (German), namely “Caballo” (Spanish). Without this article, we would have no idea what “Pferde” meant in Spanish, and thus would have no way of searching for the associated missing link. The same situation occurs with the other methods of generating queries for missing links; if there is not an article in the target language edition, there cannot be redirects to the article or

---

22 “American”

23 WikAPIdia also supports including only those anchor texts that occur with a probability greater than some threshold.

anchor texts associated with the article.

As another example, let us consider the article “Country rap” (English). This article links to “Cowboy Troy” (English), one of the more well-known country rap artists in the United States. Moving over to the Spanish Wikipedia, we find that it also has an article on the music genre, “Country rap” (Spanish), and that this article mentions Cowboy Troy, but does not link to him. If the Spanish Wikipedia had an article about Cowboy Troy, the *WikipediaTitle* strategy would work perfectly. However, the Spanish Wikipedia has no article. As such, the other Wikipedia-based query selection approaches would fail as well.

To address this issue, we turned to a large dataset of machine translated titles and redirects donated by Google Translate. This dataset consists of all articles and redirects in all 25 language editions translated into all 24 other language editions, or 882 million translations in total. We then use these translations just as we did the original titles and redirects in two analogous query selection strategies, *GoogleTranslateTitle* and *GoogleTranslateTitle+Redirect*. Using the *GoogleTranslateTitle* strategy, we find that Google accurately determined the Spanish translation of “Cowboy Troy” to be “Cowboy Troy,” as opposed to “Vaquero Troy,” “Gaucho Troy,” or “Troy, hijo de vaca”. As such, the *GoogleTranslateTitle* strategy is successful where all the Wikipedia-based strategies failed.

That said, the Wikipedia-based and Google Translate-based strategies have complementary benefits and disadvantages. The Wikipedia-based terms are manually curated and high-quality, whereas the Google Translate-based data is available in more languages but is of much lower quality. Additionally, the anchor text resource is only available directly from Wikipedia because, given its size, it was not included in our donation request to Google. In light of this complementarity, we have implemented our wikification approach and designed our experiment

such that both families of strategies can be used together in any combination (e.g  $\langle WikipediaTitle, GoogleTranslateTitle+Redirect \rangle$ ). For clarity, when only one family of strategy is used, we introduce the *WikipediaNone* and *GoogleTranslateNone* trivial strategies to serve as placeholders for the family of strategies that is not used.

We executed our recall experiment and our precision experiment using eight combinations of Wikipedia-based and Google Translate-based query selection strategies. Before beginning each experiment, we randomly selected 500 articles from each language edition and stripped them of their markup. These  $25 * 500$  articles were used for all conditions of each experiment.

Table 3.5-a shows the results of these experiments. With regard to recall (left side of the table), it is clear that despite the fact that our wikification approach is very basic, it performs quite well. As we expected, strategies that include Wikipedia anchor text outperform strategies that do not. In general, as one increases the number of queries from titles to redirects to anchor texts along both the Wikipedia and Google Translate axes, one gets better and better recall.

Indeed, the lowest recall strategy was  $\langle WikipediaTitle, GoogleTranslateNone \rangle$ , which could only recover about 80 percent of links on average, and the highest was

$\langle WikipediaTitle+Redirect+AnchorText, GoogleTranslateTitle+Redirect \rangle$ , i.e. the “kitchen sink” strategy, which recovered *almost every single link* (recall = 0.982).

It is important to note, however, that the upper-bound condition above only applies in cases where the destination of the potential missing link has a Wikipedia article in the language in which one is searching. That is, the 0.982 recall does not apply in cases like the Country rap / Cowboy Troy example above. For these situations, *WikipediaNone* is the only Wikipedia-based strategy that can be used. The maximum recall in these cases is that provided by the  $\langle WikipediaNone, GoogleTranslateTitle+Redirect \rangle$  condition, which is 0.86. In other words, on

average, our wikification algorithm using this query selection strategy will find Cowboy Troy-like links 86% of the time.

The recall section of Table 3.5-a also provides support for our hypothesis about the impact of poor indexing tools for Hebrew and Slovak. The effect is not enormous, but it is enough to make Hebrew or Slovak the language edition with the worst recall in all cases except those in which the *WikipediaNone* strategy was used. In the cases that it was, Chinese was the worst performing language edition in terms of recall, likely due to the challenges involved with Chinese machine translation.

Query Selection Strategy		Article-based Recall*			Article-based Precision			Link-based Precision		
Wikipedia	GTranslate	Recall	Min	Max	Precision	Min	Max	Precision	Min	Max
Titles	None	0.790	0.617 Slovak	0.885 Romanian	<b>0.841</b>	0.712 Slovak	0.897 Danish	<b>0.735</b>	0.566 German	0.866 Danish
Titles+Redirects	None	0.897	0.671 Slovak	0.972 Chinese	0.841	0.725 Slovak	0.897 Danish	0.727	0.575 German	0.857 Danish
Titles+Redirects +AnchorTexts	None	0.979	0.903 Hebrew	0.997 English	0.799	0.715 English	0.869 Danish	0.552	0.216 Spanish	0.759 Danish
None	Titles	0.779	0.524 Chinese	0.907 English	<b>0.807</b>	0.685 Slovak	0.879 Danish	<b>0.686</b>	0.540 German	0.836 Danish
None	Titles+Redirects	<b>0.860</b>	0.631 Chinese	0.955 Spanish	0.776	0.678 Chinese	0.831 Danish	0.625	0.486 Indonesian	0.760 Danish
Titles	Titles	0.877	0.676 Slovak	0.947 Romanian	0.828	0.719 Slovak	0.880 Danish	0.690	0.551 German	0.833 Danish
Titles+Redirects	Titles+Redirects	0.940	0.720 Slovak	0.984 Romanian	0.795	0.721 Slovak	0.851 Dutch	0.635	0.506 Indonesian	0.760 Danish
Titles+Redirects +AnchorTexts	Titles+Redirects	<b>0.982</b>	0.919 Hebrew	0.998 English	0.770	0.687 Russian	0.826 Dutch	0.524	0.217 Spanish	0.706 Danish

Table 3.5-a: The performance of our basic multilingual wikification system using a selection of query selection strategies. Rows in gray indicate performance in cases where there is no article that matches the destination of the potential missing link, leaving Google Translate data as the only available information from which to derive queries. Recall and precision are averaged across language editions.

\* Only article-based recall is reported as it was nearly identical to link-based recall. That is, there was no effect for length of article in the case of recall, whereas there was a substantial effect in the case of precision.

Moving over to the right side of Table 3.5-a, we report two types of precision: precision averaged by article and precision averaged by link. For the first type of precision, long and short articles are given equal weight, while the second type does not involve aggregating and normalizing by article. A theme that is immediately clear in the precision part of the table is the large drop from article-based precision to link-based precision. In all language editions, the article-based precision is relatively good and is actually on par with that reported in well-known monolingual wikification papers (e.g. [131, 137]), although the tasks considered have some differences. The same is not true, however, for the link-based precision.

There is only one possible cause of these precision differences: an effect for article length. Investigating the language-by-language results, we found that in most language editions, as the number of links to be re-found went up, the precision went down. Similarly, we also noticed that it was the longest articles in each language edition that nearly always had the worst precision values. Neither of these patterns existed for recall; the difference between the link-based and article-based recall was almost always less than one percent.

There are two major takeaways from these results. First, it appears that for short articles, our quite basic approach to multilingual wikification in comparable corpora works well, even across 25 different languages. This means our approach will likely prove to be a difficult (though not impossible) baseline to beat for short articles. That said, it is often the longer articles that are the most useful for any number of human or machine needs, so there is an enormous amount of work left to be done in this application space. Initial next steps should likely include improving precision by incorporating more sophisticated techniques from monolingual wikification such as semantic relatedness measures and less naïve anchor text approaches.

The second takeaway is the more important for our sub-concept-level diversity analyses.

Table 3.5-a reveals that our basic approach to multilingual wikification in comparable corpora can provide an *excellent upper-bound* for the comparison of article BOLs across language editions. Namely, the Table 3.5-a shows that the “kitchen sink”  $\langle WikipediaTitle + Redirect + AnchorText, GoogleTranslateTitle + Redirect \rangle$  strategy has near-perfect recall. The link-based precision of this strategy is relatively poor (0.55), but for many BOL comparison tasks, this does not affect the strategy’s usefulness as a robust upper-bound. As we will see many times in section (and this chapter), our ability to understand cultural contextualization in multilingual Wikipedia at the sub-concept level is greatly enhanced by being able to make statements such as, “The articles about concept  $c$  in language editions  $l_1$  and  $l_2$  share *at most* X% of their content.” In addition, this upper-bound can be combined with the inherent lower-bound that is the execution of BOL comparisons without any multilingual wikification (i.e. the “just links” strategy), allowing us to state ranges on the percentage of content shared between any two articles.

Put together, while it would be desirable to have a single, accurate value output from all BOL comparisons that is robust against missing links, the work in this section shows that it is possible to output at least a single, accurate *range* of values. For the purposes of this section, that is more than enough for us to be able to leverage the many benefits of working with BOLs, while avoiding the large pitfalls that can occur due to missing links.

It is important to point out two additional properties of our upper-bound multilingual wikification strategy. First, the upper-bound nature of the strategy is somewhat reduced when considering “Cowboy Troy”-like links as described above. The maximum recall in these cases drops from 0.982 to 0.860. While this recall is still quite high, there is a way to bring it up to the same level as above: filter out all links in the BOLs that do not link to concepts that exist in all

language editions being considered. That is, in the country rap example, this approach would involve not considering links to “Cowboy Troy” (English) at all, but only those links to concepts that have articles in both English and Spanish (e.g. Beck, Kid Rock). Below, we report our BOL-related results using both the typical upper-bound strategy as well as this “only-intersection” strategy.

Finally, due to the nature of our BOL comparison task, both the regular and only-intersection upper-bound wikification strategies will overestimate the amount of overlap between any two BOLs in almost every case. Of course, upper-bound approaches tend to overestimate in any number of domains, but this is especially true when using our upper-bound wikification strategies to do BOL comparisons. The reason this is the case is that when comparing two BOLs, our strategies’ precision errors can only serve to inflate the amount of overlap between the BOLs because the only missing links we are searching for are those that exist in one of the BOLs but not the other. As such, there will never be a mistakenly-found link that will reduce the amount of overlap, and due to the decidedly non-perfect precision of our upper-bound methods, the opposite will occur at a relatively high rate.

### 3.5.1.3 Sub-article relationships

In all language editions that appear in this thesis, when an article is considered too long by Wikipedia editors, specific topics that are a part of the original subject of the article are split off into what we call *sub-articles*. For example, in the English Wikipedia, the article “United States” has the sub-articles “History of the United States,” “Geography of the United States,” “Environment of the United States,” “American literature,” and so on. Similarly, the article “Caffeine” (English) has the sub-article “Health effects of caffeine” (English) (Figure 3.5-b). While sub-articles only occur with a small minority of articles, these articles are those that are long enough to warrant sub-articles, meaning that they tend to be about important subjects (e.g.

The screenshot shows the Wikipedia article page for "Caffeine". At the top, there's a navigation bar with links for Article, Talk, Read, Edit, View history, and a search bar. Below the title "Caffeine" and subtitle "From Wikipedia, the free encyclopedia", there are two main sections: "Health effects" and "Stimulant effects". A red box labeled "Potential Subarticle Relationship" is overlaid on the left side, pointing to the "Health effects" section. This section includes a diagram titled "Health effects of caffeine" showing a human head with arrows pointing to "Positive effects" (increased attention and alertness, decreased fatigue, lower risk of...) and "Negative effects" (anxiety and addiction, increased vasoconstriction and blood pressure). Another red box labeled "Potential Subarticle Relationships" is overlaid on the bottom left, pointing to the "History" section. This section lists "Main articles: History of chocolate, History of coffee, History of tea, and History of yerba mate". To the right of the "History" section is a small illustration of people in historical attire.

Figure 3.5-b: Potential sub-article relationships in the parent article “Caffeine” (English).

United States, caffeine).

Sub-articles represent an important problem for studies (e.g. those in this section) and systems (e.g. Omnipedia [9] and Manypedia [127]) that involve comparing articles about the same concept across different language editions. While one language edition may split an article's content into seven sub-articles, another may split its corresponding article into four, and yet another may keep all the content on a single page. To a study or system that ignores sub-articles, it could appear as if the language editions with the fewest sub-articles had the most content about a concept (i.e. the largest BOL), whereas the opposite is likely true. The editors of the “United States” (English) page, for instance, have eschewed going into great detail about historical topics, knowing that this content is covered in the “History of the United States” (English) sub-article. Another language edition that does not have an equivalent sub-article may go into more detail on its main United States page about the history of the United States, but it would be incorrect to say that this language edition covers the United States in more detail.

Making matters more complex, while every language edition considered in this thesis uses the sub-article relationship construct, each encodes these relationships using its own unique set of indicators. Additionally, in all these language editions, *not all indicated sub-article relationships represent true sub-article relationships*. That is, sub-article relationship indicators tend to be quite noisy. Significantly more so than in the Wikipedia resources discussed in Section 3.2, it appears that different editors (or even a single editor) use the various means of encoding sub-articles in different ways, some of which involve true sub-article relationships, others of which do not.

Consider the potential sub-article relationships that appear on the Caffeine (English) page in Figure 3.5-b. “History of Chocolate” (English), “History of Coffee” (English), and the other

articles that are indicated to be sub-articles at the bottom of the figure are clearly *not true sub-articles*. That is, even if article length were no issue, the content on these pages would not appear in the “Caffeine” (English) article. Compare this to the potential sub-article relationship between “Caffeine” (English) and “Health effects of caffeine” (English) article. In this case, the latter article is much more likely to be the result of an effort to split the subject of caffeine into multiple articles for length reasons. The same is true of the United States sub-article relationships discussed above.

In this sub-section, we describe how we addressed the challenges of (1) identifying potential sub-article relationships across multilingual Wikipedia and (2) determining which of these are “true” sub-article relationships. We also discuss the end result of this process: a classifier-based sub-article detection system built into WikAPIdia that we have used to support our sub-concept-level diversity research in this section and other research that appears in this thesis as well (e.g. Omnipedia in Chapter 7).

### ***Identifying Potential Sub-article Relationships***

Our overall approach to identifying all the means of indicating potential sub-article relationships in all 25 language editions was, as noted above, a brute force manual technique. A single investigator fluent in English and Spanish accessed thousands of pages in all 25 language editions, focusing on concepts that typically had sub-articles in many language editions (e.g. countries, major historical events). Although context is usually sufficient to identify a sub-article relationship, the investigator used Google Translate as an aid when necessary.

Whenever the investigator encountered a potential sub-article relationship, he recorded the *parent article* (e.g. “United States”), the potential sub-article (e.g. “History of the United

States”), and, most importantly, the Wiki markup that was used to encode the relationship. The final dataset consists of 3,083 such records. The most records for a language was 509 (English) and the least was 60 (Hungarian).

Using our dataset, we identified five general families of potential sub-article relationship indicators, four of which resemble the appearance of those in Figure 3.5-b. These four families are distinguished from one another only by the text preceding the indicated relationship and by their Wiki markup. The fifth type of indicator, however, is significantly different. Potential sub-article relationships encoded through this type of indicator are listed at the bottom of articles under a header titled “See also,” or its equivalent in another language (e.g. “Siehe auch,” “Véase También”). These potential sub-article relationships are quite common, and most Wikipedia readers will have seen examples of them, such as those in Figure 3.5-c. It was our hypothesis that while there are *some* true sub-article relationships embedded in these lists, the rate at which they occur is significantly less than is the case with the other four families. As we will see, this was a hypothesis supported by our results.

Africa – Wikipedia, the free encyclopedia

W Africa – Wikipedia, the free enc...

en.wikipedia.org/wiki/Africa

Brenthecht Talk Sandbox Preferences Watchlist Contributions Log out

Article Talk Read Edit View history Search

# Africa

From Wikipedia, the free encyclopedia

*For other uses, see [Africa \(disambiguation\)](#).*

**Africa** is the world's second-largest and second-most-populous [continent](#). At about 30.2 million

**Africa**

**See also** [edit]

- Outline of Africa
- Index of Africa-related articles
- Afro-Eurasia
- Highest mountain peaks of Africa
- List of African millionaires
- List of cities in Africa
- Urbanization in Africa

**Africa portal**

**Book: Africa**

Potential Subarticle Relationships

Figure 3.5-c: Potential sub-article relationships encoded in a “See also” list at the bottom of the parent article “Africa” (English).

Following the data collection process, the next step was to construct a large number of regular expressions based on our repository of sub-article relationship indicators. These regular expressions were then applied during the parsing of the XML database dump of each language edition, allowing us to extract nearly all potential sub-article relationships in all 25 language editions. The only error occurred with the Portuguese language edition, in which one omitted regular expression caused us to miss about 30% of potential sub-article relationships. We have fixed this bug and it will be incorporated in our next parse of the XML database dumps, which will occur after the completion of this thesis.

### **Developing a Model to Identify True Sub-article Relationships**

Once we had extracted all potential sub-article relationships, we could turn our attention to determining which of these potential relationships were “true” sub-article relationships. Our approach to this problem, as noted above, was to construct a classifier to automatically make this distinction. Below, we describe the training of our classifier and demonstrate that its accuracy is significantly better than baseline approaches.

The first step in building our classifier was the collection of training data. Our goal here was to develop a human gold standard dataset with potential sub-article relationships rated according to the extent they represented true sub-article relationships, i.e. the extent to which the sub-article would be on the same page as the parent article if length were no issue. We began this process by extracting 100 potential sub-article relationships from each the English and Spanish Wikipedias. Two English/Spanish bilingual coders were then recruited to assess each potential relationship.

Using a GUI interface we built that displays the first paragraph of each parent article/potential sub-article pair (Appendix B), our coders were asked to rate the corresponding relationship on a scale from zero to three, with three being a definite true sub-article and zero being the opposite. The exact instructions we gave with regard to the rating scheme were as follows<sup>24</sup>:

- 3: The *only* reason the potential sub-article exists is to split the corresponding main article into more manageable subtopics. The potential sub-article really *does not deserve its own page*, and the corresponding main article is the best place to put the sub-article’s content.
- 2: Same as above, but the potential sub-article’s topic is significant enough to warrant its own page.

---

<sup>24</sup> The full set of instructions given to coders can be found in Appendix C.

- 1: The potential sub-article contains information that would be useful to have on the main article, but contains its own, *unrelated (non-overlapping)* content.
- 0: The potential sub-article is on a topic that is *trivially related* to the main article or has a large amount of *non-overlapping content*.

While we anticipated that the rating task would be quite difficult given the nuance inherent to sub-article relationships, our two coders gave most potential relationships similar ratings. The Spearman’s correlation between the coders’ scores was 0.622. This is higher than the inter-rater reliability between coders of many semantic relatedness datasets (e.g. [217]), including some of the most commonly used ones (e.g. [43]).

For each of the 200 parent article/potential sub-article pairs in our ground truth dataset, we generated a number of different features that we predicted would be helpful in training a model to accurately identify true sub-article relationships. Prior to walking through some of these features, it is important to note that we interpret sub-article relationships as relationships between concepts, not just articles. That is if we determine that “United States” (English) and “History of United States” (English) have a sub-article relationship, then this relationship is applied to the versions of these articles in every language edition in which these versions exist (e.g. “Estados Unidos” (Spanish) and “Historia de los Estados Unidos de América” (Spanish)). We leverage this concept-level understanding of sub-article relationships in many (but not all) of our features. The concept-level and article-level features we considered are as follows:

- *NumLangsParent*, *NumLangsSub*, and *NumLangsRatio*: The number of languages in which the parent concept has an article, the number of languages in which the sub-article concept has an article, and the ratio between these values. For instance, the ratio in the case of the caffeine example would be the number of language editions in which there is an article about caffeine divided by the number of language editions in which there is an

article about the history of caffeine.

- $SR_{ESA}$ ,  $SR_{MW}$ , and  $SR_{WR}$ : The semantic relatedness between the parent article and the sub-article, as calculated by the Explicit Semantic Analysis ([47, 50]), Simplified *MilneWitten* ([136, 137]), and *WikiRelate* ([156, 192]) algorithms. See Chapter 6 for information about these algorithms.
- $PageRankRatio$ : The ratio of the (parseable WAG) PageRank scores of the parent article and potential sub-article (see Section 3.6 for more on PageRank scores).
- $PotSubarticleRatio$ : The ratio of the number of language editions in which there is a potential sub-article relationship to the number of language editions in which the parent concept and sub-article concept both have articles.
- $TokenOverlap$ ,  $RedirectTokenOverlap$ , and  $MaxTokenOverlap$ : These features consider the percentage of tokens in the parent article’s title contained within the potential sub-article’s title<sup>25</sup>. All values are averaged over all language editions in which both concepts exist. *TokenOverlap* just examines the titles of within-language parent/potential sub-article pairs. *RedirectTokenOverlap* does the same with redirects (the maximum such overlap is chosen in cases of more than one redirect). *MaxTokenOverlap* is equal to  $\max(TokenOverlap, RedirectTokenOverlap)$ .
- $SeeAlsoSectionPct$ : The percentage of potential sub-article relationships between the parent concept and the sub-article concept that occur in a section with the title “See also” (or its equivalent in other languages). The denominator here is the number of potential sub-article relationships overall that exist between the concepts.

In addition to identifying features, we also had to decide upon the value that would be most useful to predict. One obvious option would be to train on the mean scores from our two coders. However, this regression experiment is not well-suited to the requirements of this thesis. That is, our studies and systems need to know which articles are sub-articles of a given parent article and which are not. This implies that we need to predict a binary class rather than a continuous value. As such, we categorized each sample based on whether or not the mean score was greater than or

---

<sup>25</sup> The text in the titles is tokenized in the exact same fashion as the Wikitext resource.

Classification Model	Precision	Recall	F1	% Correct
Four-feature Logistic Regression	0.738	0.592	0.657	76.5
<i>MaxTokenOverlap</i> -only Logistic Regression	0.719	0.739	0.617	74.5
Random Baseline	-	-	-	62.0

Table 3.5-b: Performance of our final logistic regression sub-article classification models.

equal to 2.5. This means that for a sample to be understood as a true sub-article relationship, either both coders had to give it a three, or one coder gave it a three and the other a two. This threshold was chosen in order to create a relatively strict model while at the same time allowing for some variation in the coders' scores.

We investigated the performance of a number of different classifiers, focusing on decision trees and logistic regression as well as on developing a relatively parsimonious model. We did the latter to avoid overtraining of course, but also to allow for straightforward interpretation of how the model works. For training and testing of each classifier, we used 10-fold cross validation.

The most accurate model (Table 3.5-b) was a four-feature simple logistic regression classifier, which was able to predict true sub-article relationships (as defined above) with an accuracy of 76.5%, representing a 23.4% improvement in performance over the random baseline of 62.0%. The difference between the model and the random baseline was significant ( $\chi^2 = 9.87$ ,  $p < 0.01$ ). The four features included in the model are *PotSubarticleRatio*, *PageRankRatio*, *MaxTokenOverlap*, and *SeeAlsoSectionPct*.

While this four-feature model was the best-performing overall, a logistic regression classifier trained only on *MaxTokenOverlap* was nearly as accurate. It had an accuracy of 74.5%, misclassifying only four additional instances. Although the difference between these models is

Feature	Odds Ratio
<i>PotSubarticleRatio</i>	1.56
<i>PageRankRatio</i>	0.98
<i>MaxTokenOverlap</i>	10.14
<i>SeeAlsoSectionPct</i>	0.44

Table 3.5-c: Odds ratios of our final four-feature logistic regression model. The odds ratio for MaxTokenOverlap in the MaxTokenOverlap-only model was 14.06.

insignificant given the relatively small sample size, we used the more complex model in our studies and applications both because the performance hit is minimal and because of the small likelihood of an increase in performance.

The four-feature model also has the advantage of shedding light on the various properties of a potential sub-article relationship that make it more or less likely to be a true relationship. Table 3.5-c shows the odds ratios for each feature. By far the most predictive feature in the model is *MaxTokenOverlap*, which is not a surprise given the above results. Table 3.5-c indicates that a relationship with a maximum token overlap of 1.0 is over ten times as likely to be a true sub-article relationship than one with an overlap of 0.0. In the one-feature model, the ratio increases to over 14. This result is somewhat intuitive. Many of the positive examples discussed at the beginning of this section have an English token overlap of 1.0 (e.g. “United States” (English) and “History of the United States” (English)) and many of the negative examples have an English overlap of 0.0 (e.g. “Caffeine” (English) and “History of Chocolate” (English)).

Note, though, that this is not true in every one of the above examples. “United States” (English) → “American literature” (English) is a positive example, but has a token overlap of 0.0, at least in the English Wikipedia. The articles on American literature in several other language editions, however, have a token overlap of 1.0 with the corresponding articles on the

United States. This increases the maximum token overlap, whose final averaged value is 0.55, high enough that both logistic regression models correctly understand the relationship to be a parent article / sub-article relationship. Here, we see the advantages of developing features and taking an approach that understands sub-article relationships as relationships that connect two concepts rather than relationships that merely connect two articles.

The next most-predictive features were the percentage of languages in which a potential sub-article relationship exists and the percent of those relationships that occur in the “See also” list at the bottom of articles. A greater percentage of potential sub-article relationships makes a relationship between concepts more likely to be a true sub-article relationship. On the other hand, relationships that tend to occur in the “See also” list are less likely to be true sub-article relationships, as hypothesized. The odds ratios in these two cases, however, mean there will be more counter-examples than is the case with maximum title overlap. Consider the “See also” list from the “Africa” (English) page depicted in Figure 3.5-c. Table 3.5-d shows the results of our model applied to these potential relationships. Note that the majority of the relationships – buoyed by *MaxTokenOverlap* – have been predicted to be true sub-article relationships.

A higher-level understanding of the performance of our sub-article model can be obtained by examining its entire set of results for the concept of caffeine. The classifier correctly rejected

Potential Sub-article Relationship	Is Sub-article Relationship?
“Africa” → “Outline of Africa”	TRUE
“Africa” → “Index of Africa-related articles”	TRUE
“Africa” → “Afro-Eurasia”	FALSE
“Africa” → “Highest mountain peaks of Africa”	TRUE
“Africa” → “List of African millionaires”	FALSE
“Africa” → “List of cities in Africa”	TRUE
“Africa” → “Urbanization in Africa”	FALSE

Table 3.5-d: Model predictions for the “See also” potential sub-article relationships in Figure 3.5-c.

“History of chocolate,” “History of yerba mate,” and the other negative examples in Figure 3.5-b. “Health effects of caffeine” (English), on the other hand, was correctly identified as a sub-article, as was “Caffeine addiction” (English).

It is important to point out that we were able to develop significantly more accurate models for tasks other than predicting mean scores greater than or equal to 2.5. For instance, predicting scores greater than or equal to 2.0 results in approximately the same accuracy while the baseline drops to about 50%. However, for the purpose of the studies and systems below, in the end we felt a more conservative definition of sub-articles was appropriate. As such, we trained our model on the threshold of 2.5.

There are a couple of limitations to discuss with regard to our sub-article work. First and foremost, we only are able to detect potential sub-article relationships that are *explicitly* encoded by Wikipedia editors. We believe this approach is appropriate as it respects the decisions of editors, something that any Wikipedia-based system or study does implicitly when it accesses its first bit of content from the encyclopedia. That said, it would be interesting to try to discover *implicit* sub-articles in an automated fashion. This would be an excellent follow-on experiment to that in this section, as our training data and/or the output of our classifier could be leveraged as training data for this new task. A system that can execute implicit sub-article discovery successfully may be quite useful to Wikipedia editors (c.f. Weld et al. [204]) in addition to system-builders and researchers who work with Wikipedia.

A second limitation is that our model is trained on only a small set of data from just two language editions. There are likely nuances in sub-article usage that cannot be learned from this limited information. We are addressing this issue by labeling a large number of additional potential sub-article relationships. Not only will this process add 500 new samples to our ground

truth, but these 500 potential relationships come from all 25 language editions, not just two. To rate each relationship in this new dataset, our two coders used an updated version of our GUI tool that includes translations from Bing Translate. Any improved sub-article classifier that results from this new data will be implemented in WikAPIdia’s API and included in its release.

Regardless of these limitations, however, this section has demonstrated that we were able to build an accurate sub-article classifier that can be deployed in sub-concept-level diversity studies and related systems (e.g. Omnipedia in Chapter 7). In all of the four experiments below (Sections 3.5.2 - 3.5.5), we use this classifier to merge the BOLs of sub-articles with that of their parent articles prior to the execution of the experiment. As such, none of the experiments in this section mistake content about a concept that is moved to a sub-article as content that does not exist in a given language edition, providing for a much more accurate analysis of the similarities and differences of the descriptions of concepts across language editions.

### **3.5.2 Study 1: Pairwise Sub-concept-level diversity**

In our first study of the similarities and differences of articles about the same concept in different language editions, we use our BOL-based approach – enhanced by our missing links and sub-article contributions – to directly examine the question we posed in the introduction. That is, we calculate the extent to which any two articles about the same concept – e.g. “Schokolade” (German) and “Chocolate” (Spanish) – describe that concept in the same way. We accomplish this by randomly selecting same-concept article pairs and comparing the bags-of-links of these articles. Specifically, this comparison is done with a metric we call *RatioInRandom*, which is defined as follows:

$$\text{RatioInRandom}(c, l_1, l_2) = \frac{|BOL_{l_1} \cap BOL_{l_2}|}{|BOL_{l_2}|}$$

where  $c$  is a concept and  $l_1$  and  $l_2$  are chosen randomly from the set of language editions in which  $c$  has an article.

We calculated the *RatioInRandom* metric for an article pair from 20,000 concepts that exist in more than one language edition. We required that each concept have at least five links in both language editions considered in order to avoid “stub” articles. With regard to the parseability of the links in the BOLs, we took two approaches. First, we provided a best-case scenario for the global consensus hypothesis (on top of the upper-bound parameters’ best-case scenario) by using all links – parseable and unparseable – in the BOL for  $l_1$  but only the parseable links for  $l_2$ . This accounted for situations we occasionally saw in testing in which one language edition’s parseable link was another language editions unparseable link. The second approach we took was to merely compare the two BOLs based on all links (parseable and unparseable).

The results of our analysis can be found in Table 3.5-e. The table reveals that, on average, any given non-stub article in multilingual Wikipedia is missing *at least* 28% of the content of another article about the same concept in a different language edition. In other words, articles like “Schokolade” (German) omit, on average, at least 28% of the information from articles like “Chocolate” (Spanish). Table 3.5-e also shows that even when only restricting each bag-of-links to concepts that exist in both language editions (the only-intersection upper-bound), the amount of overlap between the article pairs does not increase significantly. Additionally, given that these upper-bound values are quite strict, there is likely a great deal more missing content. For instance, more moderate BOL parameters raise the amount of missing content to 36%.

The results for the same analysis in which the BOLs were both based on all links can be found in Appendix D. When considering parseable and unparseable links, the articles about the same concept in multilingual Wikipedia are even more diverse. A given article is missing *at least* 33% on average of the parseable and unparseable links of another article about the same concept in a different language edition (26% intersection-only).

Looking at our results on an article pair-by-article pair basis, we found a large number of cases in which cultural contextualization is likely the cause of the similarities and differences between the two articles. This was most obvious when one language edition went into a great deal more depth about a concept that is significantly more relevant to its corresponding language-defined culture than that of the other language edition. For instance, one of our 20,000 randomly-sampled pairs of articles was the Chinese and Japanese articles about the All Japan High School Soccer Team. The Japanese article's BOL was not surprisingly a complete superset of the Chinese article's BOL. Conversely, when the experiment sampled the concept known in the English Wikipedia as “United Kingdom general election, October 1974” (English) and set  $l_2 = \text{English}$  and  $l_1 = \text{Russian}$ , the *RatioInRandom* statistic was very low across the board regardless

<b><i>RatioInRandom</i> using Parseable Links</b>				
<b><i>Wikification Strategy</i></b>	<b><i>Mean</i></b>	<b><i>% RatioIn Random = 1</i></b>	<b><i>Mean only-intersection</i></b>	<b><i>% RatioIn Random = 1 only-intersection</i></b>
“Kitchen Sink” Upper-Bound <i>&lt;WikipediaTitle+Redirect+AnchorText, GoogleTranslateTitle+Redirect&gt;</i>	0.720	16.0%	0.776	25.8%
Moderate <i>&lt;WikipediaTitle+Redirect, GoogleTranslateNone&gt;</i>	0.637	9.43%	0.687	13.9%
“Just Links” Lower-Bound <i>&lt;WikipediaTitleNone, GoogleTranslateNone&gt;</i>	0.459	2.43%	0.521	4.68%

Table 3.5-e: *RatioInRandom* summary statistics using bags-of-links just based on parseable links.

of BOL parameters. Indeed, the article “Парламентские выборы в Великобритании (октябрь\_1974)” contains at most 1.5 percent of the content in the article “United Kingdom general election, October 1974” (English). We saw similar results when  $l_2$  was English and the concepts in consideration were Trinity College (in Toronto) and Sydney Law School.

That said, the English Wikipedia frequently played the opposite role as well. The Chinese article about the Chinese swimmer Zhang Enjian covers 100% of the content of the English article according to the upper-bound BOL parameters. The same was true for the Chinese/English article pairs about Wu Shengli (a Chinese admiral), Yamaga Station (a Japanese train station), Zengwun River (a river in China), and several others. In the case of the Zengwun River, the Chinese article is *substantially* longer than the English article. In fact, there were numerous concepts in which the English article was far shorter than the article to which it was compared: Dmitriy Svyatash (a Ukrainian politician) was far more detailed in Russian than in English, Annecy Cathedral (a Cathedral in France) was much longer in French than in English, Psalm 80 had several times more text in French than in English, and the same was true for Hokkai Gakuen University with respect to Japanese and English.

In general, for 14.3% of samples in which English was  $l_2$ , the upper-bound *RatioInRandom* statistic was 0.9 or greater, meaning that articles from other language editions not infrequently covered all or nearly all of the English articles’ content on the same concepts. When only considering linked concepts in the intersection of the two language editions, this number jumps to a full 27.0%. That said, when English was  $l_1$ , *RatioInRandom* was greater than or equal to 0.9 over 38% of the time (45.9% intersection only).

### 3.5.3 Study 2: Language-by-Language Sub-concept-level Diversity

In the previous study, we sampled randomly from across all of multilingual Wikipedia and found that on average, any two same-concept article pairs are quite different from one another. Here, we examine investigate the same issue, but on a language-by-language basis. Specifically, we investigate whether the average amount of content overlap between two articles increases or decreases with certain language pairs. The global consensus hypothesis here would predict that there are a large number of language pairs in which the articles of one language edition are nearly always subsets of another language editions' articles. This would mean, in other words, that there are a large number of language editions that have little to no unique content about concepts relative to another language edition. However, if these superset/subset language edition pairs are rare or do not exist, this would be in accordance with the global diversity hypothesis and would demonstrate that each language edition contributes a non-trivial amount of unique content about concepts that are covered in other language editions as well.

For our analysis here we only slightly adapt the *RatioInRandom* metric into our *RatioOfLang1InLang2* metric. The only difference between the two is that rather than choosing  $l_1$  and  $l_2$  randomly, in this study we hold  $l_1$  and  $l_2$  fixed and iterate through all  $l_1$  and  $l_2$  pairs. As *RatioOfLang1InLang2* is not symmetric, *RatioOfLang1InLang2* for all 600 pairwise permutations of  $l_1$  and  $l_2$  where  $l_1 \neq l_2$  had to be calculated. For each of these 600 language pairs, we calculated the average *RatioOfLang1InLang2* for 500 concepts that had articles in both language editions. As above, both articles had to have five or more links.

The results of these calculations for the upper-bound wikification strategy can be found in Table 3.5-f. The maximum average content overlap of any two language pairs was 0.93. This maximum value represents the upper-bound on the average percent of content from an

Indonesian article that is contained within a same-concept English article. This means that the oldest and largest language edition in Wikipedia is still missing on average *at least* seven percent of the content in the much smaller and less-developed Indonesian Wikipedia for all concepts that have articles in both language editions. Examining the 500 sampled concepts for English and Indonesian, it is likely that a sizable number of the concepts with the smallest *RatioOfLang1InLang2* values are the result of the cultural contextualization of encyclopedic knowledge. For instance, the article “Confederation of Indonesia Prosperous Trade Union” (English) only covers 25.0% of the content in its Indonesian equivalent “Serikat Buruh Sejahtera Indonesia” (Indonesian). The *RatioOfLang1InLang2* is exactly the same for August Melasz, a well-known Indonesian actor. Other concepts with very low *RatioOfLang1InLang2*'s in this case include Pesantren (a type of Islamic boarding school in Indonesia), the Tangerang–Merak Toll Road (an Indonesian highway), and a number of concepts related to Indonesian soccer.

Covered Language Edition

Covering Language Edition

	Cata	Chin	Czec	Dani	Dutc	Eng	Fin	Fren	Ger	Heb	Hun	Indo	Ital	Japa	Kor	Nor	Pol	Port	Rom	Rus	Slvk	Spa	Swe	Turk	Ukr	Min	Avg	Max
<b>Catalan</b>		.580	.621	.486	.615	.846	.519	.754	.716	.558	.670	.561	.730	.660	.519	.561	.652	.611	.533	.711	.410	.733	.551	.579	.671	.410	.619	.846
<b>Chinese</b>	.689		.609	.526	.654	.832	.523	.740	.708	.575	.685	.566	.711	.730	.543	.565	.658	.665	.504	.701	.556	.734	.592	.585	.666	.504	.638	.832
<b>Czech</b>	.668	.627		.527	.634	.830	.535	.747	.739	.580	.683	.581	.713	.670	.512	.567	.675	.656	.525	.731	.588	.732	.608	.582	.611	.512	.638	.830
<b>Danish</b>	.732	.702	.674		.700	.885	.589	.788	.778	.665	.734	.631	.778	.734	.599	.741	.727	.751	.588	.783	.560	.807	.681	.654	.663	.560	.706	.885
<b>Dutch</b>	.754	.584	.608	.515		.843	.514	.774	.730	.522	.642	.548	.723	.671	.493	.593	.668	.696	.509	.699	.524	.740	.612	.607	.627	.493	.633	.843
<b>English</b>	.604	.501	.470	.362	.532		.366	.626	.587	.354	.551	.455	.594	.534	.405	.456	.545	.549	.400	.562	.555	.637	.500	.463	.552	.354	.507	.637
<b>Finnish</b>	.723	.641	.672	.603	.686	.887		.787	.791	.651	.705	.629	.805	.719	.568	.637	.743	.710	.533	.792	.522	.797	.667	.631	.634	.522	.689	.887
<b>French</b>	.584	.501	.527	.421	.589	.788	.431		.667	.412	.570	.459	.680	.594	.431	.512	.592	.580	.399	.645	.369	.681	.489	.488	.518	.369	.539	.788
<b>German</b>	.549	.426	.490	.417	.561	.761	.412	.641		.419	.527	.444	.613	.566	.407	.503	.536	.496	.413	.579	.353	.625	.457	.463	.475	.353	.506	.761
<b>Hebrew</b>	.652	.597	.580	.496	.602	.855	.521	.723	.685		.633	.532	.714	.656	.504	.548	.604	.647	.515	.711	.421	.729	.540	.570	.575	.421	.609	.855
<b>Hungarian</b>	.648	.609	.601	.494	.621	.803	.540	.738	.700	.560		.532	.733	.664	.511	.575	.657	.660	.506	.705	.739	.591	.563	.609	.454	.617	.803	
<b>Indonesian</b>	.750	.713	.691	.607	.721	.930	.597	.821	.763	.660	.739		.799	.744	.622	.676	.718	.751	.568	.785	.579	.816	.673	.688	.702	.568	.713	.930
<b>Italian</b>	.616	.528	.569	.445	.586	.820	.464	.705	.664	.460	.611	.485		.573	.435	.524	.623	.623	.419	.679	.505	.682	.506	.515	.567	.419	.567	.820
<b>Japanese</b>	.591	.535	.507	.423	.578	.755	.457	.648	.654	.454	.575	.464	.659		.479	.491	.566	.552	.434	.606	.350	.645	.480	.481	.541	.350	.539	.755
<b>Korean</b>	.698	.721	.670	.554	.700	.851	.561	.745	.743	.638	.714	.616	.754	.799		.633	.693	.718	.557	.767	.517	.777	.620	.657	.644	.517	.681	.851
<b>Norwegian</b>	.719	.663	.658	.678	.707	.855	.578	.773	.793	.603	.686	.624	.769	.727	.574		.688	.702	.544	.747	.504	.779	.685	.614	.654	.504	.680	.855
<b>Polish</b>	.671	.561	.608	.520	.647	.827	.509	.744	.735	.553	.686	.561	.724	.649	.496	.595		.678	.497	.743	.519	.726	.607	.613	.649	.496	.630	.827
<b>Portuguese</b>	.775	.648	.605	.554	.760	.873	.522	.790	.764	.556	.695	.619	.778	.707	.531	.605	.676		.521	.750	.609	.813	.668	.603	.739	.521	.673	.873
<b>Romanian</b>	.813	.684	.669	.614	.715	.923	.562	.878	.818	.633	.738	.653	.819	.737	.592	.643	.736	.740		.817	.742	.810	.813	.637	.804	.562	.733	.923
<b>Russian</b>	.626	.538	.536	.440	.607	.825	.463	.706	.692	.476	.587	.498	.678	.590	.462	.504	.614	.643	.471		.441	.706	.507	.510	.687	.440	.575	.825
<b>Slovak</b>	.838	.747	.819	.605	.761	.895	.563	.874	.788	.623	.724	.628	.762	.742	.585	.656	.728	.796	.562	.784		.803	.821	.602	.827	.562	.730	.895
<b>Spanish</b>	.722	.548	.551	.447	.615	.852	.480	.756	.684	.488	.608	.540	.718	.629	.445	.518	.619	.630	.488	.694	.502		.581	.547	.659	.445	.597	.852
<b>Swedish</b>	.723	.633	.654	.571	.707	.871	.543	.807	.769	.583	.672	.607	.740	.683	.530	.668	.699	.679	.513	.753	.683	.777		.596	.760	.513	.676	.871
<b>Turkish</b>	.725	.664	.681	.581	.731	.861	.573	.792	.764	.639	.715	.650	.798	.733	.599	.630	.717	.733	.574	.811	.582	.806	.663		.698	.573	.697	.861
<b>Ukrainian</b>	.716	.627	.619	.527	.621	.841	.526	.795	.726	.553	.684	.572	.720	.679	.535	.568	.652	.634	.551	.803	.502	.718	.580	.601		.502	.640	.841
<b>Minimum</b>	.549	.426	.470	.362	.1532	.755	.366	.626	.587	.354	.527	.444	.594	.534	.405	.456	.536	.496	.399	.562	.350	.625	.457	.463	.475	.350	.494	.755
<b>Average</b>	.691	.607	.612	.517	.1652	.846	.515	.756	.728	.551	.660	.561	.730	.675	.516	.582	.658	.663	.505	.723	.514	.742	.604	.577	.647	.477	.633	.846
<b>Maximum</b>	.838	.747	.819	.678	.761	.930	.597	.878	.818	.665	.739	.653	.819	.799	.622	.741	.743	.796	.588	.817	.742	.816	.821	.688	.827	.588	.758	.930

Example: English articles cover at most  
75.5% of the content in same-concept  
Japanese articles on average.

Example: Swedish articles cover at most  
57.4% of the content in same-concept  
Turkish articles on average.

Table 3.5-f: A table analogous to Table 3.4-c, but describing sub-concept-level diversity rather than concept-level diversity. Each cell represents the average amount of content covered by the the column language edition in the row language edition's articles. In other words, each cell holds the average RatioOfLang1InLang2 metric, with  $l_1$  = row language edition and  $l_2$  = column language edition.

The article pair “Sunday School” (English) / “Sekolah Minggu” (Indonesian) is a particularly interesting case. “Sunday School” (English) has over nine times the outlinks of “Sekolah Minggu” (Indonesian), yet it only contains 63.3% of the Indonesian article’s links/content. Among the content missing from the English article but included in the Indonesian article is a detailed discussion of Sunday schools in Indonesia. This discussion contains links to Indonesia, the continent of Asia, and an organization of Christian churches in Indonesia that does not even have an article in the English language edition.

While many of the instances of low English coverage relative to Indonesian can be reasonably attributed to cultural contextualization of the encyclopedia, doing so with others is more difficult, at least without a more in-depth knowledge of Indonesian culture. For instance, despite the fact that the English Wikipedia has an extensive article on the song “White Christmas” that is eight times as long as the Indonesian article, the Indonesian Wikipedia lists the number of covers of the song that are missed by the English Wikipedia. For instance, just reading the English Wikipedia’s article, one would not know that Hanson has recorded a version of “White Christmas,” something that the Indonesian Wikipedia points out and something that we were able to independently verify.

Returning to Table 3.5-f, we see that in general, English articles cover the most information in same-concept articles, but that the English language edition is never a total superset of another language edition. In only two cases (Indonesian and Romanian) does the  $RatioOfLang1InLang2$  with English =  $l_2$  break the 0.9 threshold. In fact, relative to some of the larger language editions, the English Wikipedia is missing more than 20% of the content about shared concepts on average. English articles cover the least amount of Japanese same-concept articles, on average missing at least 24.5% of their content.

Examining the 500 article-pair samples between English and the larger language editions also reveals a great deal of cultural contextualization. The fourth-smallest  $RatioOfLang1InLang2$  with  $l_1 = \text{Japanese}$  and  $l_2 = \text{English}$  is the article about the 2005 Japanese general election, where “Japanese general election, 2005” (English) only covers at most 16.1% of the content in “第44回衆議院議員総選挙” (Japanese). Similarly, the English Wikipedia article on the Freising Cathedral only covers at most 24.3% of the much-longer “Freisinger Dom” (German).

Outside of English, Table 3.5-f reveals that some language editions are more “covered” than others. Romanian, Slovak, Indonesian, and Danish are the language editions with the least amount of unique content using our language-stratified sample. These also happen to be four of the smallest language editions with respect to numbers of articles, which, along with English being the “least covered,” reveals a relationship between number of articles and unique content in articles. There are some exceptions, however. Hebrew, the smallest language edition, is above the mean when it comes to its articles having unique content. We see a similar phenomenon with Hungarian, and an opposite phenomenon with Portuguese.

Table 3.5-f also displays some of the same language-by-language patterns we saw with concept-level diversity. For instance, despite being the second- and third-largest language editions, French and German only cover about 64-66% of each other’s content in same-concept article pairs. With regard to similar cultures having more overlap, note that the Japanese Wikipedia covers the content in Korean articles more than any other language edition. The same is almost true with respect to the Chinese Wikipedia and the Korean Wikipedia, with the Slovak Wikipedia being covered slightly more by Chinese, although it is heavily covered by most language editions. Finally, the similarity between the Scandinavian language editions persists here, with the Norwegian Wikipedia significantly more of Danish than any other language

edition.

### 3.5.4 Study 3: Percentage of Information in English

Recall that the English-as-Superset hypothesis at the sub-concept level predicts that the English Wikipedia will have all or nearly all information about a concept, with the other language editions only containing selected subsets of that information. We already saw above that this hypothesis is flawed, but the preceding study was done on a language-by-language basis. A more direct investigation of the English-as-Superset hypothesis must compare English articles to all of their other-language counterparts as a group, not just one at a time. As such, our goal in this study is to determine the average percent of multilingual Wikipedia's information about a concept that is contained within the English article about that concept (assuming there is one).

We began our analysis by randomly sampling 2,300 concepts that had articles in the English Wikipedia and at least one additional language edition. Here, we again limited our sample to concepts that had at least five outlinks in all articles (to exclude stubs), only considered parseable links, and always included sub-articles in the bags-of-links. Using these experimental parameters, we calculated the *RatioInEnglish* metric for each concept, which is defined as follows:

$$\text{RatioInEnglish}(c) = \frac{|BOL_{EN}|}{|BOL_{ALL}|}$$

where  $c$  is a concept that has an English article and an article in at least one other language edition,  $BOL_{EN}$  is the English BOL for  $c$ , and  $BOL_{ALL}$  is the union of all BOLs for  $c$ .

*RatioInEnglish* captures the extent to which an English Wikipedia article covers all of the content in multilingual Wikipedia about a concept. Another interpretation is that  $1 - \text{RatioInEnglish}$

represents the amount of information one is missing about a concept by only reading the English language edition's article about that concept.

The distribution of the *RatioInEnglish* metric over our 2,300-concept sample is found in Figure 3.5-d and summary statistics are in Table 3.5-g. Our results indicate that, on average, an English Wikipedia article about a concept that is covered in at least one other language edition is missing *at least* 29.2% of the overall information in multilingual Wikipedia about that concept. In other words, any Wikipedia reader or algorithm that uses information exclusively from the English Wikipedia is only getting *at most* 70.8% of the information available in multilingual Wikipedia<sup>26</sup> about the subject of English articles that have same-concept equivalents in other languages. In fact, in only at most 2.4% of cases does the English article have all of the information about a concept in multilingual Wikipedia.

---

26 Of course, if we were to examine more than 25 language editions, this number could only go down.

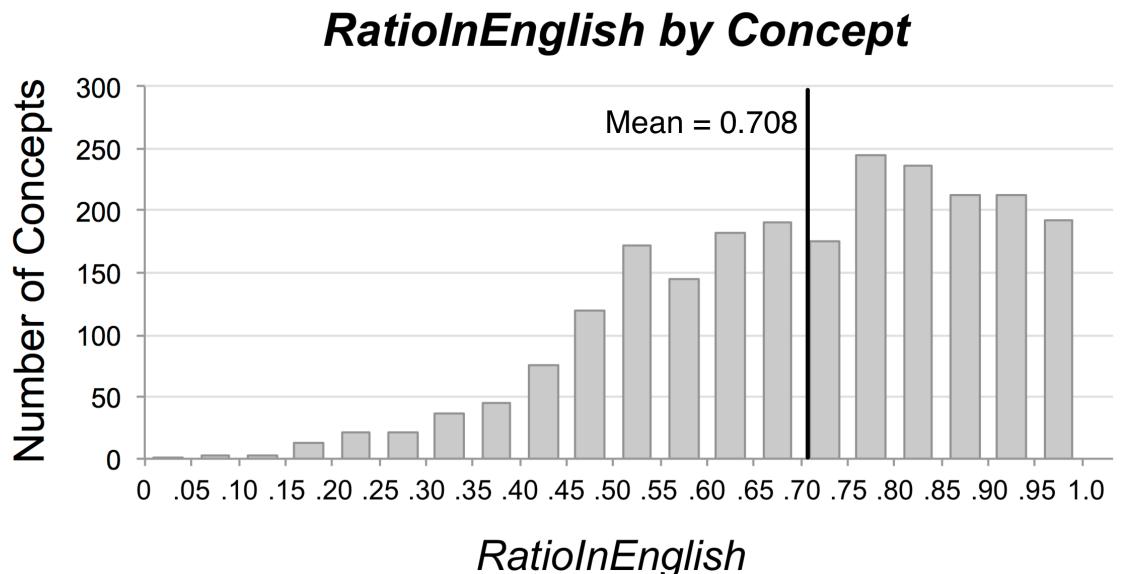


Figure 3.5-d: The distribution of RatioInEnglish values for our 2,300-concept sample. The average value is 0.707, meaning that an English article about concept c is missing 29.3% of the information about that concept that is available in the rest of multilingual Wikipedia. The median value is 0.733, which indicates that half of English articles are missing more 26.7% of multilingual Wikipedia's information.

<b>RatioInEnglish using Parseable Links</b>				
<b>Wikification Strategy</b>	<b>Mean</b>	<b>% RatioIn English = 1</b>	<b>Mean only-intersection</b>	<b>% RatioIn English = 1 only-intersection</b>
“Kitchen Sink” Upper-Bound <WikipediaTitle+Redirect+AnchorText, GoogleTranslateTitle+Redirect>	0.708	2.4%	0.754	4.6%
Moderate <WikipediaTitle+Redirect, GoogleTranslateNone>	0.584	1.1%	0.621	2.0%
“Just Links” Lower-Bound <WikipediaTitleNone, GoogleTranslateNone>	0.384	0.3%	0.426	0.4%

Table 3.5-g: Summary RatioInEnglish statistics for a variety of wikification strategies.

Looking at our 2,300-concept sample in more detail, cultural contextualization appears to be a major factor behind our results. The English Wikipedia is missing more than 75 percent of the information about concepts such as the 1994 Dutch general election, Ukrainian Premier Reserve League (a Ukrainian soccer league), Southern Basque Country (an area in northern Spain), Seweryn Krajewski (a Polish singer/songwriter), and European route E-95 (a North-South highway that runs through Eastern Europe) among others. On the other side of the *RatioInEnglish* distribution, concepts in which more than 95% of information is in English include Ricky Gervais (a well-known British comedian), the Fox News Channel, the British National Party (a far-right British political party), and the United States 29<sup>th</sup> Infantry Division.

That said, there are also more complicated cases. For instance, the English Wikipedia only contains at most 22.9% of the information about the concept of sexual minorities. This is in part because the articles on the concept in languages like Finnish discuss local issues related to sexual minorities. However, the majority of the information not in the English Wikipedia comes from the Japanese Wikipedia, which simply goes into a great deal more depth about sexual minorities than the English one does. The reverse is also true in a few situations, e.g. “Rostock Switzerland” (English) covers all the information in multilingual Wikipedia about a geologic feature in Germany.

<b>Sampled Global Concepts with Highest <i>RatioInEnglish</i></b>		
<b>English Title</b>	<b>RatioInEnglish</b>	<b>only-intersection</b>
New Jersey	0.973	0.981
Illinois	0.968	0.974
River Thames	0.968	0.977
Pittsburgh	0.960	0.966
Roy Keane (Irish soccer star/coach)	0.959	0.969
Royal Air Force	0.959	0.973
British Columbia	0.956	0.968
Birmingham (England)	0.954	0.960
Bob Hope	0.952	0.965
Andrew Johnson	0.952	0.965
<b>Sampled Global Concepts with Lowest <i>RatioInEnglish</i></b>		
<b>English Title</b>	<b>RatioInEnglish</b>	<b>only-intersection</b>
Rodent	0.187	0.187
Smolensk Oblast (Russian province)	0.223	0.476
Shogun	0.334	0.362
Osteichthyes (type of fish)	0.344	0.381
Malay Peninsula	0.354	0.391
Paleogene (geologic era)	0.367	0.390
Mecklenburg-Vorpommern (German province)	0.377	0.426
Perm Krai (Russian province)	0.421	0.605
Chernihiv (Ukrainian city)	0.431	0.577
Automobile	0.433	0.613

Table 3.5-h: The largest and smallest *RatioInEnglish* values for a sample of 500 global concepts.

In order to gain an understanding of *RatioInEnglish* among more globally-known concepts, we calculated the *RatioInEnglish* for 500 concepts that had articles in all 25 language editions<sup>27</sup>. Table 3.5-h shows the global concepts with the highest and lowest *RatioInEnglish* values. Places in the English-speaking world were the concepts that had the highest percentage of information in the English Wikipedia, e.g. (at most) 97.2% of the information in multilingual Wikipedia about the state of New Jersey is in the English Wikipedia. Similarly, concepts that had the lowest

<sup>27</sup> We set the minimum number of outlinks for all articles here to three in order to not excessively restrict our already-small global concepts dataset.

*RatioInEnglish* values were frequently – though not exclusively – concepts that were best known in cultural contexts outside of those of English speakers. For instance, the upper-bound *RatioInEnglish* statistic for Shogun is only 0.334. Other concepts in the bottom ten (according to the upper-bound statistic) include Malay Peninsula and Mecklenburg-Vorpommern (a German bundesland/first-order administrative district north of Berlin).

That said, some of the lowest upper-bound *RatioInEnglish* results could not be explained by cultural contextualization, at least not in a clear fashion. For instance, several *math and science* concepts – e.g. paleogene (a geologic era), Osteichthyes (a type of fish), Googol (a large number) – are all near the bottom as well. Indeed, comparing the “Paleogene” (English) article with “Paleógeno” (Spanish), for instance, it is clear that the Spanish article is substantially more detailed<sup>28</sup>.

In order to establish the robustness of our *RatioInEnglish* findings, we executed similar analyses on smaller concept samples using a wider set of parameters. One concern we had was that by leaving out stub articles, we may be giving an “unfair advantage” to the non-English language editions. However, after calculating *RatioInEnglish* for 500 concepts without a minimum outlinks restriction, we found that the upper-bound average was 59.7%, quite a bit lower than our original 70.8%. It seems that by removing stub articles, we were actually “benefiting” the English Wikipedia, which has a larger number of stubs than we had anticipated.

The second study we ran for robustness purposes looked at all links, not just parseable ones. The minimum number of (parseable) outlinks for all articles about a concept was raised to 10 in this case in order to avoid issues related to pages with small numbers of parseable links (likely

---

28 This finding advocates for the use of Omnipedia (Chapter 7) – which shows the similarities and differences between the language editions of Wikipedia – not as a window into the diversity between the language editions, but as a way to be able to access all of the content in multilingual Wikipedia.

automatically-generated) and enormous numbers of unparseable links. In this case we saw a mean *RatioInEnglish* of 0.704 (0.769 only-intersection), nearly identical to that of our original results using parseable links.

### **3.5.5 Study 4: Diversity When Controlling for Length**

In the examples above, we have seen two types of sub-concept-level diversity. The most common type is the diversity inherent in article depth. We saw, for instance, that a concept closely associated with a certain language-defined culture will often have a much more detailed article in the corresponding language edition of Wikipedia than in the others. However, we occasionally also encountered another form of sub-concept-level diversity: shorter articles having content that is unique relative to same-concept longer articles. That is, sub-concept-level diversity that occurs even when controlling for diversity (and cultural contextualization) in article length.

When examining the raw data from the experiments above, we found that this length-independent type of sub-concept-level diversity was not at all uncommon. For instance, recall that the Indonesian article about Sunday schools had a great deal of content not in the English article, even though the English article had nine times more links. Other examples include the Hebrew article on the concept of breakfast being many times shorter than its English counterpart, but having about 15% of its links be unique relative to English. The targets of these links are almost exclusively related to breakfast as it is understood by Hebrew-speakers, e.g. the concepts of Israeli salad, the Israeli Ministry of Health, and the Talmud. Similarly, the Spanish article “Televisión de alta definición” has many fewer links than the English article “High-definition television,” but the English article focuses on HDTV in the United States and Europe while the

Spanish article has a much more detailed content about the state of HDTV in Latin America and Spain, specifically.

In our fourth and final study, we sought to understand the extent of this second, length-independent type of sub-concept-level diversity. To do so, we use a metric known as the overlap coefficient (OC) [141]. In the context of our multilingual Wikipedia work, it is defined as follows:

$$OC(c, l_1, l_2) = \frac{|BOL_{l_1} \cap BOL_{l_2}|}{\min(|BOL_{l_1}|, |BOL_{l_2}|)}$$

where  $OC$  is the overlap coefficient of concept  $c$  and  $l_1$  and  $l_2$  are articles about  $c$  in languages  $l_1$  and  $l_2$ , respectively.  $OC$  is effectively the intersection of the two articles' BOLs divided by the size of the smaller of the BOLs. In other words,  $OC$  describes the ratio of the links of the shorter of the two Wikipedia articles about concept  $c$  also contained in the longer of the articles on  $c$ . In the context of the bag-of-links assumption, this effectively means the content in the shorter of the two articles that is not in the longer of the two articles.

For example, consider a hypothetical English Wikipedia article on a concept  $c$  that is also covered by a shorter article in the Chinese Wikipedia. If the English article has outlinks to 90% of the concepts to which the Chinese article links, we would say that the overlap coefficient for this pair of articles is 0.9. Similarly, if  $c$  is like one of those concepts discussed above for which the English Wikipedia has the shorter article and the Chinese Wikipedia the longer, the language edition of the numerator and the denominator would be swapped. If the Chinese article had 60% of English article's links in this case, the overlap coefficient would be 0.6.

We calculated the overlap coefficient for 2,000 same-concept article pairs using a variety of BOL parameters. The results of this analysis using parseable link-based BOLs can be found in Table 3.5-i. In multilingual Wikipedia, the shorter of two articles about the same concept has *at least* 11.0% unique content relative to the longer article on average (8.0% only-intersection). For only 30% (43.2% only-intersection) of article pairs was the longer article a superset of the shorter article. This means that in 70% of pairs, the longer article is missing at least some content that is available in the shorter article. Additionally, as above, these figures are the result of a very strict upper-bound and there is likely somewhat more diversity. For instance, using a more moderate strategy, the shorter article has 19% unique content (16% only-intersection). The results for all links (not just parseable ones) can be found in Appendix E and reflect a slightly smaller overlap coefficient overall.

### 3.5.6 Discussion

In this section, we have seen through four different analyses that there is a great deal of diversity in terms of how concepts are defined in multilingual Wikipedia. This sub-concept-level diversity is all the more remarkable when considering the fact that it exists *on top of* the concept-level diversity that we found in the previous section. That is, we demonstrated in Section 3.4 that

<b>Overlap Coefficient using Parseable Links</b>				
<b>Wikification Strategy</b>	<b>Mean OC</b>	<b>% OC = 1</b>	<b>Mean OC only-intersection</b>	<b>% OC = 1 only-intersection</b>
“Kitchen Sink” Upper-Bound <i>&lt;WikipediaTitle+Redirect+AnchorText, GoogleTranslateTitle+Redirect&gt;</i>	0.890	30.0%	0.919	43.2%
Moderate <i>&lt;WikipediaTitle+Redirect, GoogleTranslateNone&gt;</i>	0.807	16.0%	0.840	23.8%
“Just Links” Lower-Bound <i>&lt;WikipediaTitleNone, GoogleTranslateNone&gt;</i>	0.551	3.65%	0.619	7.15%

Table 3.5-i: Overlap coefficient averages when using bags-of-links just based on parseable links.

two language editions covering the same concept is a relatively rare event. In this section, we have seen that even when this event occurs, the language editions tend to cover the shared concept differently.

Before closing our discussion on sub-concept-level diversity, it is important to note that the compounding effect of concept-level and sub-concept-level diversity is equally important when considering the cultural contextualization present in multilingual Wikipedia. The fact that the English Wikipedia has an article about Indonesian actor August Melasz may seem like a piece of evidence in support of the English-as-Superset hypothesis. However, examining the situation at a sub-concept level, we see that indeed, cultural contextualization exists, with the Indonesian article going into substantially more detail than the English one.

### 3.6 Centrality Diversity

In the previous two sections, we examined concept- and sub-concept-level diversity across entire language editions. We now turn our attention to understanding these forms of diversity in a more nuanced, faceted fashion. That is, we begin to explore the types of concepts for which diversity is high and those for which diversity is low (relatively speaking). The first dimension of analysis we consider is that of Wikipedia Article Graph (WAG) centrality.

Centrality measures are a family of approaches in graph theory and network analysis that seeks to determine the *importance* of a particular vertex to a given graph<sup>29</sup>. A vertex with high centrality is said to be important in some fashion, while one that is less central is said to be less important. Each centrality measure defines importance in its own way. For instance, vertices with high betweenness centrality in a social network represent people who, by frequently being on the

---

<sup>29</sup> Edge centrality can also be calculated, but is not considered here.

shortest path between any two people, are integral to information flow in the network and have extensive “bridging social capital” [162].

Here we consider two different types of centrality: degree centrality and eigenvector centrality. With regard to the former, we make extensive use of *indegree centrality*, or the number of links for which a given Wikipedia article is the destination. In the Wikipedia context, the type of importance measured by indegree centrality is straightforward: an article with very high indegree centrality describes a concept that is written about frequently in the article’s language edition and an article with very low indegree centrality is more rarely discussed<sup>30</sup>. Articles with an indegree of zero are, taking the bag-of-links assumption as truth, about concepts never discussed elsewhere in the language edition.

We also consider PageRank centrality [17], which is a type of eigenvector centrality. PageRank is similar to indegree, except it assigns a weight to each link according to the centrality of its source. That is, a link from a highly central article like “United States” (English) is given more weight than a link from a peripheral article like “Mister Philippines 2008” (English), an article about the winner of a Filipino male beauty contest that has an indegree of zero.

In the previous section, we saw that the WAGs of each language edition can differ extensively from one another. After a brief discussion of methods, we begin this section by investigating the effects of this diversity on the resulting WAG centrality measures. Here we ask questions such as, “What are the most central concepts in each language edition?” and “Is this set of concepts roughly consistent in multilingual Wikipedia or does it vary widely from language

---

<sup>30</sup> This statement is based on the bag-of-links assumption, which breaks down in this context in certain cases. For instance, in the German and English Wikipedias, years are written about frequently, but are typically not linked to when they are mentioned.

edition to language edition?” We then look at concept-level diversity at varying degrees of centrality, examining whether concepts with high centrality in a language edition tend to have greater conceptual coverage than concepts with low centrality. Finally, we perform a similar analysis with sub-concept-level diversity.

### 3.6.1 Centrality Methods

The first step in determining the centrality diversity across the 25 language editions of multilingual Wikipedia was executing the algorithms that calculate centrality on all 25 WAGs. Given the size of multilingual Wikipedia – recall that our 25-language dataset has over one billion links – this was a non-trivial task.

We began by selecting the type of links we would consider. Thanks to WikAPIdia’s support for many link (edge) properties such as the location of the link in an article and the parseability of the link (Section 3.2.2), we could calculate centrality measures on various versions of each language edition’s WAG. In this chapter, we mainly only consider *parseable WAGs*, or the article graph that only consists of parseable links. Unparseable links tend to be of lower informational value than parseable links (see Section 3.2) and, due to their number, significantly increase the computational complexity of centrality measure algorithms such as PageRank. We do, however, briefly consider other types of WAGs later in the section, for instance showing the effect of including unparseable links in our analyses using smaller language editions as test cases.

Calculation of the indegree of each article (vertex) in each WAG is a simple process. The bulk of what we had to do here was iterate once through the parseable links in a given language edition, counting the number of times each article was the destination of a link. There were a few other Wikipedia-specific considerations, for instance adding the indegrees of redirect pages to

the pages that were the target of the redirects. However, for the most part, this was a very straightforward process.

Calculating the PageRank score of every article in all 25 language editions was a bit more involved. Although executing the PageRank algorithm on large graphs can be done relatively easily using distributed approaches, doing so on a single machine – even one with 64GB of memory like ours – requires more careful consideration and optimization. As is discussed in more detail in Section 3.12, we were able to implement the graph interface of the popular JUNG network analysis software library [145] directly in WikAPIdia, meaning that we could take advantage of WikAPIdia’s unique approach to resource management. Doing so allowed us to execute PageRank on an in-memory version of the entire multilingual Wikipedia article graph, which meant that JUNG’s implementation of PageRank became processor-bound rather than I/O-bound. With our machine’s two 2.4Ghz six-core Intel Xeon processors, we were able to run 100 iterations of PageRank on all 25 WAGs considered in this thesis in approximately one day. The resulting PageRank scores for each article are what is considered below.

### **3.6.2 Centrality Diversity in the WAGs of Each Language Edition**

After calculating the indegree and PageRank centrality of all articles in every language edition, we were able to compare the centrality assigned to articles about the same concept in different language editions. The global consensus hypothesis in this case suggests that the language editions will largely agree as to which concepts are the most central, or, in other words, that there will be a consensus as to the *most important concepts to encyclopedic world knowledge*. More specifically, the global consensus hypothesis predicts a situation in which a significant majority of the  $n$  most-central concepts in any two language editions are the same.

The global diversity hypothesis, on the other hand, suggests that the cultural contextualization that appears at the concept- and sub-concept level will cause the centrality/importance of each concept to vary in each language edition. The outcome of the global diversity hypothesis in this context is the  $n$  most central concepts in each language edition differing extensively.

From the first analysis we performed – comparing the 100 most-central concepts in each language edition – we found that there was much more support for the global diversity hypothesis than the global consensus hypothesis. On average, any two language editions shared only 54.9% of these concepts for indegree centrality and 51.7% for PageRank. In other words, if one were to ask, “What are the most important concepts in all of Wikipedia?”, the only correct answer is, “According to which language edition?”

We performed an identical analysis on the 1,000 most-central concepts and the 10,000 most-central concepts in each language edition. Table 3.6-a shows that our results were nearly the

<b>INDEGREE TOP-N SET OVERLAP</b>				
<b>Set</b>	<b>Mean</b>	<b>Stdev</b>	<b>Min</b>	<b>Max</b>
Top 100	0.549	0.109	0.300 ja/sv	0.820 da/fi, fi/no
Top 1,000	0.500	0.127	0.231 ja/sv	0.745 it/hu
Top 10,000	0.464	0.063	0.295 ja/sk	0.637 es/ca
<b>PAGERANK TOP-N SET OVERLAP</b>				
<b>Set</b>	<b>Mean</b>	<b>Stdev</b>	<b>Min</b>	<b>Max</b>
Top 100	0.517	0.096	0.270 ja/nl	0.79 fi/no, it/no
Top 1,000	0.540	0.069	0.348 ja/nl	0.72 hu/cs
Top 10,000	0.522	0.054	0.34 ja/sk	0.68 es/ca

Table 3.6-a: Pairwise agreement of the n-most central concepts between language editions, excluding the no-year-link language editions of German and English. In all cases, the n-most central concepts overlap by about 50 percent.

same as with the top 100: only around 50 percent of the  $n$  most-central concepts were shared between any two language editions on average. We did, however, see a noticeable decrease in the range as we increased the number of concepts under consideration. The minimum overlap between most-central concepts was 34% with 10,000 concepts as opposed to the 17% with 100, and the maximum decreased in a corresponding fashion.

Several additional patterns appear in Table 3.6-a and in the full dataset it describes. First, Table 3.6-a reveals that some of the same pairs of language editions responsible for the strongest similarities in other sections in this chapter are responsible for the strongest similarities here as well. The largest overlap between the lists of the 100 highest-indegree concepts of any two language editions was a tie between that of Norwegian and Finnish and that of Norwegian and Danish. Norwegian and Finnish are also tied with Norwegian and Italian (a result of less-obvious cultural causes) for the equivalent position in the 100 top PageRank score analysis. Table 3.6-a additionally shows that Spanish and Catalan have the most overlap in the top 10,000 PageRank score analysis. When Catalan was one of the two language editions being considered, the Spanish Wikipedia frequently provided the largest overlap, even if this overlap was not the largest globally. The same was true of Japanese with respect to the other two East Asian languages. Korean and Chinese were the only two language editions to share more than 50% of their top 10,000 PageRank concepts with Japanese, for instance.

It is important to note that for all the above results, the German and English Wikipedias were not included due to the editor communities in these language editions strongly discouraging links to articles about years<sup>31</sup>. Since years feature prominently in the top 100, 1,000, and 10,000 most central articles in all other language editions, we removed English and German to avoid

---

<sup>31</sup> In contrast to the findings of Zlatić et al. [221], we did not see the same issue with the Polish Wikipedia.

confounding community practice diversity with centrality/importance diversity. If we had included English and German, the means would all shift down slightly and the minimum overlap would involve one of these two language editions in every case.

Given that German and English both eschew links to years, we can, however, compare these two language editions with each other. Indeed, since these are the largest and oldest language editions, this comparison is quite valuable. The 100 most-central concepts in English and German only have 53 concepts in common as determined by indegree and 47 as determined by PageRank. This is approximately in line with the figures reported in Table 3.6-a. Looking at the PageRank results more closely, both language editions' top 100 include concepts like association football (soccer), many European countries, both world wars, and a number of other types of concepts mostly geographic in nature. The German XOR of these sets, on the other hand, includes German political parties, German cities, the Pope, ice hockey, Napoleon, while the English XOR includes the American Civil War, Ontario, South Africa, New York, Member of Parliament, and so on. The full intersection and XOR of these sets can be found in Appendix F.

Table 3.6-b, which shows the top 10 most-central concepts in nine language editions according to both indegree and PageRank, provides a finer-grained view of our results. A number of patterns can be seen in this table (and in the equivalent data for the other 16 language editions). First and foremost, with only one exception in all 25 language editions, one or more of the home countries of each language-defined culture appears in the top 10 for both indegree and PageRank centrality<sup>32</sup>. In fact, in the large majority of language editions, the home country is the most central or the second-most central concept. In some Wikipedias, even a prominent city in a

---

32 The one exception occurs in the Turkish Wikipedia's indegree list where Turkey is #14, but Turkey is the *most* central article in the Turkish Wikipedia when it comes to PageRank. This suggests that the Turkish Wikipedia's indegree rankings are affected by links from relatively insignificant articles.

home country makes it in the top ten, with some ranking as high as number three. For instance, Paris is the third most-central concept in the French Wikipedia for both indegree and PageRank, and the same is true for Copenhagen (Danish, indegree), Budapest (Hungarian, PageRank), Helsinki (Finnish, indegree), Oslo (Norwegian, indegree), and Prague (Czech, indegree). This home country/prominent city pattern is a primary driver of the centrality diversity between the language editions.

<b>Catalan</b>		<b>Czech</b>		<b>English</b>	
<b>PageRank</b>	<b>Indegree</b>	<b>PageRank</b>	<b>Indegree</b>	<b>PageRank</b>	<b>Indegree</b>
France	United States	United States	United States	United States	United States
United States	France	Czech Rep.	Czech Rep.	France	List svrn states
Municipality	Median	France	Prague	United Kingdom	Animal
Spain	Animal	Bohemia	France	Germany	English
Barcelona	2007	Germany	Czechoslovakia	Canada	France
2007	Species	Prague	Germany	England	Assoc. Football
Italy	Municipality	United Kingdom	Bohemia	World War II	United Kingdom
Catalonia	Chordate	Latin	2006	India	Germany
Sovereign State	Family (biology)	Europe	2007	Assoc. Football	Canada
Germany	2009	Pope	2009	List svrn states <sup>33</sup>	World War II
<b>German</b>		<b>Hebrew</b>		<b>Japanese</b>	
<b>PageRank</b>	<b>Indegree</b>	<b>PageRank</b>	<b>Indegree</b>	<b>PageRank</b>	<b>Indegree</b>
United States	United States	Israel	United States	Japan	Japan
Germany	Germany	United States	English lang.	United States	2007
France	France	English lang.	Israel	2006	2006
World War II	World War II	France	France	2007	United States
Latin	Berlin	Europe	2006	2005	2008
Austria	Italy	Jerusalem	2005	English lang.	2005
Switzerland	Austria	Germany	2007	Tokyo	2009
Berlin	Switzerland	United Kingdom	Germany	2008	2010
Italy	World War I	Hebrew lang.	World War II	2009	2004
English lang.	United Kingdom	World War II	2008	United Kingdom	Tokyo
<b>Portuguese</b>		<b>Russian</b>		<b>Spanish</b>	
<b>PageRank</b>	<b>Indegree</b>	<b>PageRank</b>	<b>Indegree</b>	<b>PageRank</b>	<b>Indegree</b>
Brazil	United States	Russia	Russia	United States	United States
United States	Brazil	United States	United States	Spain	Spain
Portugal	Square kilometer	Soviet Union	Soviet Union	France	Species
Square kilometer	Census	France	Ukraine	2008	2008
France	English lang.	Germany	Moscow	English lang.	Animal
Germany	Animal	Ukraine	Germany	Animal	Family (biology)
English lang.	2007	Moscow	2001	Argentina	Square kilometer
Animal	1999	Italy	France	Germany	2000
Pop. density	2004	2001	2007	Italy	2001
Census	Portugal	United Kingdom	2006	Species	France

Table 3.6-b: English-language titles of the 10 most-central concepts in nine language editions according to indegree and PageRank centrality

33 “List of Sovereign States”

An equally visible pattern present in Table 3.6-b is the prominence of the United States and to a lesser degree, France. The United States appears in the top 10 in all but two cases and it is the most central article in many language editions (in terms of both indegree and PageRank centrality), beating out the home countries of the corresponding language-defined cultures. In other words, the most significant agreement between the language editions in terms of concept centrality is that the United States is one of the or is *the* most important concepts in all of world knowledge. While the United States is of course a prominent player on the world stage, this result is likely due in part to mass translation of geographic articles from the English Wikipedia into many other language editions as was observed by Worten-Wang et al. [203], and has substantial implications for the ability of each language edition to customize content for its language-defined culture. This is an issue we discuss in detail in Section 3.11.1.

Geography in general features prominently in all top 10 lists and it is easily the most common domain of the concepts in Table 3.6-b. Geographers frequently work to communicate the importance and widespread nature of spatial information (e.g. [45]). In Section 3.4, we showed that geographic concepts are common in the global core of multilingual Wikipedia, backing up the geographers' argument. Table 3.6-b provides even more resounding evidence of the importance of geography. Not only do geographic concepts feature prominently in the global core of encyclopedic knowledge, they are globally determined to be the most central concepts within this core. In other words, they are the “core of the core.”

Like space, time also appears frequently in Table 3.6-b, as noted above. Years, particularly those that are recent (but not too recent) make up a large portion of several languages’ top 10s, especially that of Japanese. With regard to Japanese, not only do specific years have high centrality, but certain eras in Japanese history do as well. The Shōwa period and the Meiji period

are ranked numbers #21 and #25 respectively in terms of PageRank centrality. The Edo and Heisei periods are also in the top 100. More generally, it appears that around 2007 seems to be the “sweet spot” for centrality when it comes to years. This suggest that it takes five years or so for knowledge about a year to fully integrate itself in Wikipedia-based encyclopedic world knowledge.

Automatically created links also cause an important trend in Table 3.6-b. As noted in Section 3.2, while the parseable WAG contains a much lower percentage of non-manually-created links, it certainly contains some. As described by Lih [120], this most famously first occurred with geographic articles, but it also occurs in several other well-defined domains. These links tend to have an outsized effect on centrality measures as they cause a large number of articles to link to a small number of articles (e.g. many geographic articles in the English Wikipedia link to the U.S. Census in order to cite demographic statistics). While in most language editions these automated links only determine a relatively small portion of the  $n$  most-central articles as defined by PageRank and indegree, in several it is substantial. Most notably, the article with the highest PageRank score in the Dutch Wikipedia is “Kevers” (Dutch), which is about the same concept as “Beetle” (English). It is unlikely that this represents the most central concept to world knowledge as understood by Dutch speakers. Rather, this, like the United States’ centrality, is due to automated content production processes, a point to which we return later in this chapter (Section 3.11)

There are other commonalities between the lists in Table 3.6-b (e.g. some biological concepts), but it is also clear that a large number of concepts do not fit into any extensive pattern and, in doing so, create a great deal of diversity. In a few language editions World War II is in the top 10, in most others it is not. Some language editions have the Pope in the top 10, but this is

<b>PageRank Rank Correlations</b>			
<b>Set</b>	<b>Mean r</b>	<b>Min r</b>	<b>Max r</b>
Top 100	0.566	-0.074 (n.s.) Hebrew / Slovak	0.834 Korean / Norwegian
Top 1,000	0.663	0.255 Indonesian / Japanese	0.835 Hungarian / Czech
Top 10,000	0.657	0.461 Hebrew / Slovak	0.772 Czech / Polish

*Table 3.6-c: Summary statistics describing the correlation coefficients of the ranks of concepts in all top n PageRank set intersections. For instance, the average correlation between the ranks of concepts in the intersection of any two language editions' sets of top 10,000 PageRank centrality concepts was 0.657.*

not the case for the majority of them. Interestingly, the former Soviet states most relevant to the language-defined cultures of some of these language editions are in the top 10. This is not true of the language editions where this relationship is not applicable.

While comparing the sets of the top  $n$  most-central concepts in each language edition is one way to gain a descriptive understanding of the similarities and differences between the language editions in terms of concept centrality, performing correlations on the centrality ranks of each concept is another. We calculated these pairwise correlations on sets of the  $n$  highest PageRank concepts where  $n$  again was 100, 1,000, and 10,000. Only concepts that appeared in both sets were considered. The results of this analysis are in Table 3.6-c. We did not consider indegree in this case due to the large number of ties that occurred when  $n$  was greater than 100, especially for the smaller language editions.

The story in Table 3.6-c is similar to that in Table 3.6-a: although the language editions do have some agreement in the centrality rankings of concepts, this agreement is far from a widespread consensus. The mean pairwise rank correlation of any two language edition's top  $n$  sets was only 0.566 for  $n = 100$ , and it increased only by small amounts for  $n = 1,000$  and  $n = 10,000$ . As before, we excluded German and English from these statistics. Examining Appendix

F, it is clear that there is some variation in the PageRank score ranks of concepts that had one of the top 100 PageRank scores in these language editions. Austria is #6 in German, while it is #56 in English; California is #28 in English and #52 in German; and Canada has English's fifth-highest PageRank score but only Germans 22nd-highest. That said, even with these clear examples of the effect of cultural contextualization, English and German have a significantly higher correlation at  $n = 100$  than most other language editions with  $r = 0.65$ .

### **3.6.3 Centrality and Concept-level Diversity**

We now turn our attention to the effect of centrality on concept-level diversity. Examining concept-level diversity with a centrality lens allows us to assess whether high-centrality concepts tend to be covered more widely than low-centrality concepts, and whether this occurs for centrality as defined by all language editions. If we find this is the case, this would suggest that concept-level diversity is greatest at the periphery of each language edition.

The data in Table 3.6-b hints that, indeed, the most-central concepts in each language edition tend to be covered by more language editions than less central concepts. Namely, all of the concepts in Table 3.6-b have an English name. Since, as noted above, machine translation is not used in this thesis without an explicit declaration saying so, these names had to come from English Wikipedia articles about each concept, meaning that such articles exist. For all or nearly all concepts in Table 3.6-b, this is to be expected. It would be surprising if English, or any other language edition for that matter, did not have articles on concepts like the United States, France World War II, any recent year, and so on.

To explore the effect of centrality on concept-level diversity in a more robust fashion, we first divided up all concepts that appear in each language edition into twenty centrality quantiles.

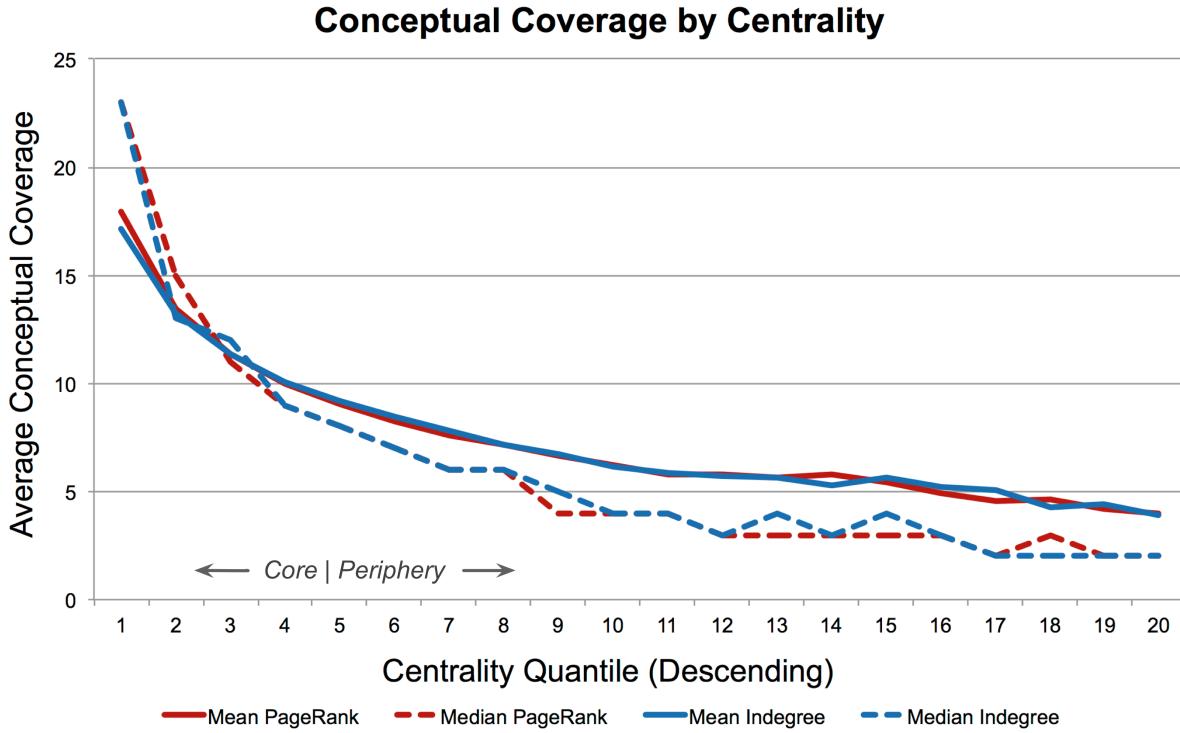


Figure 3.6-a: The grand mean/median of conceptual coverage for each centrality quantile. The value at  $x = 1$  of each line indicates the mean of the mean conceptual coverage at the 95<sup>th</sup> centrality percentile in our 25 language dataset (or median of the medians).

These divisions were language-specific because, as we found above, a concept may appear in a top quantile in one language edition, and a lower quantile in another. For each language edition, we then calculated the mean and median conceptual coverage of all the concepts in each quantile.

The results of this analysis can be found in Figure 3.6-a, which shows the grand mean (mean of the means) of the conceptual coverage in each quantile. A clear pattern is evident in Figure 3.6-a: as centrality decreases, so does conceptual coverage. Put simply, this means that more central concepts are covered by more language editions. Across all Wikipedias, a concept with a centrality in the top 5% will, on average, be covered by just over 17 (indegree) or just under 18 (PageRank) language editions. Moreover, the average such concept (median concept) in

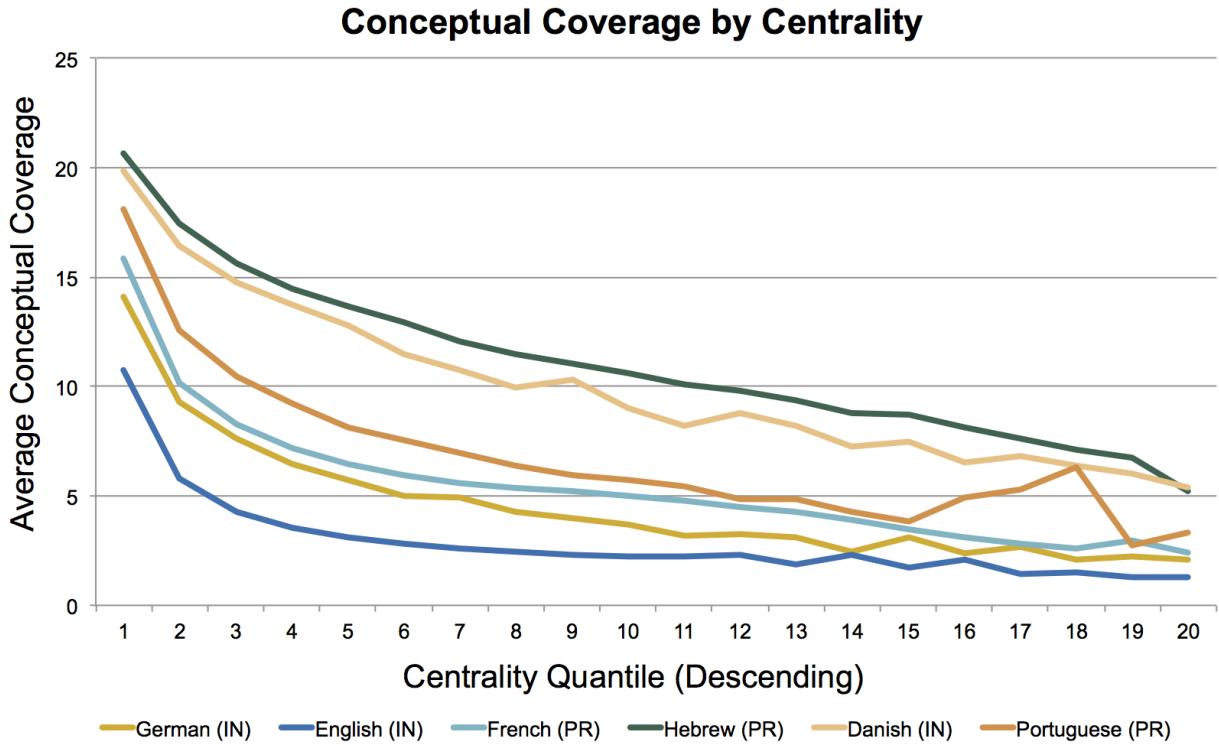


Figure 3.6-b: Conceptual coverage by centrality quantile in a selection of language editions and centrality metrics (IN = indegree, PR = PageRank). The tendency for smaller language editions to have higher values at all quantiles can be seen here. All values are conceptual coverage means.

the top 5% will be covered by exactly 23 language editions in both cases. Concepts in the bottom 5% have articles in many fewer language editions. A concept in this set is covered in fewer than four language editions on average, and the average such concept (median) is covered in just two languages<sup>34</sup>.

Figure 3.6-b zooms in on Figure 3.6-a by providing a language-specific view for a selection of Wikipedias. Several patterns are noticeable here. First, the values in the top quantile ( $x = 1$ ) are illustrative. In general, the larger language editions have a smaller value at  $x = 1$  than the smaller language editions. In fact, the smaller language editions tend to have higher mean conceptual coverage at all quantiles, a reflection of the fact that the average conceptual coverage

<sup>34</sup> Recall that we are considering grand means. This means that the mean of the mean number of language editions for a bottom 5 percent concept is less than four. The mean in English and other large language editions is smaller, while the mean in the smaller language editions is larger, as can be seen in Figure 3.6-b.

for a concept in these language editions tends to be higher than those in the larger language editions (Section 3.4). Another interesting phenomenon in Figure 3.6-b is the Portuguese spike in the lower quantiles. This indicates that some set of articles that is peripheral in Portuguese are covered by a large number of language editions. We intend to investigate this phenomenon in future work.

While the average conceptual coverage in centrality quantiles is one way of examining the role of centrality in concept-level diversity, another is to look at the sum of the centrality scores at different levels of conceptual coverage. For instance, if we sum the PageRank scores in a given language edition for single-language concepts, those for two-language concepts, and so on, we can get a different sense of how these concepts are situated in the various WAGs.

Table 3.6-d shows the aggregated PageRank score sums for single-language, non-global, and global concepts in a variety of language editions. This can be interpreted as the likelihood one would land on an article about a single-language concept, non-global concept, or global concept by randomly surfing around each language edition. Due in part to the fact that it has the

<b>Aggregate PageRank Score Sums by Conceptual Coverage</b>			
<b>Language Edition</b>	<b>Single-Language</b>	<b>Non-global</b>	<b>Global-Concepts</b>
Japanese	0.236 (Highest)	0.727	0.257
English	0.2	0.803	0.188
Swedish	0.140	0.651	0.326
Chinese	0.132	0.628	0.344
French	0.115	0.685	0.293
Spanish	0.089	0.621	0.361
Hebrew	0.09	0.552	0.435
Romanian	0.058 (Lowest)	0.488	0.492

*Table 3.6-d: The PageRank score sums for various types of concepts in a selection of language editions. The English row, for example, shows that only slightly more aggregated centrality is attributable to single-language concepts than global concepts, despite single-language concepts' massively larger numbers.*

highest ratio of single-language concepts, the English language edition has the second-largest aggregate PageRank score for single-language concepts (0.2), meaning that there is a 20% chance of arriving at a page on a single-language concept in the English language edition by randomly clicking a link. The corresponding number for global concepts is slightly less (0.18).

At 0.058, the probability of landing on a page about a single-language concept in the Romanian Wikipedia is the lowest. The equivalent number for global concepts in Romanian is 0.492, despite the fact that global concepts make up only 5.2% of Romanian concepts. It is important to note, however, that the PageRank score share for single-language concepts is not perfectly correlated with language edition size. Most notably, Japanese reprises its role as an outlier in this context by being the language edition with the ninth-most articles, yet having the highest aggregate PageRank score for single-language concepts. This is, perhaps, not a surprise given the fact that Japanese has almost the same percentage of single-language concepts as English (Section 3.4).

### **3.6.4 Centrality and Sub-concept-level Diversity**

In this section, we investigate if, like concept-level diversity, sub-concept-level diversity tends to be greatest in the periphery. Our basic approach here is adapted from that in the previous sub-section. Namely, we sample concepts from quantiles of each language edition's centrality distribution and calculate the sub-concept-level diversity for each quantile.

We use two sub-concept-level diversity metrics in the analyses below: *RatioInLang* and concept-averaged overlap coefficient ( $OC_c$ ). *RatioInLang* is analogous to the *RatioInEnglish* metric (Section 3.5.4), except generalized to all language editions. That is, for any language edition, *RatioInLang* reports the percentage of content in all of multilingual Wikipedia about a

given concept that is in that language edition's article about the concept.<sup>35</sup> More formally,  $RatioInLang$  is defined as follows:

$$RatioInLang(c, l) = \frac{|BOL_l|}{|BOL_{ALL}|}$$

where  $c$  is a concept that has an article in  $l$ , and  $BOL_l$  and  $BOL_{ALL}$  are the bags-of-links of the article about  $c$  in  $l$  and the union of all BOLs of articles about  $c$ , respectively. The concept-averaged overlap coefficient  $OC_c$  is the pairwise overlap coefficient as defined in Section 3.5.5, averaged for all pairs in a concept. As such, with  $OC_c$ , each concept  $c$  is given a score, rather than each article pair.

$RatioInLang$  and  $OC_c$  capture different types of sub-concept-level diversity.  $RatioInLang$  provides a perspective from the language edition under analysis  $l$ . That is, it measures the diversity of multilingual Wikipedia relative to  $l$ .  $OC_c$ , on the other hand, captures sub-concept-level diversity as it occurs across all language editions that cover a concept. Both metrics together provide a fuller picture of the relationship between sub-concept-level diversity and centrality.

Given the concept-level diversity results we saw earlier, one reasonable hypothesis is that sub-concept-level diversity will decrease as centrality increases, or that sub-concept-level diversity is mostly relegated to the periphery of multilingual Wikipedia. If this were the case, we would expect  $RatioInLang$  to be higher in very central concepts and lower in less central concepts. That is, in a situation where there is less sub-concept-level diversity in the core, each language edition should contain a higher percentage of the content in multilingual Wikipedia about concepts it considers highly central than about concepts it considers peripheral.

---

<sup>35</sup>  $RatioInLang$  is undefined when a concept does not exist in the language edition in question.

With regard to  $OC_c$ , if sub-concept-level diversity were relegated to the periphery, the  $OC_c$  should be lower in the periphery than in the core. Recall that  $OC$  measures the amount of content in the shorter of two articles about the same concept that is also in the longer of the two articles. As such, if the average  $OC$  is high in a group of concepts – e.g. core concepts – there are more cases of longer articles being complete supersets of shorter articles. If average  $OC$  is low – e.g. on the periphery – shorter articles tend to have more unique content, meaning there is likely to be more diversity overall.

Before reporting our *RatioInLang* and  $OC_c$  findings, it is important to note that in this study and in the similar studies in Sections 3.7 and 3.8, we use heuristic approaches to bag-of-link generation. In Section 3.5, we reported our results using a variety of permutations of the various parameters that go into generating bags-of-links. However, utilizing our current single-machine setup, this is a very time-consuming process. For Section 3.5, we were able to use relatively small numbers of concepts in our experiments, reducing the impact of this problem. Here, however, we needed to calculate the BOLs for hundreds of concepts in each centrality quantile in each language edition. This meant that we needed to develop heuristics.

Examining our data from Section 3.5, we found that the results using the different sets of parameters were often quite correlated, especially in their ranks. The “kitchen sink” upper-bound and “just links” lower-bound results for the *RatioInEnglish* metric, for instance, displayed an  $r_s$  of 0.83 and an  $r$  of 0.79. Given this high correlation, in the case of *RatioInLang* we felt it was reasonable to use the lower-bound strategy as a proxy for the upper-bound strategy, even though the metrics were slightly different. While this prevents us from obtaining absolute values, we retain our ability to make claims about relative values of sub-concept-level diversity. Given that the goal of this section is to determine if sub-concept-level diversity rises or falls with centrality,

this was more than enough for our purposes.

The situation with  $OC_c$  was a bit more complex. We calculated the  $OC_c$  for 200 randomly selected concepts and found that the correlation between the upper-bound and lower-bound was not as strong ( $r_s = 0.54$ ,  $r = 0.57$ ). As such, we complemented our heuristic  $OC$  results with results generated on a smaller sample but using middle-ground BOL parameters that were more correlated with the upper-bound ( $r_s = 0.80$ ,  $r = 0.81$ ).

After finalizing our sub-concept-level diversity metrics, we divided up the PageRank score distribution of each language edition into 10 bins of equal width and then sampled 1,000 concepts from each bin. All 1,000 concepts had to have an article in more than one language edition, and no minimum number of outlinks was set. We did not consider the indegree distribution here because we saw that the concept-level results for indegree and PageRank were quite similar and calculating the two metrics for all  $1000 * 10 * 25 = 250,000$  for both PageRank and indegree would have excessively taxed our limited computational and I/O resources.

For all 1,000 concepts in each PageRank score quantile, we calculated the  $RatioInLang$  and  $OC_c$  values and averaged these values over each bin. The results of this analysis can be found in Figure 3.6-c, which shows the average  $OC_c$  of PageRank quantiles by language edition, and Figure 3.6-d, which shows the same for  $RatioInLang$ .

It is clear from the results in both figures that sub-concept-level diversity *increases* as centrality increases, which is the exact opposite of what we found with concept-level diversity. In Figure 3.6-c, the most central quantile is the quantile with the lowest average  $OC_c$  in all but one language edition. This means that longer articles cover the *smallest* amount of shorter same-concept articles in the core, a strong indicator that sub-concept-level diversity is greatest in the core. While there is individual variation among the language editions, the average amount of

sub-concept-level diversity as measured by  $OC_c$  slowly decreases as one moves towards the periphery, with the exception of a small spike that occurs in the last centrality quantile.

The purple line in Figure 3.6-c shows that we were able to repeat the above  $OC_c$  results using the more resource-intensive wikification strategy whose correlation with upper-bound strategy is higher. Like the black line, the purple line represents averaged ranks across several language editions, but instead of all 25 language editions, we only calculated the  $OC_c$  in this case for eight language editions (English, German, French, Dutch, Italian, Russian, Spanish, Italian). In addition, only 100 concepts per quantile were considered. Regardless, however, the fact that the overall trend is the same using this wikification strategy means that it is quite likely the upper-bound strategy would result in similar findings.

## Overlap Coefficient by Centrality

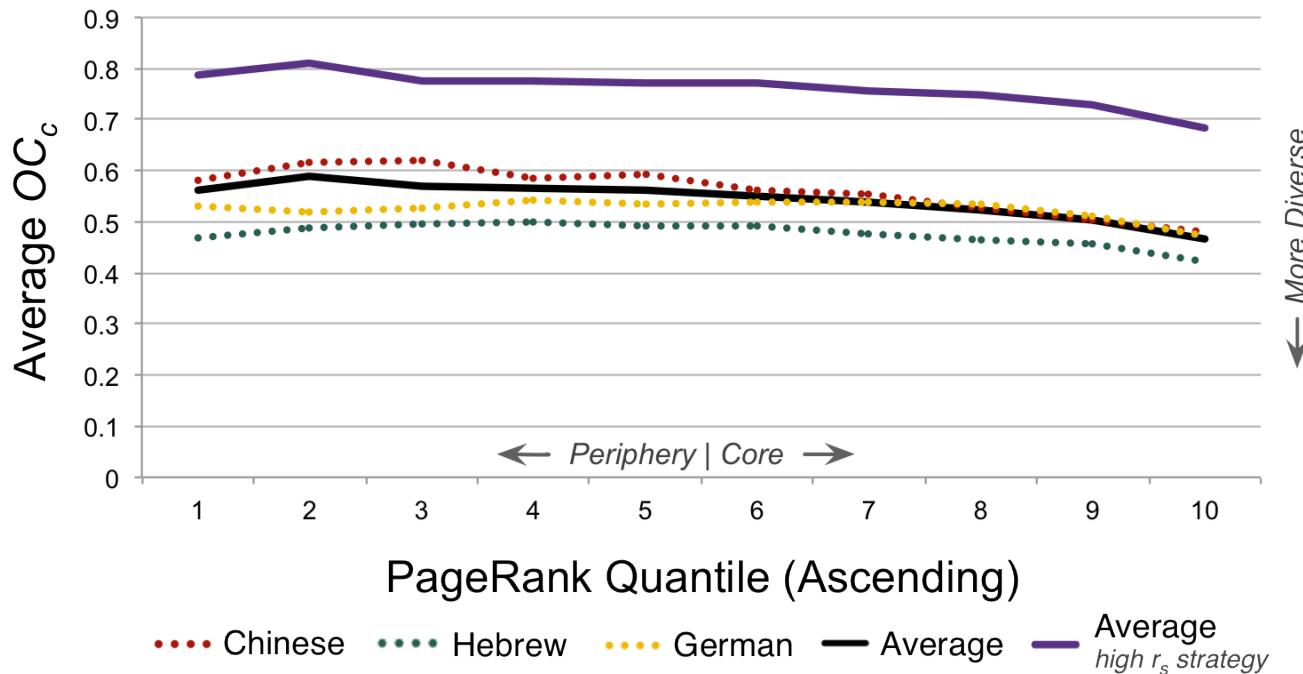


Figure 3.6-c: The average  $OC_c$  for each of 10 centrality quantiles for several language editions, the mean of all language editions, and the mean of all language editions using a wikification strategy with a higher correlation with the upper-bound. The most-central quantile has, on average, the lowest-ranked  $OC_c$  value, meaning that longer articles cover the smallest amount of shorter articles in the core. As more and more peripheral quantiles are examined, the amount of unique information in shorter articles decreases. In other words, according to  $OC_c$ , diversity is greatest at the core and weakest at the periphery.

Roughly the same pattern with *RatioInLang* can be seen in Figure 3.6-d. This figure reveals that articles about a language edition's central concepts are actually missing *more* information than articles about peripheral concepts. In other words, there are in general more diverse views in multilingual Wikipedia about concepts that are in the core of each language edition than about those that are on the periphery.

Here, the tendency for there to be more sub-concept-level diversity in the core is less of a surprise. We saw above that central concepts tend to be covered in more language editions, meaning there will be more opportunities for unique information to be added to multilingual Wikipedia about these concepts. However, this does not take away from the overall finding that, on average, a reader of a single language edition's article about a central concept is getting a lower percentage of multilingual Wikipedia's knowledge about that concept than she would if the concept were on the periphery.

It is also important to note that while the tendency for the core to have more sub-concept-level diversity than the periphery was representative of all language editions in the case of  $OC_c$ , this was not true with *RatioInLang*. Namely, in English and the East Asian language editions, there did not appear to be a relationship between centrality and *RatioInLang*<sup>36</sup>. These Wikipedias' divergence from the norm here is worthy of future study, especially given the fact that all three East Asian encyclopedias exhibit this property.

---

36 With  $n = 10$ , it is difficult to use techniques like Spearman's correlation to determine if this is the case in a more robust fashion.

## Ratio of Information in Language by Centrality

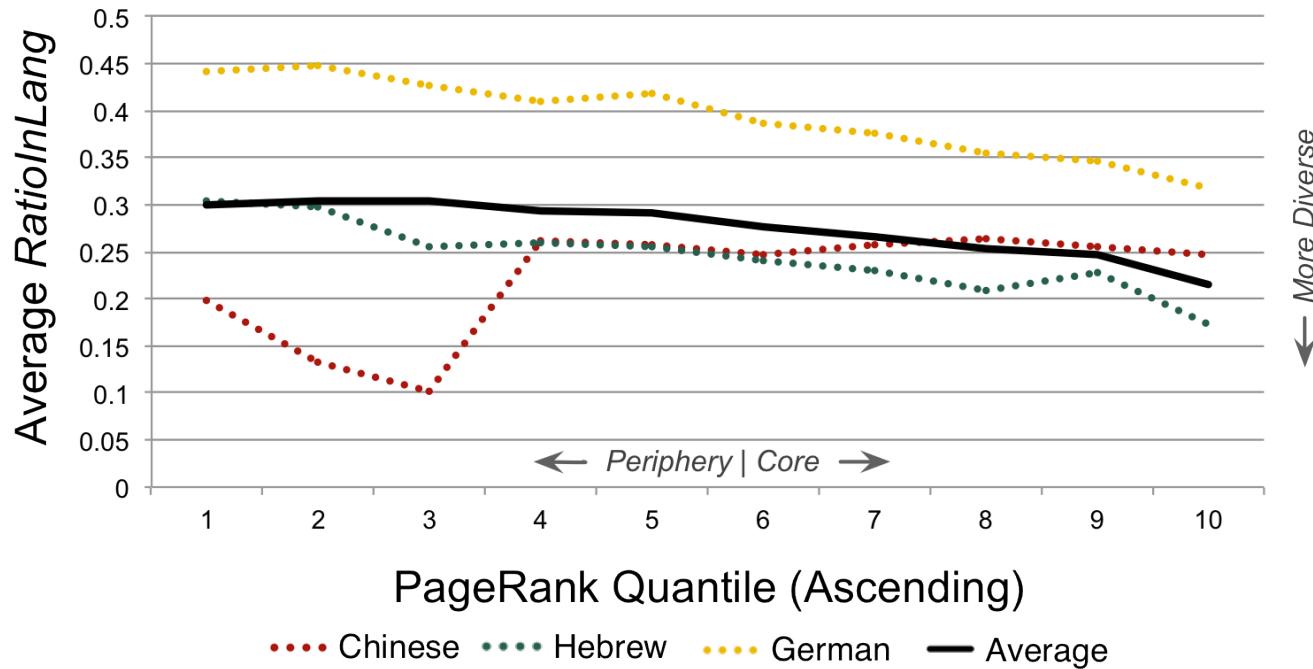


Figure 3.6-d: The share of multilingual Wikipedia's information about a given concept that appears in a language edition by the centrality of that concept in the language edition. On average, an article about a concept that appears in the core of a language edition has a smaller share of multilingual Wikipedia's total information about that concept than would be the case if the concept were in the periphery. In other words, there is more sub-concept-level diversity in the core than in the periphery according to this metric as well. Note: Since we use lower-bound wikification strategies in this study, the absolute values have little meaning. The relative values – i.e. the trends – are where the information lies.

### 3.6.5 Discussion

In this section, we have made three contributions that help to shed light on the role of centrality in encyclopedic world knowledge diversity:

1. We showed that the concepts that are most central to each language edition’s article graph vary extensively from language edition to language edition. Just because a concept is in the central “core” of one language edition does not mean it is in the core of another language edition.
2. We established that highly central concepts are covered by a much larger number of language editions than concepts that are less central. In other words, concept-level diversity is much greater in the periphery than in the core. This occurs for the core/periphery as defined by all language editions.
3. We demonstrated that the opposite is true for sub-concept-level diversity. Core concepts are described in a more diverse fashion in each language edition than periphery concepts.

In most other sections of this chapter, all or nearly all findings point in a single direction supporting the global diversity hypothesis. With centrality, our results demand more nuanced interpretation. In particular, the fact that concept-level diversity is less common in the core of each language edition than in the periphery provides probably the best support for the global consensus hypothesis in this entire thesis. Centrality is a proxy for how “important” a language edition considers a concept, and it is reasonable to argue that if concept-level diversity occurs most often with less important concepts, then concept-level diversity itself is less important.

However, this argument must be considered in the context of three additional important

findings. First, we showed that the very definition of what is “important” varies somewhat extensively from language edition to language edition. Second, the fact that concept-level diversity is less common in the core just means that sub-concept-level diversity becomes more prominent and it is more prevalent in the core than in the periphery. Finally, centrality is only one way of measuring the importance. Another method is to use content consumption metrics like page views, which are the focus of Section 3.8. In that section, we will see that when adopting this more reader-focused metric of importance, this one substantial piece of support for the global consensus hypothesis quickly breaks down.

It is important to note the limitations of the work in this section. First, we only consider two forms of network centrality. There are many other centrality measure types and variations on those types, and it is possible that these could provide a different perspective on the findings above. Second, as noted above, our sub-concept-level metrics are heuristics. While we established that the heuristics are highly correlated with the upper-bound values, there is a small chance that using the true upper-bound values would cause any small-effect results to be slightly different.

The third limitation of the work in this chapter is that we do not consider “missing links” (Section 3.5.1.2) in our network centrality calculations. Including missing links would have involved “wikifying” all 17+ million articles in our 25-language dataset, something that would have required a substantial increase in available computational and I/O resources. While it is possible that using the “wikified WAG” of each language edition could result in different conclusions, there are a number of reasons to believe this is unlikely. Primarily, as noted above, we showed that our “just links” heuristic that does not include any wikification at all is quite correlated with our most-complete wikification strategy. Since our centrality approach is also

“just links,” this suggests that while the absolute indegree values might change with wikification, the relative values will stay roughly the same. Moreover, PageRank scores are always relative. As this chapter is almost entirely focused on comparing relative centrality across language editions, it is likely that the lack of missing links in our network centrality calculations has little effect on our overall conclusions<sup>37</sup>.

---

37 The one exception might be in the case of systemic difference in linking behavior, the only confirmed instance of which occurs with years. That is, with wikification, English and German might also consider years to be highly central concepts like the other language editions.

## 3.7 Topic Diversity

In the sections above, we often informally group concepts by topic in order to better understand an underlying phenomenon. For instance, in Section 3.4 we discussed how countries and other geographic topics appear to form an important part of the global encyclopedic core. The goal of this section is to perform more formal analyses of the role of topics in Wikipedia concept-level and sub-concept-level diversity and, in doing so, to shed light on the relationship between topic areas and the cultural contextualization of user-generated content more generally.

Our primary research question here is “Does the same level of concept-level and sub-concept-level diversity exist across all topics, or do certain topics display more diversity – and corresponding cultural contextualization – than others?” In other words, the experiments in this section will seek to establish if, for example, concepts related to American football tend to be covered in fewer language editions and have less diversity in their articles than, say, concepts related to soccer. If this proves true, it would suggest that information about American football tends to be siloed in a very small number of language editions (e.g. English), while soccer information is shared more widely.

We begin this section with a brief discussion of how we assigned Wikipedia concepts to topics by leveraging the work of other researchers in the Wikipedia space. We then investigate concept-level diversity across a large number of topics, again using conceptual coverage as our primary metric. Finally, we perform a similar analysis with regard to sub-concept-level diversity.

### 3.7.1 Assigning Concepts to Topics

The major methodological challenge in the analysis of Wikipedia diversity across topic areas is the assignment of concepts to well-defined topics. Unfortunately, this challenge is

currently an open one, at least with respect to multilingual Wikipedia. We investigated two approaches to developing solutions or workarounds to this problem. First, we attempted to adapt Kittur et al.’s English-only Wikipedia Category Graph (WCG)-based topic assignment method [103] to a multilingual context. The Kittur approach works by giving an English Wikipedia article weights according to its relative membership in each top-level English Wikipedia category. For instance, the article “Albert Einstein” (English) has non-zero weights for the top-level categories “Science,” “People,” “Technology,” and so on. As Kittur et al.’s algorithm is largely based on path length in the WCG – the same technique used by the *WikiRelate* semantic relatedness measure discussed in Chapter 6 – we were able implement it in WikAPIdia relatively easily. Unfortunately, we found that the top-level categories of each language edition were sufficiently different in number and semantics so as to make the language-defined topic schema mostly incompatible with one another. We were also concerned that even though Kittur et al.’s approach was determined to be accurate in the English language edition through a Mechanical Turk-based evaluation, the uncertain quality of the WCG in the other language editions might significantly reduce its efficacy outside of the English context.

The other approach to concept topic assignment we considered was to leverage the active area of research that involves integrating Wikipedia into existing ontologies and/or developing a new ontology around a formalized version of Wikipedia. One of the most well-known successes in this area of work is YAGO2s [92, 194], a “semantic knowledge base” that merges together Wikipedia, WordNet, and GeoNames, a prominent web-based gazetteer. For the purposes of this section, YAGO2s has many advantages. First, it provides topic assignments of variable granularity; using YAGO2s, we can assign concepts to WordNet synsets for an understanding of diversity across relatively fine-grained topics and WordNet domains (as defined by the synset-to-

domain mappings introduced by Magnini and Cavaglià [125]) for a broader perspective. Second, YAGO2s is a finished product and has undergone a robust evaluation.

That said, YAGO2s also has a critical drawback in the context of our research: while it is claimed that YAGO2s is a *multilingual* semantic network, YAGO2s's version of multilingual is that which conforms to the English-as-Superset hypothesis. In other words, no concept that does not appear in the English Wikipedia appears in YAGO2s. This means that none of the more than four million concepts in our multilingual Wikipedia dataset that are not described by an English article can be assigned topics using YAGO2s.

In the end, however, we decided that while YAGO2s's English-as-Superset drawback would significantly and meaningfully constrain the types of analyses we could perform and the conclusions we could draw from these analyses, it was better than struggling to find ways to formally integrate the top-level categories across 25 language editions and to evaluate the quality of the non-English WCGs. Both of these would have been required if we had attempted to adapt the Kittur approach to a multilingual context, and both of these are unsolved problems. Of course, we could have used Kittur's approach in an English-as-Superset fashion as well, but the main advantage to using Kittur et al. here was the potential for it to allow us to execute analyses free from the constraints of the English Wikipedia. Compared directly in the English Wikipedia context alone, YAGO2s is a much more recent, much more ambitious effort to develop a general-purpose semantic network around information in Wikipedia, whereas Kittur et al.'s approach was intended (and was successfully used) to provide ballpark figures to help researchers understand, for instance, the types of articles where editor conflict most frequently appears [103].

Once the decision to use YAGO2s was made, it was a straightforward matter to download the entire YAGO2s network and use the relationships in this network that are incident with

English Wikipedia articles to assign topics to the set of concepts in our dataset that have an English article. As noted above, we considered two types of YAGO2s-based topics – WordNet synsets and WordNet domains – and we report our results using both below.

When we matched a concept to a synset, we also matched that concept to all of the synset’s “parents” according to the “is-a” hypernymy/hyponymy relationships in WordNet (as included in YAGO2s). We then used all of these synsets and the synset-to-domain mappings discussed above to assign WordNet domains to that concept. For instance, the concept that is described by the article “Andorra” (English) is assigned to the WordNet synsets “Country,” as well as to its parents “Administrative district”, “Region”, “Location”, “Physical entity” and so on. Using these synsets and the synset-to-domain mappings, the same concept is then assigned to the WordNet domains “Geography,” “Administration,” “Town planning,” and “Factotum” (which is another way of saying “miscellaneous”).

The final two of these domain assignments – particularly the “Town planning” assignment – raise an important issue: even though YAGO2s has been robustly evaluated [92, 194], it is not perfect, and Magnini and Cavaglià’s synset-to-domain mappings introduce additional error. Moreover, topic assignment can be a highly context-sensitive activity. In certain situations, Albert Einstein may be most usefully categorized as a prominent American Jew, while in others his achievements in science or nuclear non-proliferation might be most relevant ([103]). Below, we address these issues in two ways. First, we largely do not consider domains with more than 20,000 concepts, which eliminates cases like “Town planning.” Second, we are careful to include in the text a small set of randomly selected concepts for almost all topics that we discuss in any detail. The goal here is to provide the reader with an idea of how these topics are interpreted by YAGO2s.

### 3.7.2 Concept-level Diversity by Topic

Once we had mapped all possible concepts to synsets and domains, we could begin using these synsets and domains as topic groups with which to analyze the average concept-level diversity on a topic-by-topic basis. Due to YAGO2s's English-as-Superset assumption, concept-level diversity in this chapter means something different than in the chapters that precede and follow it. Here, when a concept has minimum concept-level diversity, it is covered by the English Wikipedia and no other language editions. All single-language concepts from language editions other than English are left out of the analyses, as are concepts with any level of conceptual coverage that do not have articles in the English Wikipedia.

Following our centrality quantiles approach in Section 3.6, to understand the role of topic areas in concept-level diversity, we first grouped concepts belonging to each topic into bins. We then calculated the average conceptual coverage of each bin. We did this separately for YAGO2s's WordNet synsets, which provide a more fine-grained view, and its WordNet domains, which are much broader in scope.

With regard to WordNet synsets, we found that topic plays a critical role in concept-level diversity, with some topics being covered widely by many language editions and other topics being largely the sole province of the English Wikipedia. Moreover, the average coverage for many topics of both types is exactly what would be expected under the hypothesis that user-generated content reflects the cultural contexts of its contributors. Topics that are in many language-defined cultural contexts tend to have high average conceptual coverage, while the reverse is true for topics that tend to be parochial to the world of English speakers.

Tables 3.7-a and 3.7-b show the synsets with the highest and lowest conceptual coverage, respectively, along with several randomly-selected example concepts for each synset. Only

synsets with 25 or more concepts are shown. Numerous themes are present in both tables. First, Table 3.7-a makes it clear that, indeed, certain types of geographic concepts are some of the most-well covered concepts in all language editions. Topics like provincial capitals and prefectures have some of the highest average conceptual coverages, with provincial capitals appearing in 15.99 language editions on average and prefectures appearing in 14.59<sup>38</sup>. If we were to extend the length of Table 3.7-a, we would find topics like “City” with very large numbers of concepts yet with relatively high average coverage (“City” has 40,369 concepts yet still has an average coverage of 8.81). Moreover, the “Economy” synset, which has the sixth-highest conceptual coverage overall, is predominantly made of countries, an example of the difficulties in topic assignment discussed above.

Another theme in Table 3.7-a is the extensive coverage of royalty-related topics. The synsets “Crown prince,” “Crown princess,” “Grand duke,” and “Dauphin” are all at the very top of the conceptual coverage spectrum. Additionally, just barely missing the cut for Table 3.7-a were the topics “Shogun” and “Czarina,” indicating that robust coverage of this subject area is not limited to Western subjects. That said, also barely missing the cutoff were other Western royalty-related synsets such as “Queen Mother,” “Queen Dowager,” “Grand duchess”, “Archduchess,” and so on.

The largest synset in Table 3.7-a is that of laureates, which is made up of a large number Nobel laureates but also contains laureates of other prizes as well. Over 87 percent of laureate concepts are covered in language editions other than English and 197 of them are global concepts (e.g. Václav Havel and Eugene O’Neill).

---

<sup>38</sup> For context, recall from Section 3.6 that the average conceptual coverage for a concept that has an English article is 2.91.

Christianity-related synsets are also featured prominently in Table 3.7-a. From a culturally contextualized UGC standpoint, this is not a surprise given the language editions considered here. Christianity is not a predominant religion in only four of the 25 corresponding language-defined communities. As such, it would follow that popes<sup>39</sup> and epistles tend to be widely covered. Moving from religion to science, Table 3.7-a also provides several examples of fairly specific, high-coverage natural science-related topic areas, of which there are hundreds in the overall dataset.

---

39 This perhaps is a bit more of a surprise as some language-defined cultures considered here are predominately Protestant.

Synset	#	Mean Coverage	Global concept examples	English-only examples
Crown prince	25	20.68	George IV of the United Kingdom; Wilhelm II	Henveyru Ganduvaru Manippulu
National flag	279	19.73	Flag of Afghanistan; Flag of Liechtenstein	Canadian Parliamentary Flag Program
Commander in chief	29	16.97	Kim Il-sung; Joseph Stalin; Francisco Franco	
Statesman	387	16.90	Nicolas Sarkozy; Jóhanna Sigurðardóttir	Hamid Bin Ahmad Al-Rifaie; Frederick C. Alderdice
Colossus	29	16.86	Atlas(mythology); Themis; Oceanus; Mnemosyne	T-Rac; Titans Tomorrow; Team Titans
Economy	333	16.82	East Germany; Puerto Rico; Malta; Uruguay; Tunisia	Amsterdam Entrepôt; Legal origins theory
Crown princess	48	16.23		
Pope	436	16.07	Pope Leo X; Pope Leo III; Pope John Paul I; Pope Pius XII	Pope Gabriel VIII of Alexandria
Provincial capital	311	15.99	Mecca; Nanjing; The Hague; Shenyang; Cape Town	Attapeu; Hà Tĩnh city; Bac Kan; Svay Rieng (town)
Bird family	149	15.89	Columbidae; Grebe; Kiwi; Penguin; Heron; Swift	Anseranatidae; Tree kingfisher; Water kingfisher
Grand duke	130	15.42	Jogaila; Henri, Grand Duke of Luxembourg	Grand Princes of Tuscany
Dauphin	35	15.40	Henry II of France; Louis XV of France	Beatrice of Albon
State capital	161	15.21	San Jose, California; New York City; Trenton, New Jersey	Isanlu Isin; Omupo; Warrap, South Sudan; Issele-Uku
Deist	31	14.90	Marlon Brando; Jean-Jacques Rousseau; Tupac Shakur	Thomas Davison; Fenwicke Holmes; Hal Bidlack
Protectorate	42	14.71	Zanzibar; Federated States of Micronesia; Cook Islands	Malagasy Protectorate; Ashanti Protectorate
Prefecture	375	14.59	Limoges; Poitiers; Marseille; Chiba Prefecture; Strasbourg	Prefecture Apostolic of Kaiservilhelmsland;
Laureate	1550	14.47	Eugene O'Neill; Günter Grass; Hermann Hesse	Inamullah Khan; Julij Betetto; Aleksander Zorn
County town	121	14.33	Cork(city); Oxford; Durham; Leicester; Cambridge; Nottingham	Llanfachreth
Tyrannosaur	29	13.83		
Seal	40	13.68		Leptophoca; Ragged-jacket; Lobodontini; Freshwater seal
Antipope	43	13.53		Pope John (numbering)
Wading bird	109	13.51		Druridge Bay curlew
Demigod	32	13.50	Theseus; Helen of Troy; Heracles; Minos	Amphitheus I; Carmanor; Cyamites
Epistle	50	13.40	Epistle to the Romans; Second Epistle to the Corinthians	First Letter (Plato); Epistle of Pseudo-Titus; Faithful saying

Table 3.7-a: The synsets with the highest average conceptual coverage (and 25 total concepts or more). For each synset, we provide several randomly selected global and English-only concepts where possible.

Synset	#	Mean Coverage	Example Concepts
Senior high school	16837	1.065	Lakeview Academy; J.J. McClain High School; Eaton Community College; Old Rochester Regional High School
Anchor	32	1.063	Jim Vance; Johnny Mountain; Sylvia Perez; Sniggle; Glenn Brenner; Kevin Corke; Bill Bonds; John Hambrick; Cynthia Gouw; Sally Thorne
Belemnite	92	1.054	Zugmontites; Buelowiteuthis; Youngibelus; Pseudohastites; Protoaulacoceras; Raphibelus; Belemnocamax; Belemnelloamax; Belemnitina
Masquerade	40	1.050	The World Tossed at Tennis; The Triumph of Beauty; The Sun's Darling; Pleasure Reconciled to Virtue; The Gypsies Metamorphosed
Church school	42	1.048	Seymour College; Scotch Oakburn College; Kingswood College (Box Hill); Forest Lake College; Scotch College, Adelaide
Icefall	48	1.042	Minnehaha Icefalls; Catcher Icefall; Cranfield Icefalls; Cherry Icefall; Sledgers Icefall; Cooper Icefalls; Wild Icefalls; Shackleton Icefalls
Barn	164	1.037	Tim Thering Octagon Barn (Plain, Wisconsin); Miller Round Barn; Connected farm; Frank Senour Round Barn
Airstrip	29	1.034	Deblois Flight Strip; Accomack County Airport; Napa County Airport; Battle Mountain Airport; Leadville Municipal Airport
Grammar school	481	1.031	Bedlingtonshire Community High School; Queen Elizabeth's High School; Canberra Grammar School; Rugby High School for Girls
Log cabin	71	1.028	Abner Williams Log House; Thomas Brown House (Inwood, West Virginia); Ipsut Creek Patrol Cabin; Adsit Log Cabin
Session	74	1.027	Sixty-ninth Texas Legislature; Twentieth Texas Legislature; Tenth emergency special session of the United Nations General Assembly
Electorate	297	1.027	Pensioner Settlements (New Zealand electorate); Temuka (New Zealand electorate); Hawkes Bay (New Zealand electorate)
Barbershop quartet	82	1.024	Oriole Four; Nightlife (quartet); The Ritz (quartet); Michigan Jake; The Jazz Firm; Four Teens; Revival (quartet); The Suntones
Hurdles	96	1.021	Spring Juvenile Hurdle; Prestige Novices' Hurdle; Martin Pipe Conditional Jockeys' Handicap Hurdle; Sharp Novices' Hurdle
Hang glider	57	1.018	Icaro Laminar; Flight Design Exxtacy; Ellipse Zenith; Aeros Combat; Helite Tsunami; Flugschule Wings Alfa; Europe Sails Special Dimension
Training school	59	1.017	Newlands Girls' School; Sacred Heart High School (London); Penair School; Coombe Girls' School; Katharine Lady Berkeley's School
Comprehensive school	1375	1.014	The Skinners' Company's School for Girls; Cardinal Griffin Catholic High School; Ashlyns School
Football season	75	1.013	Georgia Bulldogs football under Harry Mehre; Minnesota Golden Gophers football under Murray Warmath
Charter school	440	1.007	Integrated Day Charter School; Guajome Park Academy; Heritage Academy (Arizona); Fountain Square Academy
Weatherman	62	1.000	Lukwesa Burak; Chris Fawkes; Daniel Corbett; Suzanne Charlton; Richard Edgar; Bert Foord; Tori Lacey; Nick Miller (weather forecaster)
Sorority	29	1.000	Kappa Zeta Phi; Sigma Psi Zeta; Alpha Kappa Delta Phi; Delta Phi Lambda; Kappa Delta Chi; Sigma Alpha Omega; Sigma Iota Alpha
Secondary modern school	28	1.000	Cottesloe School; Lostock College; Wye Valley School; Mandeville Upper School; Waddesdon Church of England School; Great Marlow School

Table 3.7-b: The synsets with the lowest average conceptual coverage (and 25 total concepts or more).

We now turn our attention to the synsets with the lowest average conceptual coverage, shown in Table 3.7-b. From a cultural contextualization standpoint, these synsets and others with very low coverage are of three varieties. First, some of these synsets are comprised of concepts that have articles only in the English Wikipedia most likely not because they are outside of the cultural contexts of the other language editions, but for other reasons, for example the English Wikipedia’s status as having the largest group of editors. This is probably the case for the “Belemnite” synset in Table 3.7-b. Belemnites are an order of long-extinct animals.

The two other varieties of low-coverage synsets, however, depend heavily on cultural contextualization as a causal factor and appear to represent a much larger number of synsets. To understand the first of these two varieties, which we call *topic-parochial synsets*, consider the extremely low-coverage “Sorority” synset, which is tied with “Weatherman” and “Secondary modern school” for the lowest average conceptual coverage in our entire dataset (among synsets with at least 25 concepts). Kappa Zeta Phi, Sigma Psi Zeta, and other concepts belonging to this synset are, by and large, only notable in the English-speaking world, and in the United States more specifically. The rare “international” sorority is one that generally has several chapters in Canada (e.g. Alpha Phi [5]). The low conceptual coverage of synsets such as these can be explained by the *entire subject* of the synset – i.e. the entire topic – being relevant only to English speakers. The domains they describe simply do not exist in the home regions of the other language-defined cultures. Table 3.7-b’s “Barbershop quartet,” among others, can also be interpreted in this fashion. Moving just past the cut-off for Table 3.7-b, we would find additional topic-parochial synsets such as “Savings and Loan” (mean = 1.08) and “School district” (mean = 1.09). Similarly, if we remove the 25-concept requirement, we would find a set of topic-parochial synsets with a mean coverage tied with that of “Sorority” that reads like a “who’s who” list of

English-speaker-centric topics: “Glee club,” “Homestead,” “Dude ranch,” “Cricket match,” and so on.

The general area of sports can provide a particularly clear understanding of topic-parochial synsets. Table 3.7-c shows a selection of sports-related synsets from across the conceptual coverage spectrum. Synsets that are related to soccer, the most popular sport in the world [164], tend to have high average conceptual coverage, even though the number of concepts in each synset can be quite large. Moreover, the average of these synsets is likely pushed down significantly due to YAGO2s’s disambiguation issues with soccer and American football. Note that the randomly selected English-only concepts for the synset “Football team” are both teams that play American football, not soccer. Moving towards the middle of the conceptual coverage spectrum, we find synsets like “Basketball league” and “Hockey league,” both of which

Synset	#	Mean Coverage	Global concept examples	English-only examples
Sweeper	27	12.07	Franz Beckenbauer; Arsène Wenger; Franco Baresi	Chad Gibson; Diogo Matos
Football team	1558	5.41	Belgium nat'l football team; Morocco nat'l football team;	Northeast Ohio Panthers; Sacramento Rush
Football league	1265	5.14	Serie B; Gambrinus liga; Ekstraklasa; Tippeligaen	Trelawny League; Thai Super Cup
Basketball league	215	4.27	National Basketball Association	National Alliance of Basketball Leagues
Hockey league	377	3.22	National Hockey League	Golden Horseshoe Junior Hockey League
Tight end	978	1.44		Robert Royal; Steve Bush; Richard Dickson; Kerry Cash
Cornerback	1346	1.42		Nate Allen (cornerback); Thom Darden; Weldon Brown
Ballpark	48	1.10		Joe Wolfe Field; Cranston Stadium; Robbie Mills Field
Wicket-keeper	957	1.08		Hector Hyslop; Alec Davies; Farokh Engineer; William Deane (cricketer)
Cricket match	10	1.0		Bicentennial Test; The University Match (cricket)

Table 3.7-c: The conceptual coverage of various sports-related synsets.

correspond to sports that experience moderate global popularity, but certainly not to the same extent as soccer. Finally, as we saw in Table 3.7-b, the low end of the conceptual coverage spectrum is replete with synsets related to sports that are only well-known in the English-speaking world, primarily American football and cricket. Over 75% of American football tight ends in the English Wikipedia (at least as categorized by YAGO2s) do not have articles in any other language edition, a number that increases to over 95% for wicket-keepers. There is a case to be made that cricket is the second most-popular sport in the world next to soccer, yet because this popularity is limited to English-speaking countries like the United Kingdom, India, and Pakistan, cricket-related articles rarely appear in any non-English language editions in our dataset. If we had included the Urdu or Bengali language editions, however, this would likely change, especially as these language editions grow.

The second type of culture-dependent low-coverage synsets we call *instance-parochial synsets*. Instance-parochial synsets describe an overall idea that *is* relevant outside of the English-speaking world, it is the concepts contained within these synsets that tend to be much less well-known and cause the low average coverages. Consider, for example, the “Airstrip” synset in Table 3.7-b. Obviously, the concept of airstrips is well-known outside of the United States, Canada, the United Kingdom, and other English-speaking countries<sup>40</sup>. However, the same cannot be said of *specific* airstrips. It would be surprising if, say, the Dutch Wikipedia had no article about airstrips in general and no articles about airstrips in the Netherlands, but it is hardly a shock that the Dutch Wikipedia does not have articles about the Accomack County Airport in Virginia or the Deblois Flight Strip in Maine. Other synsets in Table 3.7-b that follow a similar pattern include “Barn” and “Weatherman.”

---

<sup>40</sup> In fact, the airstrip concept has an article in all 25 language editions.

Synset	#	Mean Coverage	Global concept examples	English-only examples
<b>Highest Average Coverage Domains</b>				
Astronautics	1276	8.290	Michael Collins (astronaut); Neil Armstrong; Vladimir Komarov	Apstar VI; CRRES; Aleksey Ovchinin; MidSTAR; Rock (comics)
Astronomy	4421	6.640	Isaac Newton; Canopus; Uranus; Enceladus (moon)	SU Andromedae; United States imprisonment rate; Farzana Aslam;
Betting	1863	5.910	Günter Grass; Oscar Niemeyer; Bertolt Brecht	Tip jar gaming; Ellen Goodman; Marie de Sabrevois
Tennis	4411	5.740	Justine Henin; Steffi Graf; Pete Sampras; Maria Sharapova	Markus Günthardt; Melissa Brown (tennis)
Skiing	4244	5.420		Snowshoe Thompson; Patrizia Bassi; Joanne Duffy
Heraldry	19358	5.030	Tarja Halonen; Pope Pius XII; Lee Myung-bak; Bob Hope	Viscount Goderich; Cann baronets; Muir baronets;
Chess	2368	4.990	Alexander Alekhine; Marcel Duchamp; Bishop (chess)	Bruno Edgar Siegheim; Louis Uedemann; Adolf Seitz
<b>Lowest Average Coverage Domains</b>				
Pure Science	2788	2.335	Massachusetts Institute of Technology; Boeing; CERN	Watson Institute for International Studies; Baran Unity
Fencing	3612	2.317		Jarosław Kisiel; Ralph Chalmers; Ali Murat Dizioğlu;
Finance	2862	2.298	Who Wants to Be a Millionaire?; Viktor Yushchenko;	Joshua Peter Bell; Philip Daly; John Cox Bray; John Barker Church
Golf	4985	2.226	Bunker; Linus Torvalds; Ella Fitzgerald; Leonard Bernstein	Yeh Wei-tze; John Coleman (Australian footballer)
Baseball	2969	1.960	Hamburger SV; Nazca Plate	Rome Braves; Niagara Stars; Hamilton Thunderbirds
Racing	2321	1.602		Hood to Coast; A G Hunter Cup; Long Walk Hurdle
Cricket	15363	1.092	Samuel Beckett; Arthur Conan Doyle	Peter Phelps (cricketer); Pramila Bhatt; Frank Hayes (cricketer)

Table 3.7-d: The domains with the highest and lowest average conceptual coverage.

Moving away from synsets, Table 3.7-d shows the results of an identical analysis using the WordNet domains resource in YAGO2s. As noted above, we did not consider domains with more than 20,000 concepts as we found that these domains were subject to excessive noise. We also limited our analysis to domains that had at least 1,000 concepts in order to focus on the key benefit of domains: topic assignments at a higher granularity.

The key take-away from Table 3.7-d is the quantitative confirmation of some of our qualitative grouping of synsets above. For instance, the domain “Cricket” has the single lowest average conceptual coverage of all domains that met our analysis’ parameters. This is not an effect of small amounts of data; the “Cricket” domain has over 15,000 concepts. Sports overall are a prominent theme in Table 3.7-d, which introduces tennis and skiing as two quite globally known topics. One quite curious finding in Table 3.7-d is the low conceptual coverage of fencing, a sport not typically associated with the English-speaking world. We initially believe this to be the result of an error in WordNet domain assignment, but we found that, indeed, there are a very large number of English-only articles about fencers. However, as we will see in the sub-concept-level diversity study that follows, the story of these articles is quite complex.

Domain assignment errors are also suggested by some of the example concepts in Table 3.7-d. However, examining many of these potential errors, we frequently found they were issues related to context rather than true mistakes. For instance, Hamburger SV is most well-known for its soccer team, but the club is also the home of a baseball team (the Hamburger Stealers). Other times, however, there were actual errors. “Nazca Plate” (English), for instance, is about a geologic feature rather than a baseball-related subject.

### **3.7.3 Sub-concept-level Diversity by Topic**

We now turn our attention to the relationship between the sub-concept-level diversity present in a concept and that concept’s topic memberships. To better understand this relationship, we executed an analysis that draws from both the concept-level work above and the work we did with sub-concept-level diversity and centrality in Section 3.6.

For this analysis, we first sampled concepts belonging to each synset topic by randomly

sorting each synset’s concepts, selecting the first 25, and sampling every tenth concept after the first 25. We then calculated the mean *RatioInEnglish* metric for each synset’s sample, skipping any sampled concept that only has an article in English. Recall that *RatioInEnglish* measures the share of the content about a concept available in all of multilingual Wikipedia that is in that concept’s English article (if it has one). *RatioInEnglish* was calculated in the same fashion as *RatioInLang* was in Section 3.6. That is, it was done using the lower-bound wikification strategy as a proxy for the upper-bound, with which it is closely correlated.

After executing our sampling procedure and calculating *RatioInEnglish* for all concepts selected by this procedure, we had an estimate of the extent to which the English articles in a given synset describe all the information about that synset in multilingual Wikipedia. Table 3.7-e shows the synset topics with the highest percentage of content in English and the lowest percentage of content in English among synsets with 25 or more concepts. Immediately obvious in the table is that many of the synsets with the greatest amount of information in English also have the lowest average conceptual coverage. That is, the left side of Table 3.7-e is replete with topics we observed in the previous subsection when discussing topics that have the fewest articles outside of the English Wikipedia. The interpretation of this phenomenon is quite clear: for concepts in English-speaker focused synsets like “Fraternity,” “Quarterback,” and “Halfback,” even when a concept does have an article in a language edition other than English, that article tends to have very little information that is not already in the English article. Indeed, the English dominance of some of these synsets is so extreme that the *minimum* amount of content not available in the English article is over 80 percent. That is, even without any wikification, the English language edition covers over 80 percent of the content in the non-English articles about these topics. Performing any wikification at all is very likely to

significantly increase this number.

The right side of Table 3.7-e reveals that the inverse of the phenomenon we saw with “Fraternity” and “Quarterback” also occurs. That is, there are many cases of English articles having a relatively small percentage of the information about topics with high average conceptual coverage. For instance, there are many skiing-related synsets on the right side of Table 3.7-e, with skiing being a very high conceptual coverage domain. Similarly, the “National Flag” synset has the second-highest average conceptual coverage and appears on the list of synsets with the lowest average *RatioInEnglish* metric.

Overall, we found that the Spearman’s correlation between average conceptual coverage and *RatioInEnglish* is a relatively high -0.478 ( $p < 0.0001$ ). Roughly speaking, we can use this correlation coefficient to characterize the relationship between concept-level and sub-concept-level diversity among topics. When concepts in a topic are covered by many language editions, the share of sub-concept information about that topic in a single language edition (in this case English) goes down. Conversely, when a concept is covered in just a few language editions, there is less diverse information about that concept across the language editions overall.

<b>Highest RatioInEnglish</b>		<b>Lowest RatioInEnglish</b>	
<b>Topic</b>	<b>RatioInEng</b> (Lower Bound)	<b>Topic</b>	<b>RatioInEng</b> (Lower Bound)
School district	0.889	Junior college	0.091
Public school	0.861	Fixed charge	0.124
Preparatory school	0.858	Fee	0.130
Inauguration	0.846	Retainer	0.136
Caste	0.845	Shogun	0.160
Police	0.823	Ski jumper	0.164
Halfback	0.821	Vicariate	0.167
Bartender	0.821	Commune	0.192
Gastropod	0.818	Snowboarder	0.217
Fraternity	0.815	Kibbutz	0.220
Outline	0.810	Threadfin	0.227
Suburb	0.809	Collective farm	0.230
Residential district	0.805	Tetra	0.230
Air base	0.804	Pika	0.236
Byway	0.802	Barbu	0.238
Newsreader	0.796	College	0.240
Newscaster	0.795	Bullfighter	0.240
Coordinator	0.794	National flag	0.243
Cub	0.792	Herbicide	0.245
Constituency	0.791	Characin	0.245
Joint venture	0.788	Regency	0.247
Public house	0.785	Adventure story	0.248
Quarterback	0.783	Skier	0.249
Screenplay	0.782	Surveying instrument	0.249
Mollusk	0.779	Diocese	0.251
Tavern	0.776	Jurisdiction	0.251
Ghat	0.774	Bony fish	0.252
Linguistic process	0.772	Cypriniform fish	0.253
Professorship	0.771	Teleost fish	0.254
Cornerback	0.768	Pteridologist	0.254
Tight end	0.762	Avenue	0.255

Table 3.7-e: Topic synsets with the highest and lowest average RatioInEnglish metrics.

Examining the outliers of the concept-level/sub-concept-level diversity relationship revealed several interesting special cases. High *RatioInEnglish*/high conceptual coverage outliers include synsets like “State capital,” which is disproportionately made up of U.S. state capitals (“Provincial capital” is a separate synset). We saw in Section 3.5 that concepts describing places in English-speaking regions tended to have the highest *RatioInEnglish* values. Given that administrative district capitals tend to have high conceptual coverage (Table 3.7-a) and given that the particular capitals in question are mostly U.S. capitals, the outlier status of this synset makes sense.

The low *RatioInEnglish*/low conceptual coverage outliers were also worthy of examination. One of these outliers is the synset “Fencer,” by far the largest synset in the “Fencing” domain. We noted above that the “Fencing” domain’s low conceptual coverage is quite odd given the fact that Fencing is not known to be a sport dominated by English speakers like cricket or American football. The “Fencer” synset’s status as a low *RatioInEnglish*/low conceptual coverage outlier provides an explanation for this peculiar result. The fact that the average conceptual coverage and the average *RatioInEnglish* are quite low for “Fencer” suggests that there are many English-only articles about fencers, but that they are very short articles. As such, when another language edition covers a fencer, that language edition’s article is likely to have significantly more information than its English counterpart.

A survey of English articles about fencers revealed a great deal of support for this interpretation. For instance, “Giuseppe Delfino” (English), a short article about an Italian fencer whose counterparts in other language editions are much more detailed, is a contributor to the low *RatioInEnglish* value. On the other hand, the very short English-only article “Knut Enell” (English) about a Swedish fencer contributes to the low conceptual coverage. Investigating the

articles about fencers in more detail, we found that they had mostly been created by a single user and were quite formulaic in nature, suggesting automation was involved.

Finally, just as we did with concept-level diversity, we also examined the relationship between sub-concept-level diversity and topic using domains in addition to synsets. As before, the results reinforced our earlier findings. “Cricket” is the domain with the highest average *RatioInEnglish*, “Baseball” has the sixth highest, and so on. On the other side of the distribution, “Skiing” has the lowest average *RatioInEnglish*, with “Chess” and “Tennis” also in the bottom five. Indeed, the domain rankings for sub-concept-level diversity looked a lot like those for concept-level diversity. At -0.583, the Spearman’s correlation coefficient between mean *RatioInEnglish* and mean conceptual coverage was even stronger for domains than for synsets.

Our sub-concept-level analysis of domains also revealed an important finding that problematizes a very common practice in the multilingual Wikipedia literature. As noted above, a large percentage of the research projects in this area use biographies as a supposedly representative sample of all articles in an entire language edition. Examining our *RatioInEnglish* domain results, we noticed that the one-million-concept “Person” domain had a low mean value relative to other large-scale domains. We mentioned at the beginning of this section that very large domains tend to be rather noisy, but a survey of a sizable number of concepts that were assigned to the “Person” domain revealed this not to be so in this case.

Using a two-sample t-test, we established that biographies do indeed have a significantly lower mean *RatioInEnglish* value than articles covering concepts from other domains ( $t(13164.88) = 16.11, p < 0.001$ ). Moreover, examining the results from our concept-level domain study, we found that the “Person” domain also has a relatively low mean conceptual coverage compared to other large domains. Biographical articles in the English Wikipedia belong

to concepts that are on average in 2.84 language editions. This is less than the English Wikipedia-wide average conceptual coverage, which is 2.91. Although our concept-level experiment data collection process does not afford formal analyses of this difference's significance, it is highly likely to be significant due to the enormous sample sizes involved.

Put together, these results strongly suggest that biographical articles are not representative of all articles in a language edition when it comes to encyclopedic world knowledge diversity. The implication of these results is that multilingual Wikipedia researchers should end the practice of focusing on biographies unless their conclusions are limited to this domain. Instead, studies should use all articles and associated resources from each language edition or, where this is not possible, employ a random sampling procedure.

### **3.7.4 Discussion**

In this section, we demonstrated that the amount of concept- and sub-concept-level diversity in multilingual Wikipedia varies extensively from topic to topic. We also showed that this variation can often be explained using cultural context as a framework. At the concept-level, this means that topics that are parochial to the English-speaking world (e.g. fraternities, American football) and instances of more general topics that occupy a similar cultural position (e.g. various American air strips, various American weatherpeople) tend to not have many articles in non-English language editions. At the sub-concept-level these cultural effects get compounded, with the few articles in other language editions about topics like sororities and American weatherpeople having very little information that is not already in the English article.

Due to the English-as-Superset assumption built into the YAGO2s dataset, our conclusions from this section are limited to concepts that have an English article. It is possible that the

relationships between topics and diversity are quite different for concepts that do not meet this requirement. These differences could occur at the high level – e.g. a much greater rate of instance-parochialism – or, much more likely, at the level of individual topics themselves. For instance, while we found that royalty that are covered in English tend to be covered in many language editions, the opposite could be true for royalty that do not have English articles. In order to investigate the extent to which these divergences occur, it will be first necessary to develop a sufficiently accurate multilingual topic assignment algorithm. We view the development of such an algorithm to be an important area of future work.

### **3.8 Diversity in the Consumption of Content**

In the previous sections, we focused on the cultural contextualization inherent to the output of peer production processes when peers are clustered into cultural groups. In this section, we turn our attention to the language-defined cultural differences in the *consumption* of that output. More specifically, rather than using a cultural lens to examine the properties of Wikipedia content, we do the same with Wikipedia *page views*. To the best of our knowledge, this series of studies represents the first formal investigation of the similarities and differences in the consumption of content across the language editions of Wikipedia.

The first goal of this section is to characterize the amount of content consumption diversity in multilingual Wikipedia. Specifically, we investigate whether there is substantial variation in the content people access from language edition to language edition. Our second goal is to use page view information to better understand concept-level and sub-concept-level diversity. Here we ask questions such as, “What percentage of page views go to single-language concepts? What percentage go to global concepts?” We also investigate the extent to which concepts with large

numbers of page views have more or less sub-concept-level diversity than those that are less commonly accessed. Finally, we close this section with a comparison of page view diversity with that of centrality diversity. Our focus here is on whether the content in multilingual Wikipedia reflects the cultural interests of its readers or whether it exaggerates or mutes the diversity in these interests. Before investigating any of these issues, however, we must first discuss the basics of our page view methods, focusing on our data collection and aggregation process.

### **3.8.1 Content Consumption Diversity Methods**

The source of our page views data is a publicly available raw dump of the Wikipedia page view logs made available by the Wikimedia Foundation<sup>41</sup>. The logs report hour-by-hour page view information dating back to December 2007 for every page in multilingual Wikipedia, including redirect pages, disambiguation pages, categories, user pages, and so on. Because this information is formatted in plaintext, the logs are many terabytes in size. As such, we built into WikAPIdia a page view processor that streams the raw logs from the Wikimedia Foundation, aggregating all page views into daily totals. These totals were then aggregated on a month-by-month basis, which is the information used in our studies. We restrict our focus to page views that occurred from January 1, 2010 to December 21, 2012, the final date of our data collection process. Immediate future work involves incorporating the available data prior to 2010 and from December 21 through December 31, 2012.

It is important to note that the page view data can be somewhat noisy. Page views have the inherent disadvantage relative to the unique visitors metric in that they are subject to enormous outliers, usually the result of automated processes. For instance, the second most-viewed article in the English Wikipedia is the article “23” (English), a statistic that is clearly not the product of

---

<sup>41</sup> <http://dumps.wikimedia.org/other/pagecounts-raw/>

“natural” traffic to Wikipedia. However, the signal in the page view data is a strong one and Wikipedia page views have been used to determine the pharmaceutical drugs about which patients most frequently seek information [113], to provide context for a well-known study on the relationship between the quality of an article and the number editors it has [104], and to train accurate models of English Wikipedia article popularity [196], among other applications. Moreover, some of the statistics that may seem like noise may be the product of a concept being featured on the main page of a language edition, becoming a featured article, and similar phenomena. Of course, one near guarantee [83] of the strong signal in our page view data is that, as we will show below, Justin Bieber is a very popular concept in many language editions.

### 3.8.2 Basic Content Consumption Diversity

We now turn our attention to using page view data to examine diversity in content consumption in multilingual Wikipedia. In this section, we adopt methods similar to those in the centrality study in Section 3.6.2, focusing on comparing the top  $n$  concepts in each language edition. Recall that the top  $n$  centrality study involved comparing the set of 100, 1,000, and 10,000 most-central concepts in each language edition. Here we do same, using page views aggregated over the entire 2010-2012 dataset instead of centrality.

Table 3.8-a shows the aggregate statistics for each of these page view top  $n$  comparisons. It

PAGE VIEW TOP-N SET OVERLAP				
<b>Set</b>	<b>Mean</b>	<b>Stdev</b>	<b>Min</b>	<b>Max</b>
Top 100	0.214	0.111	0.020 uk/zh	0.490 no/sv
Top 1,000	0.308	0.106	0.047 ja/uk	0.535 es/pt
Top 10,000	0.354	0.089	0.193 ja/uk	0.577 es/pt

Table 3.8-a: Pairwise agreement of the n-most viewed concepts between language editions.

is clear from the table that there is extensive diversity in the  $n$  most-viewed concepts in multilingual Wikipedia. On average, two language editions share only 21.4% of their 100 most-viewed pages, 30.8% of the 1,000 most-viewed pages, and 35.4% of the 10,000 most-viewed pages. We see again in Table 3.8-a that most of the superlative similarities occur between the language editions of closely associated language-defined cultures such those rooted in Scandinavia and Portuguese/Spanish.

The large role cultural context plays in causing page view diversity becomes clear when examining the  $n$  most-viewed concepts in non-aggregate form. Table 3.8-b shows the top 10 most-viewed concepts in twelve language editions. Consider the left-most column, which shows the top 10 for the Catalan, Hebrew, and Spanish Wikipedias. In all three cases, the majority of the displayed concepts have an obvious cultural component. In Catalan, we see that (1) places in which Catalan speakers live are featured prominently and (2) two of the top 10 most-visited concepts are so isolated to the cultural context of Catalan speakers that they do not have an English Wikipedia article. “Regles d'accentuació del català” (Catalan) is an article about Catalan accent usage. Mario Conde is a notorious former bank CEO in Spain. Despite the fact that he has been called “Spain’s Machiavelli” [142], there is no article about him in English (although there is in Spanish, Galician, and Portuguese). Even more culturally contextualized than the Catalan Wikipedia is the Hebrew Wikipedia, whose top 10 list could be mistaken for a list of concepts most important to Israeli culture and history. David Ben-Gurion, Theodor Herzl, and Yizhak Rabin are all included, as is the country of Israel, the United States, and the city with the largest Jewish population in the United States: New York. Moving on to Spanish, note that Mexico, not Spain, is the most-viewed concept in the Spanish Wikipedia. Other top 10 Spanish concepts include the Maya civilization, another Latin America-related concept, as well as concepts less

unique to the cultural context of Spanish speakers (e.g. Justin Bieber).

At a higher level, there are several clear themes in Table 3.8-b. As we saw with content metrics like centrality, home cultural regions are also prominent in patterns of content consumption. That said, popular culture is significantly more present in the table below than it is in the equivalent table for centrality (Table 3.6-b). The Japanese Wikipedia's top ten, for example, is predominantly made up of anime/manga and girl bands. Another theme in Table 3.8-b is the popularity of adult-themed concepts. In the Japanese Wikipedia, one of these concepts is the most-viewed concept in the entire language edition, and in the Finnish Wikipedia an adult concept is the second-most-viewed.

Social networks and other Web 2.0 sites like YouTube are also quite present in Table 3.8-b. The social network that is in the top 10 is contextualized for each language-defined culture. Odnoklassniki is a popular Russian social network and appears in the Russian Wikipedia's top 10, while Facebook does not. A similar phenomenon occurs with Orkut and the Portuguese Wikipedia.

Lastly, it is important to note that while several of the concepts in Table 3.8-b most likely represent noise in the page view dataset, this noise is unrepresentative of at least the top 100 most-viewed articles. For instance, in the English Wikipedia, the only concepts whose appearance in the top 100 is most likely due to noise (as determined anecdotally) appear in the top ten. This is not surprising given some of the likely causes of the noise, such as various automated processes.

Catalan	Czech*	English*	German
Catalonia	Wiki	Time <sup>†</sup>	Germany
Barcelona	Czech Republic	23 <sup>†</sup>	Cul-de-sac <sup>†</sup>
Middle Ages	Prague	Wiki	How I Met Your Mother <sup>†</sup>
Wikipedia	Germany	Facebook	Two and a Half Men <sup>†</sup>
Catalan language	European Union	Sitemaps <sup>†</sup>	The Big Bang Theory <sup>†</sup>
Spain	World War II	United States	Wikipedia
Physical education	United States	Google	Facebook
<i>Regles d'accentuació del català</i> "Stress rules of Catalan"	List of historical anniversaries	YouTube	Berlin
<i>Mario Conde</i> "Mario Conde"	Europe	Justin Bieber	Hamburg
Catalan Wikipedia	Facebook	Glee (TV series)	United States
Hebrew	Japanese	Portuguese	Russian
Israel	Av 女優一覧 "List of Pornographic Actors"	Orkut	Russia
United States	AKB48	Webserver directory index <sup>†</sup>	Wikipedia
Facebook	One Piece	Bullying	Odnoklassniki
Yitzhak Rabin	Wikipedia	Google	250 лучших фильмов по версии IMDb "250 best movies according to IMDb"
Jerusalem	海賊 (ONE PIECE) "Pirate (One Piece)"	Volleyball	YouTube
Wikipedia	Neon Genesis Evangelion (anime)	Justin Bieber	Moscow
Korban	A Certain Magical Index	Portugal	Human sexual activity
New York City	Girls' Generation	United States	Ukraine
David Ben-Gurion	Arashi	World War II	Bittorrent
Theodor Herzl	List of One Piece characters	Industrial Revolution	Harry Potter (character)
Spanish	Korean	French	Finnish
Mexico	South Korea	Facebook	Wiki
@ (At sign) <sup>†</sup>	Naneun Ggomsuda	Wiki	Alastonsuomi.com "Nude Finland" (website)
Spain	Masturbation	France	Justin Bieber

Spanish	Korean	French	Finnish
Maya civilization	Animal female reproductive system	Paris	Human sexual activity
World War II	Girls' Generation	YouTube	Salatut elämät
Science	World War II	United States	Helsinki
Google	중 2 병 “Middle School 2nd Year Syndrome”	Google	German
Justin Bieber	Tsundere	Victor Hugo	James Bond
Water	Park Chung-hee	List of One Piece episodes	Halloween
United States	Japan	Justin Bieber	Facebook

*Table 3.8-b: The English titles of the 10 most-viewed concepts in 12 language editions of Wikipedia. Concepts that do not exist in the English Wikipedia are shown with their native language edition title below which an English translation from Google Translate (or Urban Dictionary, where necessary) is provided in gray text. While there is clearly some noise present in the page view dataset, the cultural context of each language edition is visible in the table, which is replete with home countries and cities, as well as local pop culture and history. The Hebrew Wikipedia is a particularly good example.*

<sup>†</sup> Indicates that a single investigator determined that noise almost certainly caused this concept to be in the top 10. The investigator used Internet search engines and consultation with members of the language-defined culture in question to make this determination.

\* Indicates the “Main Page” article has been removed from the top 10 in this language edition. JWPL’s parsing engine – the core of WikAPIdia’s parsing approach (Section 3.12) – only identified a “Main Page” for a subset of language editions, possibly due to variation in its namespace membership or other technical concerns. (Note that the final product of the parsing process passed a number of robust evaluations described in Section 3.2.3).

### 3.8.3 Content-level Diversity and Content Consumption

Above, we saw that there is a great deal of diversity in the concepts viewed by the audiences of each language edition. Our goal in this section is to understand how this diversity interacts with concept-level diversity. If we find that very few page views go to single-language concepts and that most of the content that is consumed in multilingual Wikipedia describes global or near-global concepts, this would decrease the significance somewhat of the extensive concept-level diversity we identified in Section 3.4 and would do the same for the corresponding cultural contextualization. If, on the other hand, we find that a sizable number of page views go to single-language or near-single language concepts, this would suggest that the content that is contextualized for each language-defined community plays an important role in multilingual Wikipedia. In this scenario, single-language concepts would not be esoteric curiosities, but rather concepts that are important to the needs of people seeking encyclopedic world knowledge.

To determine for which of these scenarios there is more support, we first grouped all 8.67 million concepts by conceptual coverage level. We then summed the page views over the entire 2010 – 2012 period for each group. For instance, we added up all page views from 2010 – 2012 for all articles that are single-language concepts, did the same for articles that are part of two-language concepts, and so on. Our goal here was to understand the percentage of page views that go to single-language concepts versus two-language concepts, versus three-language concepts, and so on.

Figure 3.8-a, which shows the share of total page views that went to each level of conceptual coverage, provides significantly more evidence for the second scenario above than the first. 16.00% of page views go to single-language concepts while 16.96% of concepts go to

global concepts. In other words, when it comes to content consumption, single-language concepts are nearly as important as global concepts. While there are many, many fewer global concepts than single-language concepts and global concepts receive far more page views on a per-concept basis (Figure 3.8-b), Figure 3.8-a reveals that, in aggregate, single-language concepts rival global concepts in terms of Wikipedia reader interest. Flipping our analysis around, Figure 3.8-a also demonstrates that over 83% of page views go to concepts that are missing from at least one language edition.

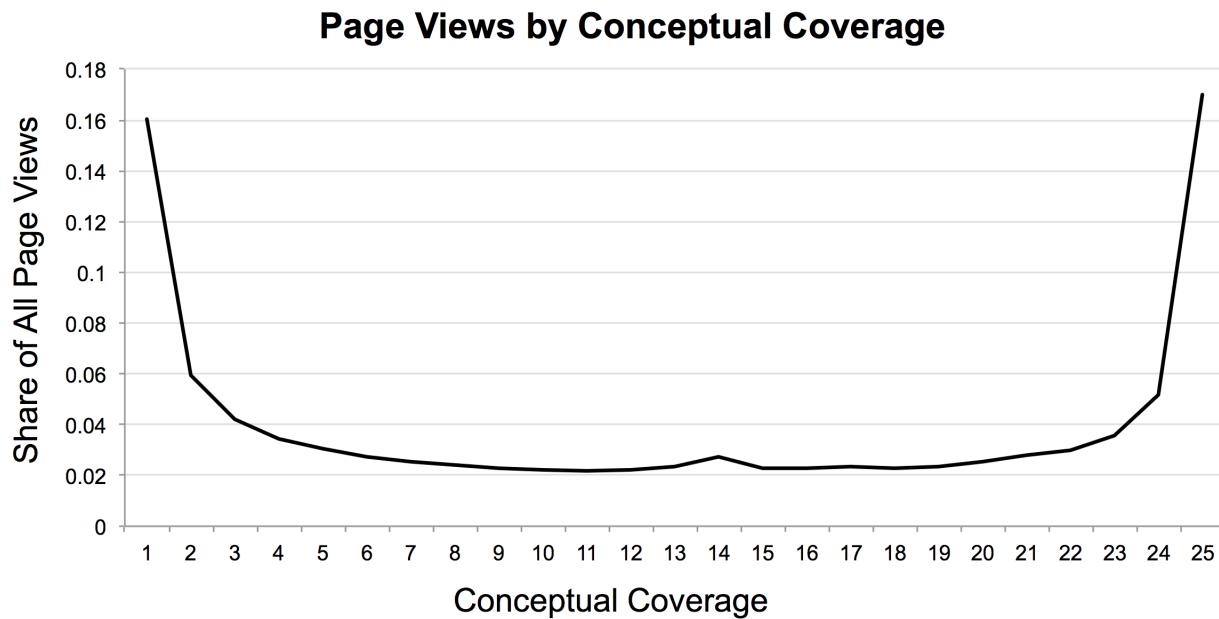


Figure 3.8-a: The share of page views that go to each level of conceptual coverage. About 16% of page views go to all single-language concepts, while approximately 17% go to the much smaller number of global concepts.

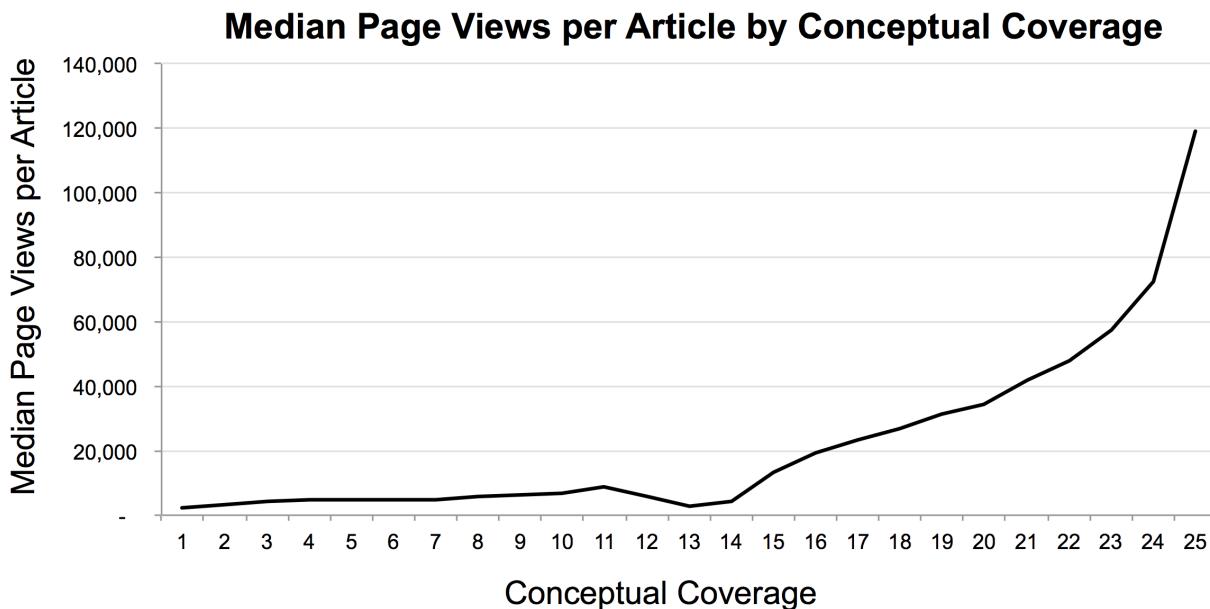


Figure 3.8-b: The median number of page views that go to articles belonging to concepts at each level of conceptual coverage. Global concept articles receive far more page views than single-language concept articles, but there are far more single-language concepts than global concepts (see Section 3.4), causing the results in Figure 3.8-a.

<b>Language</b>	<b>% Single-language</b>	<b>% Non-global</b>	<b>% Global</b>
Czech	8.16%	56.06%	43.94%
Spanish	8.52%	71.50%	28.50%
German	13.54%	82.42%	17.58%
English	16.21%	85.81%	14.18%
Norwegian	15.23%	75.68%	24.32%
Indonesian	21.60%	70.21%	29.79%
Japanese	29.12%	90.97%	9.03%

*Table 3.8-c: The percentage of page views that go to single-language, non-global, and global concepts in a representative selection of language editions.*

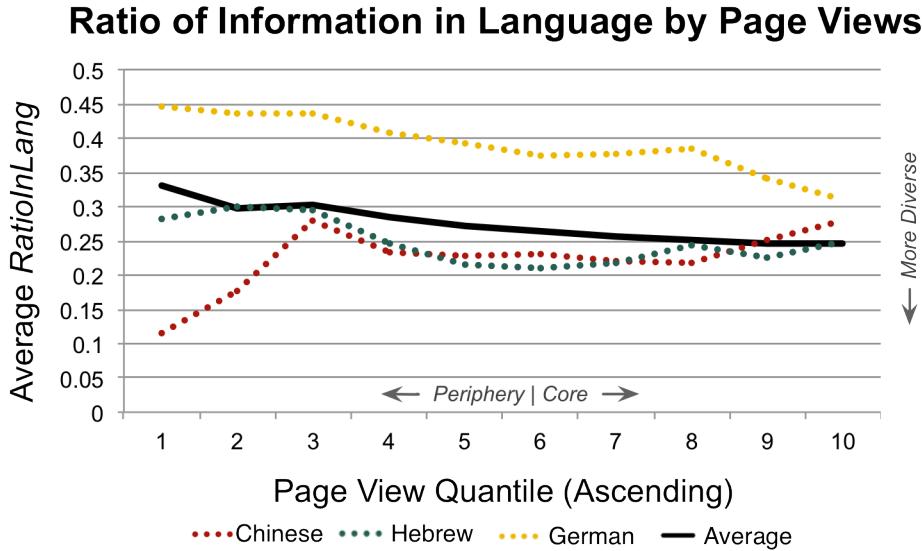
In order to test the robustness of the findings in Figure 3.8-a to the set of language editions considered, we also ran an identical study, leaving out the English Wikipedia and its large number of single-language concepts. The results of this study were nearly exactly the same as the first: 18.44% of page views went to single-language concepts while 19.91% went to (24-language) global concepts.

Examining these findings on a language-by-language basis, it is immediately clear that the percentage of page views that go to single-language versus global concepts is not correlated with the number of articles in each language edition. Table 3.8-c shows the share of page views by conceptual coverage in a representative selection of language editions. Here, we see that Japanese, the ninth-largest language edition in terms of number of articles, has by far the largest share of page views that go to single-language concepts (29%). Moreover, some of the smallest language editions we consider such as Korean and Indonesian have single-language concept shares above 20%. In other cases, we see larger language editions having a relatively small percentage of views go to single-language concepts. The second-smallest such percentage belonged to the Spanish Wikipedia (8.5%).

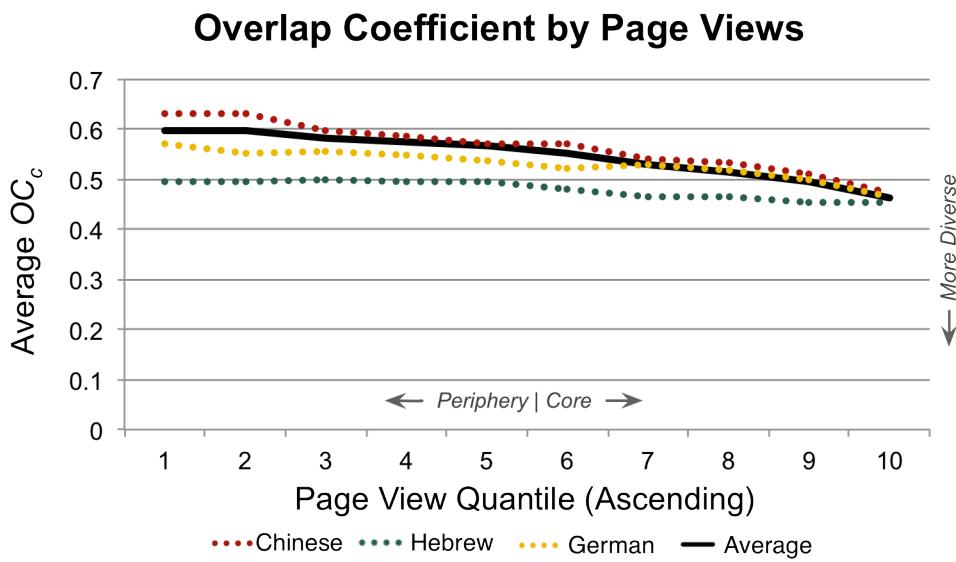
Interestingly, the share of page views going to single-language concepts and the percentage of a language edition that is made up of those concepts were not strongly correlated; the Spearman's correlation between these two variables was only 0.35. English was a significant outlier, with it having the highest percentage of single-language concepts but only the seventh-highest percentage of page views going to those concepts. Slovak was the largest outlier in the other direction. The Slovak language edition has the second-lowest percentage of single-language concepts, but it has the sixth-highest percentage of single-language page views.

### **3.8.4 Sub-concept-level Diversity and Content Consumption**

In the preceding section, we inquired as to whether concept-level diversity “matters” to Wikipedia readers and found that the overwhelming answer is “yes”: single-language concepts as a group get almost as many page views as global concepts, and 83% of page views go to concepts that do not appear in all 25 language edition editions. In this section, we ask the analogous question for sub-concept-level diversity: does sub-concept-level diversity “matter?” Specifically, our research question here is, “Do the vast majority of page views go to concepts for which there is a global consensus on the definition or do Wikipedia readers frequently access articles about concepts that are described differently in each language edition?” If the former is true, this would serve to decrease the importance of sub-concept-level diversity in multilingual Wikipedia. If the latter is true, it would mean that not only does each language edition describe each concept differently, but also that the readers of each language are actually consuming different information about each concept and are doing so quite often.



*Figure 3.8-c: The share of information about a concept available in a language edition by the number of times that concept is viewed in the language edition. More popular concepts have a lower share of information on average, meaning that sub-concept-level diversity is highest among the most popular concepts.*



*Figure 3.8-d: The average amount of a shorter article's content available in a longer same-concept article by concept page views. Shorter articles have more unique information in more popular concepts.*

Due to the computational and I/O demands of our wikification algorithm, our methods for sub-concept-level diversity diverge somewhat from those we leveraged in the previous section. Here we adopt the approach we used to investigate the relationship between centrality and sub-concept-level diversity in Section 3.6. Specifically, we divide up each language edition's concepts into quantiles and calculate the average heuristic  $RatioInLang$  and  $OC_c$  metrics for each quantile. However, whereas in Section 3.6 we used quantiles based on PageRank scores, here we use quantiles based on page views.

The results of the page view  $RatioInLang$  analysis using a sample of 1,000 concepts in each of 10 quantiles can be found in Figure 3.8-c. Averaged across all language editions, the  $RatioInLang$  metric nearly monotonically decreases from the lowest page view quantile to the highest page view quantile. This means that in a given language edition, the most popular articles are those that have *smallest* share of the information about their corresponding concepts. In other words, sub-concept-level diversity is highest in the most-viewed concepts.

The results for our  $OC_c$  metric analysis can be found in Figure 3.8-d. Here again, we see that sub-concept-level diversity is highest among the concepts with the greatest popularity and lowest among the concepts with the least popularity. More specifically, the amount of unique content in shorter articles relative to longer same-concept articles is greatest in the most-popular concepts and smallest in the least-popular ones.

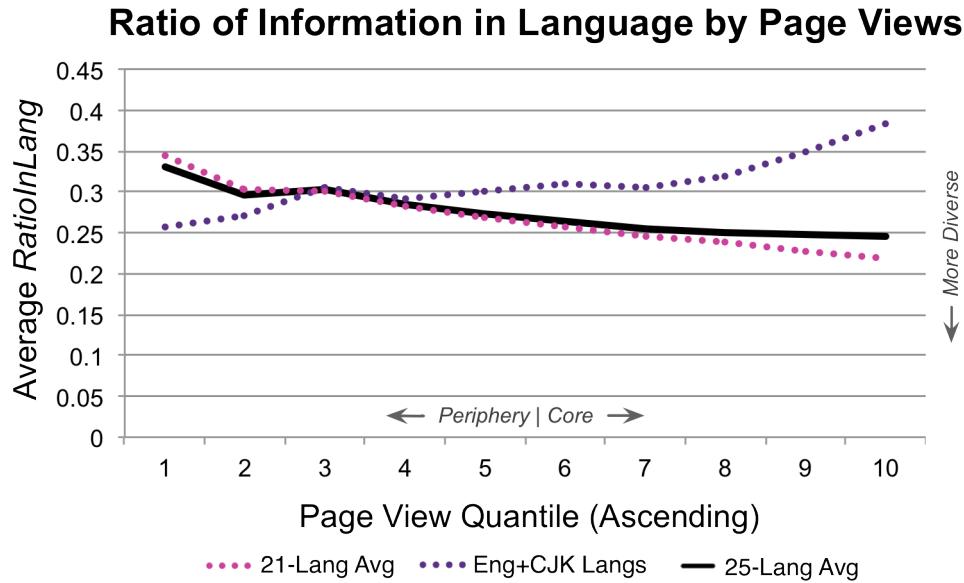


Figure 3.8-e: RatioInLang for the four-language English + Chinese + Japanese + Korean set of languages vs. that of the other 21 languages and all 25 languages.

As was the case with centrality, the  $OC_c$  results were more consistent across language editions than the *RatioInLang* results. English and the three East Asian languages again bucked the trend of the rest of the language editions with regard to *RatioInLang*. In fact, for these four language editions, the trend is the opposite: the most-viewed concepts are those that have the *highest* share of information in the language edition's articles about them (Figure 3.8-e). As we noted above, exploring the unique behavior of English, Chinese, Korean, and Japanese with regard to *RatioInLang* is an important area of future work.

### 3.8.5 Discussion: Page Views vs. Centrality

In Section 3.6, we discussed how centrality measures are one way of assessing the importance of an article in a language edition. Here we argue that page views are another way of doing the same thing. Whereas centrality-based importance is derived purely from the content – specifically the article graph structure – of each language edition, page view-based importance

is assigned *en masse* by the readers of each language edition.

Given that centrality and page views can both be considered importance metrics, it is striking, then, to consider the divergent nature of their output. Comparing Table 3.6-b and Table 3.8-b, it is clear that while adult topics are quite important to readers, the article graphs do not reflect this. Similarly, there are no popular culture-related concepts in Table 3.6-b, but Table 3.8-b is replete with them.

Even more informative than comparing centrality-based importance and page view-based importance for individual concepts is doing the same for entire language editions. In particular, juxtaposing the results in this section with those in Section 3.6, it is clear that there is a great deal more diversity between the language editions according to page view-based importance than there is according to centrality-based importance. For instance, at the beginning of this section, we showed that the average overlap between any two language editions' set of 100 most-viewed concepts is 21.4%. In Section 3.6, we found that the equivalent number for most-central concepts is well over 50% for both PageRank and indegree centrality. Moreover, the page view top  $n$  concepts reflect more diversity when  $n = 1,000$  and  $10,000$  as well.

These findings suggest that the content of multilingual Wikipedia, diverse as it is, *mutes* the diversity of its readers' interests. While the content of the Catalan Wikipedia suggests that France and the United States are the two most important concepts in all of world knowledge, the readers of the Catalan Wikipedia suggest that Catalonia and Barcelona are. While the content of the Japanese Wikipedia puts the United States as the fourth-most important concept, the readers of the Japanese Wikipedia do not rank the United States in the top 10. While the Spanish Wikipedia's content indicates that Spain is its most important concept, the readers of the Spanish Wikipedia imply with their behavior that Mexico is more important. Indeed, a list of similar

examples can go on and on and on.

The tension between the less diverse content of Wikipedia and the more diverse nature of its readers' interests has important implications, in particular for the future directions of Wikipedia. This topic is a major focus of in Section 3.11, and we defer detailed discussion until then. However, for the remaining sections of this thesis – especially Section 3.10 – this tension is important to keep in mind.

### 3.9 Diversity over Time

Throughout this chapter, we have seen that there is a great deal of diversity across the language editions of Wikipedia both in terms of what concepts are covered in each language edition as well as how those concepts are covered. We have made these conclusions based on a static snapshot of 25 language editions. Multilingual Wikipedia, however, is far from static. Each language edition is edited on a second-by-second basis. This raises an important question: as the language editions grow and change, are they growing and changing toward one another or away from one another? That is, is the diversity in multilingual Wikipedia decreasing over time, as the global consensus hypothesis might suggest, or is it conforming to the global diversity hypothesis and remaining constant or increasing?

In this section, we compare concept-level and sub-concept-level diversity across two database dumps of the 25 language editions considered here: one from September 2009 and one

Statistic	2009 (millions)	2012 (millions)	Increase (%)
Number of Articles	11.07	17.88	61.5
Number of Parseable Links	304.03	517.64	70.3
Number of Concepts	5.81	8.67	49.0

Table 3.9-a: Growth from 2009 to 2012 in our 25-language multilingual Wikipedia dataset.

from October/November 2012, which we have been using for most of this chapter. Our primary approach is to repeat experiments we executed in our CHI 2010 paper on Wikipedia diversity [82], which used the 2009 data, but do so using the data from 2012.

As shown in Table 3.9-a, the 2009 dump had a total of just over 11 million articles and 304 million parseable links. By the time of the 2012 dumps, the 25 language editions had almost 17.9 million articles and more than 517 parseable million links. In other words, the number of articles grew by 61.5% and the number of links in those article grew by even more: 70.3%.

Below, we first look at how this growth has played out in terms of concept-level diversity across language editions. That is, we evaluate the differences in the conceptual coverage distribution over time. Following that, we examine sub-concept-level diversity over time by investigating the extent to which same-concept articles are more similar now than they were in 2009.

### **3.9.1 Concept-level Diversity**

The concept-level global consensus hypothesis, applied in this temporal context, would suggest that the language editions are growing largely due to the translation of articles, in particular from the English language edition to the other language editions. In other words, the global consensus hypothesis would posit that the almost seven million new articles seen in Table 3.9-a will have shifted the conceptual coverage distribution further to the right, with the 2009 distribution having significantly more single-language concepts and fewer global concepts, relatively speaking. On the other hand, another possibility, tantamount to the temporal global diversity hypothesis, is that the new articles are creating more diversity, or at least following the same high-diversity conceptual coverage distribution as the articles that preceded them.

In addition to showing the growth in articles and links, Table 3.9-a also depicts the growth in concepts from 2009 to 2012. While articles grew by 61.5%, concepts grew by a substantial amount as well: 49.0%. This is a clear initial indication that a large percentage of the new articles are about new concepts that were not described in 2009’s multilingual Wikipedia. A more in-depth analysis, however, is needed in order to understand in detail how new articles affect the conceptual coverage character of multilingual Wikipedia. This analysis is the focus of the remainder of this subsection.

In our CHI 2010 paper, we performed a study almost identical to that in Section 3.4: we calculated the conceptual coverage of each concept in the 2009 25-language dataset and examined the conceptual coverage distribution across multilingual Wikipedia. There is one significant difference between the two studies, however: the concept identification algorithm we used in 2009 was less sophisticated than *Conceptualign*. Our previous algorithm did, like *Conceptualign*, identify all connected components in the interlanguage link graph. However, it did no “splitting” in cases of interlanguage link conflicts. That is, it aggregated into the same concepts the English articles “Rain gutter,” and “Canal,” “High school,” and “Secondary school”, and so on.

While the older algorithm results in less precise concepts, it was necessary to recreate the algorithm and apply it to our 2012 dataset in order to compare concept-level diversity over time. Because we constructed WikAPIdia such that it can dynamically switch between concept alignment algorithms (so as to facilitate concept alignment research), this process was a simple one. As noted above, interlanguage link conflicts are not the norm (although they tend to occur in more “global” and more significant concepts), so applying the CHI 2010 algorithm to our 2012 data did not result in a massive drop-off in the number of concepts: it output 8.41 million

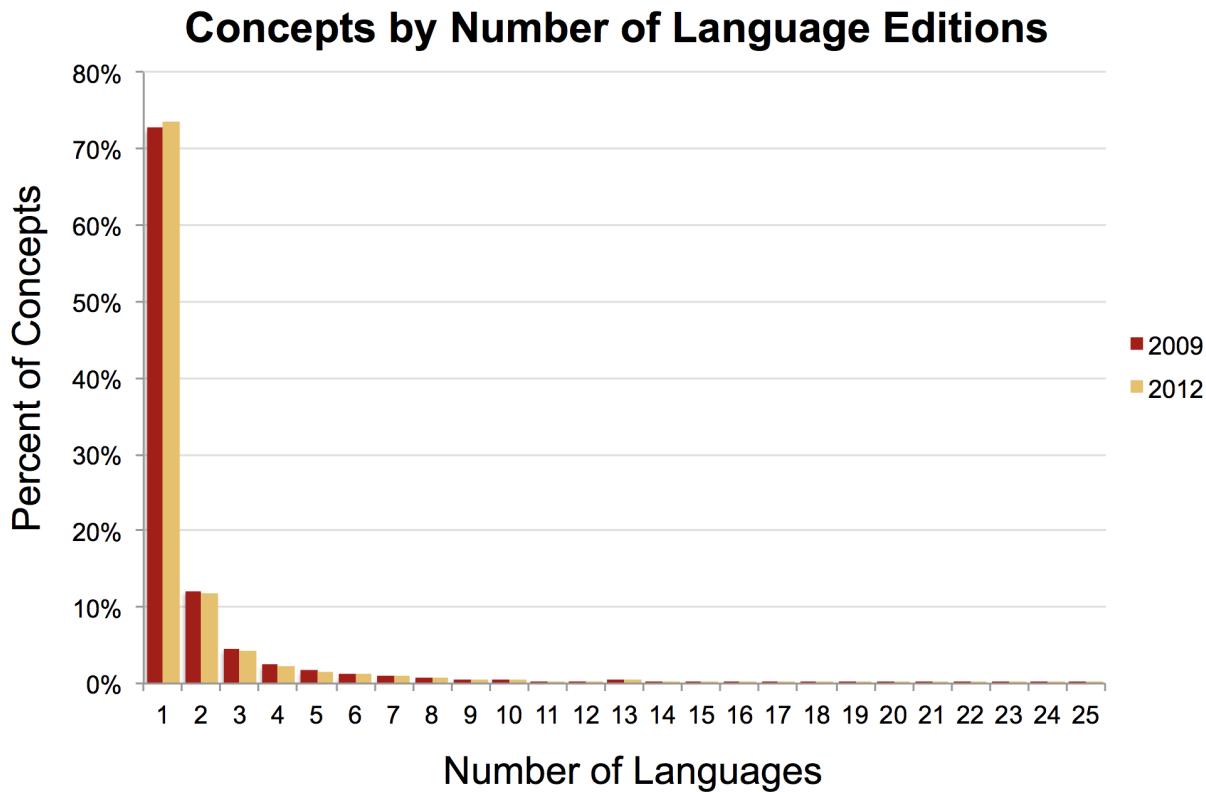


Figure 3.9-a: The conceptual coverage distribution in 2009 (red) versus that in 2012 (orange). Almost the exact same percentage of concepts appear in each number of language, with the percentage of single-language and global concepts being very similar.

concepts to *Conceptualign*'s 8.67 million.

With the older concept alignment algorithm applied to our 2012 data, we could compare the conceptual coverage distribution in 2009 to that in 2012. Figure 3.9-a shows that comparison. As can be seen clearly in the figure, the distribution is almost exactly the same in 2012 as it was in 2009. In other words, significant growth in the number of articles does not appear to reduce (or increase) the extensive amount of concept-level diversity in multilingual Wikipedia. Where an article gets created about a concept that already is covered by a language edition, it is counterbalanced by another article on a concept that is new to multilingual Wikipedia.

Figure 3.9-b provides a closer look at the differences in the conceptual coverage

distributions, nearly all of which are too small to identify clearly at the scale necessary to show the whole distributions in Figure 3.9-a. Each bar in Figure 3.9-b shows the change in the *relative* share of conceptual coverage (i.e.  $1 - (\% \text{ in 2012} / \% \text{ in 2009})$ ). As such, the first bar on the left indicates that the ratio of single-language concepts to all concepts in 2012 is one percent higher than it was in 2009, and the bar furthest to the right shows that the ratio of global concepts in 2009 was four percent higher than it is now.

Although they certainly occur at the margins of the overall trend of nearly constant conceptual coverage, Figure 3.9-b does show several interesting small-scale divergences. For instance, there appears to be an increase in the percentage of concepts that are covered by large

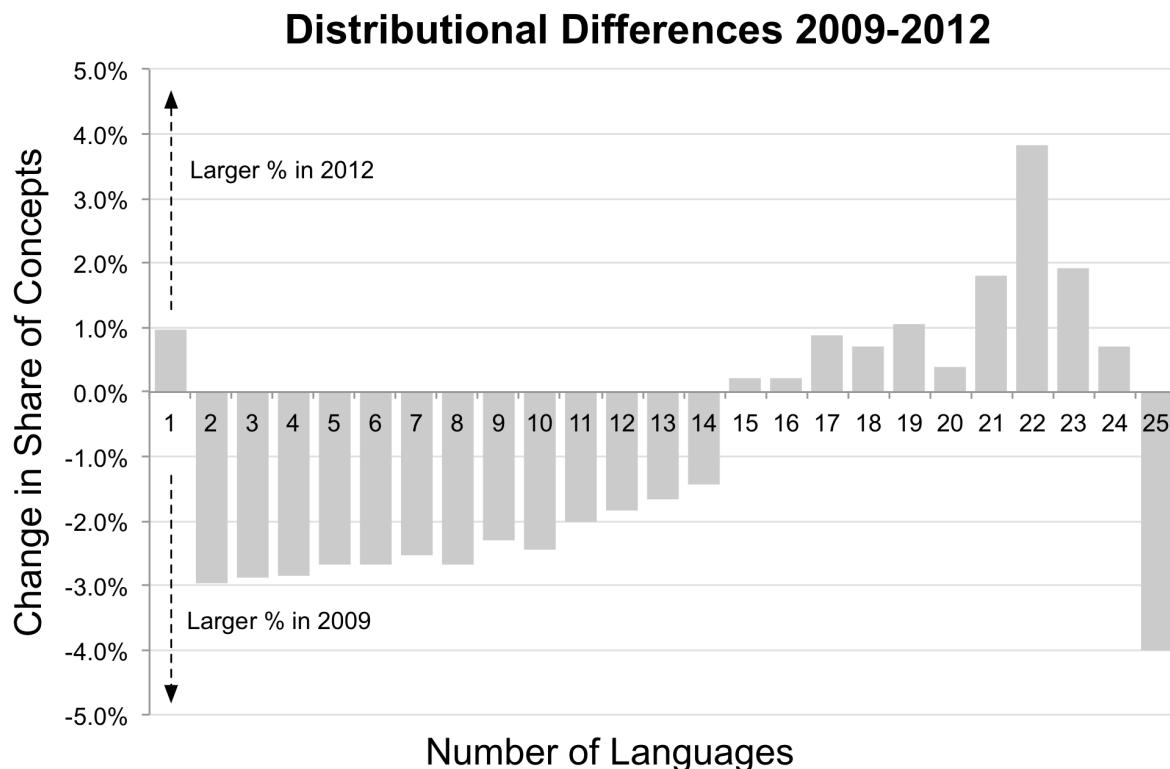


Figure 3.9-b: The relative difference between the conceptual coverage distribution in 2012 and that in 2009. Positive numbers indicate a greater share of concepts in 2012, negative numbers indicate the opposite. As such, the first bar on the left indicates that the ratio of single-language concepts to all concepts in 2012 was 1% higher than it was in 2009.

numbers of language editions, a growth that is substantially reversed for the special case of global concepts. There are a number of possible causes of this trend. For instance, the smaller language editions could be covering more concepts in the larger language editions, but not the same set of concepts, so the global concepts' share is reduced. This same phenomena could be behind the complimentary reduction in the share of low- to mid-range conceptual coverage concepts. We leave exploring these perturbations in more detail to future work.

### **3.9.2 Sub-concept-level Diversity**

In the previous subsection, we saw that the amount of concept-level diversity has stayed the same over the past three years. In other words, the language editions covered extensively different sets of concepts in 2009, and they do now as well. But what about how those concepts are covered? Has the average pair of articles about the same concept in two language editions grown more similar or more different over time?

One field of argument akin to the global consensus hypothesis is that once information appears in one language edition, it will eventually be transferred to the other language editions, creating a situation in which the language editions become more equal over time. Alternatively, while the above may be occurring, it could be at least outpaced by the language editions adding novel unique content, much of it culturally contextualized as we have seen above.

To compare sub-concept-level diversity in 2009 to that which exists in our current dataset, we again repeat an experiment we performed in our 2010 CHI paper. As above, our methods were less sophisticated at the time, so we had to adapt our more advanced methods from Section 3.5 to make them compatible with our 2010 sub-concept-level analyses. In the language of Section 3.5, the 2010 paper compared the bags of links between two articles using the “just

links” lower-bound wikification approach, using parseable links only, and ignoring sub-articles.

The experiment in our 2010 paper involved using bags of links as defined above to calculate the overlap coefficient ( $OC$ ) – defined in Section 3.5 – for *all* pairs of articles in all global concepts that had a minimum of three parseable outlinks and a minimum of three parseable inlinks for all languages. These thresholds were established in order to ensure that each article examined was sufficiently developed and sufficiently integrated into its language edition.

With our 2012 data, we calculated the overlap coefficient exactly as we did in our 2010 paper and did so for all concepts that met the exact same sampling guidelines. After the overlap coefficient ( $OC$ ) for each eligible 2012 article pair was calculated, we found that the mean  $OC$  was 0.4558. In 2009, the mean  $OC$  was 0.4641. Figure 3.9-c shows the nearly-identical  $OC$  distributions behind these nearly-identical means. In over three years of significant growth in multilingual Wikipedia, very little has changed with regard to the extent to which longer articles

### Overlap Coefficient Distribution over Time

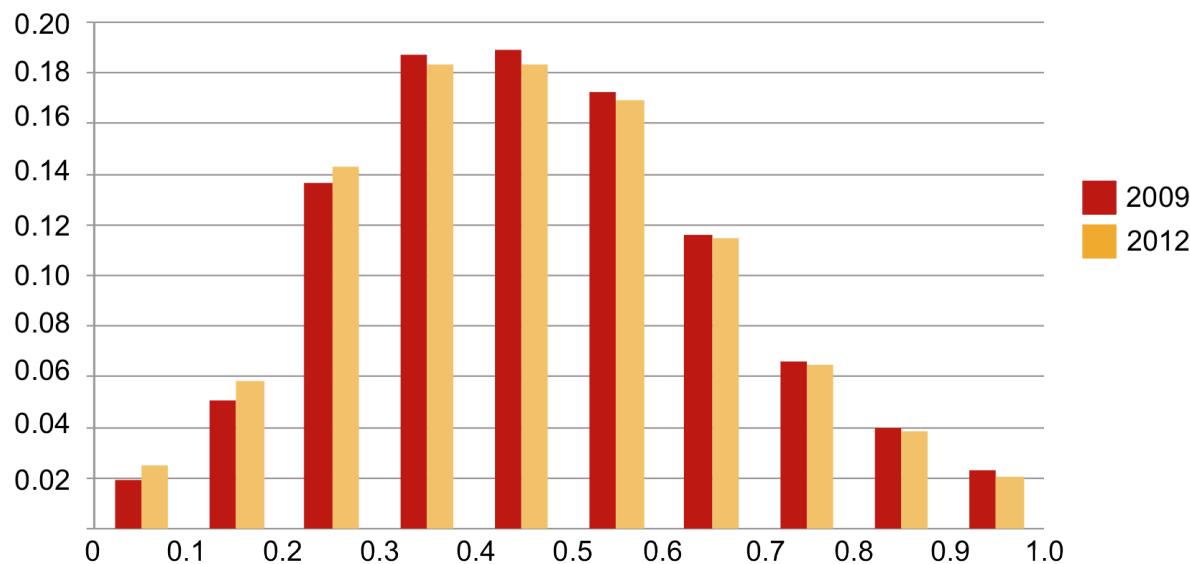


Figure 3.9-c: The overlap coefficient ( $OC$ ) distribution in 2009 (red) and that using our current dataset from 2012 (orange). Despite three years of growth, the sub-concept-level diversity has remained at almost exactly the same level.

contain the links of shorter articles about the same concepts. If we adopt the bag-of-links assumption, we can say that little has changed with regard to the extent to which the longer articles cover the same content as the shorter articles.

Taking a step back, this finding strongly suggests that the amount of sub-concept-level diversity in the 25 language editions has remained remarkably constant, despite three years of massive growth in multilingual Wikipedia. This consistency is all the more remarkable considering that the article graphs of each language edition (WAGs) have grown even more than the number of articles, with the graphs having an enormous role in the overlap coefficient calculation. While there has likely been information transfer across the language editions, this is almost exactly balanced out by new, unique content being written in each language edition, both in terms of new articles and new content on existing articles.

### **3.9.3 Discussion**

The concept-level and sub-concept-level results in this section all point to the same conclusion: the extensive amount of diversity in multilingual Wikipedia has not been affected by multilingual Wikipedia's substantial growth. The implications here are significant. First, at least according to data from the past three years, the language editions of Wikipedia are not growing more and more similar, as some might predict, but they are also not diverging from one another. This is true despite substantial efforts to translate information from the English Wikipedia to the other language editions (e.g. [35, 198, 208]). Second, the surprisingly static nature of our concept- and sub-concept-level results points to the diversity between the language editions being a property of the underlying content generation process in multilingual Wikipedia. The shape of the concept-level diversity distribution suggests preferential attachment may be at play,

but a great deal more work is needed to establish if this is the case, as well as to better understand the lack of change in sub-concept-level diversity.

### **3.10 Cultural Context and Multilingual Wikipedia Diversity**

Throughout this chapter, we have frequently used small sets of intuitive examples to illustrate the role of cultural contextualization in the diversity across multilingual Wikipedia. However, in order to definitively show that culture is the cause of at least some of the similarities and differences between the language editions, more robust approaches are required. In particular, we need to adopt a method that both (1) is informed by social science theory rather than intuition and (2) scales to all of multilingual Wikipedia rather than relying on a small set of examples.

In this section, we first present a geographic information science-inspired cultural context mining method that has both of these properties. Next, we apply this method to our 25-language dataset, demonstrating quantitatively that each language edition contextualizes encyclopedic world knowledge for its own language-defined culture, and does so in a substantial fashion. Finally, building on the discussion in Section 3.8, we compare the cultural contextualization in content to that in its consumption and discuss implications.

#### **3.10.1 Methods: Mining Cultural Context**

##### **3.10.1.1 Theoretical Motivation**

Strong motivation from social science theory is essential to the validity of any method that claims to mine cultural context. Our approach draws on two fundamental theories from the field of human geography. First, human geographers have long known that human populations tend to be spatially autocorrelated, or spatial clustered. The entire idea of regions – one of the five

themes of geography [112] – depends on this phenomenon. The spatial autocorrelation of human populations applies to a large variety of population types, including that of language speakers. More formally, if a person  $a$  is near a person  $b$  who speaks language  $l$ , person  $a$  is likely to also speak language  $l$ . Scaling this process up to the entire world results in a heavily regionalized geography of world languages [116]. That is, people who speak Japanese tend to live in Japan, people who speak Finnish tend to live in Finland, and so on. For many languages, the regions are not contiguous – English, Spanish, and Portuguese for example – but this in no way violates the assumption of spatial autocorrelation and the resulting regionalization.

The second human geography theory that motivates our cultural context mining approach describes the typical relationship between distance and spatial interaction. Namely, holding other factors constant, as distance increases, spatial interaction decreases. This theory is sometimes referred to as “distance decay” [40], and is fundamental to cornerstones of human geography such as central place theory [23, 28]. In the framework of cultural context, distance decay suggests that geographic features nearby a cultural group will be more likely to be in that cultural group’s shared expertise (cf. Clark [24]) than places that are farther away. Indeed, Clark frequently uses geographic features as examples of cultural shared expertise, writing that certain types of cultural communities – namely, those that are regionalized – share expertise such as “local geography, civil institutions, practices, argot, [and] national cultural practices” [24].

Putting these two theories together, we can safely assume that geographic features nearby a given language-defined community’s *home cultural regions*<sup>42</sup> are more likely to be in the shared expertise, or cultural context, of the language-defined community. Indeed, this is merely Clark’s

---

42 Cultural geography has a variety of terms to describe the result of culture-related spatial processes (e.g. ‘core’, ‘culture hearth’, ‘domain’). We use the term “home cultural region” as a synonym to “culture region” that is more descriptive to a general audience.

statement, applied to language-defined cultures. Based on this assumption, we are able to develop quantifiably testable hypotheses that allow us to investigate the extent to which the user-generated content in multilingual Wikipedia reflects the cultural contexts of its contributors.

First, if each language edition does in fact contextualize encyclopedic world knowledge for its corresponding language-defined culture, we would expect the culture's home regions – its geographic shared expertise – to play a disproportionately prominent role in the language edition relative to that of other geographic areas. For instance, there might be single-language concepts about places in the home regions that are not covered in other language editions. Similarly, when describing a non-geographic concept – say highways – the language edition might discuss general ideas related to highways in the framework of examples drawn from the shared expertise of its readership. In the framework of this chapter, this scenario can be interpreted as a version of the global diversity hypothesis in which cultural context is an explicit factor.

On the other hand, we might also hypothesize that there will be widespread agreement – or a “global consensus” – among the language editions as to the most prominent geographic regions in the world. This would suggest that the information in each language edition does *not* reflect the geographic shared expertise – or cultural context – of its audience. For instance, highways might be described using the same world-famous highways in most or all language editions, and geographic single-language concepts might not display any relevant spatial pattern in each language edition.

### **3.10.1.2 Connecting Wikipedia to Geography**

Prior to determining the support for either of the above two hypothetical scenarios, it is first necessary to *geospatially reference* multilingual Wikipedia. That is, concepts about geographic

features need to be connected to the locations of those features. For instance, the multilingual Wikipedia concepts about Troy, Michigan; the United States; California; Northwestern University; the MacKenzie River; Tufts University; Omaha, Nebraska and so on, need to be attached to spatial data representations such as latitude/longitude coordinates and/or (multi)polygons.

#### *Latitude/Longitude “Geotags” in Wikipedia articles*

Many Wikipedia articles about geographic features in a variety of language editions include latitude/longitude coordinates for the subject of those articles. For instance, in Figure 4-a, the editors of the English Wikipedia have tagged the “Troy, Michigan” (English) page with the latitude and longitude of the geographic center of the city. While our early work in this area involved extracting these lat/lon values ourselves, our later work has relied on DBpedia’s [13] dataset of extracted coordinates<sup>43</sup>, a dataset that is utilized frequently in the literature and in geographic Wikipedia applications. In its most recent version, the DBpedia dataset includes coordinates for approximately 665,000 concepts.

However, even though the DBpedia dataset is used widely and has a large number of concepts, through careful examination we recently established that it is missing a substantial number of important lat/lon “geotags.” For instance, the concepts San Francisco, Houston, the Great Wall of China are all omitted, even though each has coordinates on its English Wikipedia page. Although researchers have identified other problems with the DBpedia dataset [97], the important omissions in the dataset have yet to be reported. We also noticed additional, non-trivial problems with DBpedia’s lat/lon geotags, such as a substantial portion of the longitudes extracted

---

43 <http://wiki.dbpedia.org/Datasets#h18-17>

from the Spanish and Catalan Wikipedia being reversed (e.g. 100°E being reported as 100°W), a problem we had to fix in order to allow for comparison of content about geographic features across language editions.

Due to the omissions in DBpedia, it was necessary to seek out an additional source of Wikipedia geotags. For this purpose, we turned to another database of Wikipedia lat/lon coordinates called Wikipedia-World [211]. With geotags for 1,348,792 concepts, Wikipedia-World is significantly larger than the DBpedia dataset, but it is also utilized significantly less often (although it has leveraged for a few research projects and applications, e.g. [70, 212]). One reason Wikipedia-World is employed more rarely is that it includes *all* latitude and longitude coordinates, not just those that describe the subject of an entire page. This makes the interpretation of these geotags less straightforward. For instance, the article “List of windmills in the United States” (English) has many lat/lon coordinates – one for each windmill. As such, none of these lat/lon tags individually describe the geospatial footprint of the overall concept, which should be the aggregate footprint of windmills in the United States. This is a non-issue with DBpedia; all of its lat/lon tags have a 1:1 relationship with articles.

In order to leverage the rich geographic information in Wikipedia-World while at the same time avoiding the hazards of multiple lat/lon coordinates per page, we used a simple but esoteric spatial data representation called a *multipoint* [202] that is supported by most geographic information systems. Multipoints are what they appear to be: a geometry that consists of a collection of points. Across the entire Wikipedia-World dataset, we took all the lat/lon tags on pages like “List of windmills in the United States” (English), and aggregated them into a multipoints. Doing so was essential to the accuracy of the geospatial operations most central to this research in this section. For example, without aggregating into a multipoint all the

coordinates on the “List of windmills in the United States” (English) article, spatial containment operations – so important to the work below – would fail to report accurate information. Namely, these operations would assign all properties of the article (inlinks, etc.) to all U.S. states in which just one of the lat/lon pairs was located, when in fact the article is about entities in a variety of states. With the multipoint, spatial containment operations only consider the concept to be contained within the United States, an appropriate assessment given the nature of the concept as a list of U.S. entities.

Once we had created all the necessary multipoints, we merged the Wikipedia-World data with that of DBpedia, creating or adding to multipoints where the geometries did not agree. Since both datasets include tags from multiple language editions, we also created multipoints when the tags disagreed across language editions, a phenomenon which is a topic of future work. In the end, 1,369,365 concepts in our multilingual Wikipedia dataset were associated with geographic point representations, a group of concepts that did include San Francisco, Houston, and the Great Wall of China. The group also included over 20,000 concepts not in Wikipedia-World, meaning that Wikipedia-World is missing a small amount of the DBpedia dataset as well.

### *The Geoweb Scale Problem in Wikipedia*

After attaching latitude and longitude coordinates to all 1.37 million concepts in our merged geotag dataset, successfully connecting Wikipedia to geography required overcoming one additional major obstacle: the Geoweb Scale Problem (GSP) [79, 84]. We define the GSP to occur when the representations of geospatial footprints mined from the Web are inappropriate for the scale of analysis required to investigate a given geographic research question. The simplicity, ubiquity, and impact of the GSP cannot be understated, the latter two of which are the result of

point-based spatial data representations' near-total dominance in user-generated content [97]. In UGC, one often finds regions as large as entire states, provinces, and countries encoded as single lat/lon points. This creates serious and often-ignored hazards for analyses that rely on spatial operations ranging from those involving distance to those requiring containment relationships. The spatial data in Wikipedia is no exception here. All 50 states in the U.S., for instance, are represented as a lat/lon coordinate (corresponding to their geographic center) in our geotags dataset. In fact, the entire United States is encoded in the same way.

In our textbook chapter on doing geographic virtual communities research [79], we write:

“...there is no easy solution [to the GSP]. The two approaches used in the literature are either to (1) redefine your study around the spatial representation limitations of your data or (2) filter your data to remove the most egregious cases.”

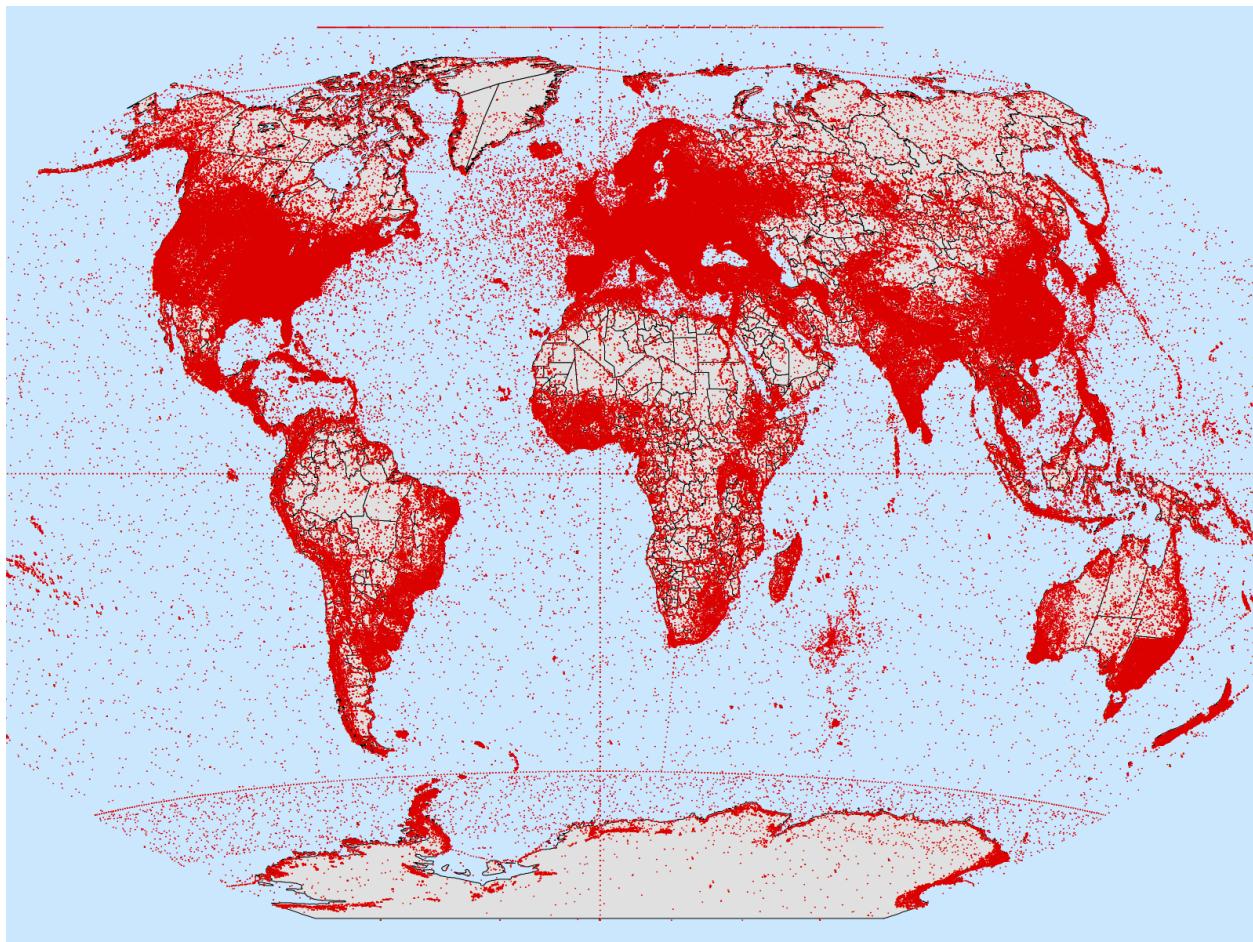
In this work, we use both approaches. With regard to the latter, we manually connected all countries and all first-order administrative districts (states, provinces, oblasts, etc.) to detailed polygonal models of their geospatial footprints, filtering out their point representations<sup>44</sup>. The polygonal representations were extracted from ESRI datasets<sup>45</sup> and the GADM database of Global Administrative Areas<sup>46</sup>. Doing the same for larger-scale geographic features like second-order administrative districts, cities, natural parks, legislative regions, however, is intractable without extensive manual labor resources. This is where the first approach mentioned in our textbook chapter comes in. In this chapter, we only perform analyses at the scale of countries and first-order administrative districts. To do anything else with even our improved spatial data

---

<sup>44</sup> We also connected their subarticles to these representations. This was a very important step, as it prevented our methods from interpreting concepts like those described by “Geography of the United States” (English) and “History of the United States” (English) from being associated with a single lat/lon point within the United States.

<sup>45</sup> <http://www.esri.com/data/data-maps>

<sup>46</sup> <http://www.gadm.org/>



*Figure 3.10-a: A map showing the distribution of geographic concepts. Red dots indicate concepts associated with lat/lon tags from DBpedia and Wikipedia-World. The grey polygons are first-order administrative districts. Countries are not shown as they are covered by the districts.*

representations would be to introduce extensive Geoweb Scale Problem-induced error. Future work may involve using crowdsourcing approaches (or, hopefully, datasets developed by other groups) to improve large-scale spatial representations enough for us to use them in similar types of analyses.

Once we had combined our dataset of geotags with the improved spatial data representations for countries and first-order administrative districts, we had our final geographic representation of Wikipedia concepts. A high-level descriptive map can be found in Figure 3.10-a. It is clear that our 1.37 million geographic concepts cover much of the world, even rural areas.

### 3.10.1.3 Cultural Context Metrics

At the start of this section, we hypothesized that if multilingual Wikipedia reflects the cultural contexts of its contributors, each language edition should give the corresponding language-defined community’s home cultural region a “disproportionately prominent role” in encyclopedic world knowledge relative to that of other geographic areas. Above, we showed how we can connect geographic concepts in Wikipedia to representations of their spatial footprints, setting the stage for formal geographic analyses. In this section, we focus on methods related to the final part of the above hypothesis. That is, we introduce techniques to measure the prominence of geographic regions in each language edition, techniques we use in the next section to compare the prominence of home cultural regions against that of other regions.

Our approach to evaluating the language edition-specific prominence of geographic regions around the world is a two-stage process. First, we apply a prominence score to the subset of the 1.37 million geographic concepts that are covered in each language edition. We then use spatial containment operations to aggregate these scores by region.

The use of the term “prominence” in our hypothesis is purposefully flexible. The prominence of a given concept in Wikipedia can be measured in any number of ways, and we use several prominence metrics throughout this section. That said, our focus is on two such metrics, indegree and PageRank scores, the two network centrality measures that were leveraged extensively in Section 3.6. Indegree and PageRank scores have three major benefits for the purposes of this study:

- They are accurate indicators of the importance of a geographic concept as determined by the link graph of each language edition.
- They are strong measures of how much a given concept is discussed in each language

edition (cf. Section 3.6).

- Rather than just focusing on a single article, they provide an understanding of how a geographic concept is integrated into an entire language edition. While a single editor can drastically change article-based statistics herself, this is much more difficult in the case of language edition-based statistics like indegree and PageRank scores (cf. Section 3.6).

Thanks to these benefits, if we find that one geographic concept has a higher relative indegree or PageRank score than another in a given language edition, we can say that (1) it is considered more important by the language edition, (2) it is discussed more often throughout the language edition, and (3) this finding is reflective of how the entire language edition accounts for the concept, rather than simply reflective of the information on a single page about that concept, which is subject to outlier effects. Combined together, these three statements make a strong argument for the overall prominence in the language edition of the geographic concept with the high indegree / PageRank score.

For the purposes of completeness and to help understand any outlier effects, we also adopt two additional measures of geographic concept prominence: outdegree and article count. The outdegree of a geographic concept in a language edition is the number of links in the language edition's article(s) on the concept. Following the bag-of-links assumption, outdegree is a good proxy for the amount of content in an article. Article count is simply a function that returns one if a language edition covers a concept, and zero otherwise. When aggregated over regions, the article count metric provides a basic understanding of the geography of the language edition's conceptual coverage.

Regardless of the prominence metric used, an approach for aggregating these metrics over regions must be adopted. The most trivial such approach is to approximate the prominence of a

region by setting it equal to the prominence value of the single concept about the region. For instance, we could assess the indegree prominence of Finland in each language edition by setting it equal to the indegree of the article about Finland in each language edition. This approach is flawed, however, in that it ignores the potentially massive amount of shared expertise about places *in* Finland that might be reflected in some of the language editions, particularly Finnish. For instance, this “1:1” approach ignores the indegree of major Finnish cities like Helsinki, small Finnish towns, Finnish high schools, and so on. Indeed, following Clark [24], it is likely that while the country of Finland is in the shared expertise of many language-defined cultures, Finnish high schools are likely not, and the high schools and related concepts make up an important component of Finnish speakers’ shared expertise. Additionally, omitting the articles about smaller geographic features mutes the effect of concept-level diversity in our geographic prominence assessments; all language editions have articles about Finland, but only Finnish has an article about Brännskär, an island off the southwest coast of the country<sup>47</sup>.

To address this issue, we use an aggregation approach based on spatial containment. For each region considered, we perform what is known in the geographic information science community as a “spatial join.” This process involves summing<sup>48</sup> the prominence of all geographic concepts located within a region and setting the geographic prominence of the region equal to the result. We do not use strict spatial containment, allowing for the concept representing the region itself (e.g. Finland in the example above) to be included in the summation. This spatial containment-based approach means that for each language edition, the

---

<sup>47</sup> It is important to point out, however, that not all concept-level diversity would be muted. Non-geographic articles that only exist in the Finnish Wikipedia but link to Finland would still have a substantial effect on the indegree of Finland in Finnish.

<sup>48</sup> A spatial join can use any aggregation operation. We use a sum.

indegree of the article about Finland *and* the article about Brännskär, which is zero in the case of all Wikipedias but Finnish, are included in Finland's prominence score.

Due to the geospatial aggregation inherent in the prominence score of each region, in our past work we have referred to our final prominence metrics as *spatial indegree sums*, *spatial PageRank score sums*, and so on. This is nomenclature we also adopt here. For instance, if we find that all articles about places in Finland have a total indegree of 500,000 in the English Wikipedia, we would say that the English spatial indegree sum for Finland is 500,000.

We aggregate prominence scores for concepts over two types of regions: countries and first-order administrative districts. We use these region types for two reasons. First, the nature of the spatial autocorrelation of language-defined cultures results in countries and first-order administrative districts being ideal data points for our study. Language-defined cultures tend to cluster in countries (e.g. Finland, Japan, etc.) and states/provinces (e.g. Québec, East Flanders), which means that countries, states, provinces, etc. are units of analyses with minimal noise in this context. The second impetus behind our selection of region types was of course the Geoweb Scale Problem, which prevents us from doing more local-scale analyses even if we desired to do so.

Reflecting on the Geoweb Scale Problem in greater detail, let us imagine the effect of totally ignoring the GSP on prominence score sums. The United States provides an excellent example in this context. As noted above, the United States is tagged with the lat/lon coordinate of its geographic center in many language editions. Since the statehood of Alaska and Hawaii, this center has fallen in the middle of the state of Kansas. Had we not upgraded the representation of the United States' geospatial footprint to a multipolygon, the indegree sum for Kansas in each language edition would include the indegree of the language edition's article

about the United States. This is a severe mistake not only because it mischaracterizes the role of Kansas in the encyclopedias, but also because it results in an enormous outlier. Without correcting for the GSP, Kansas quickly becomes one of the most important and most-discussed first-order administrative districts in the world in all language editions. While Kansas is the boyhood home of Dwight D. Eisenhower [36], contains the World's Largest Hand Dug Well<sup>49</sup>, and is notable in the geography literature for being flatter than a pancake [44], correcting for the GSP results in Kansas becoming much less prominent in all language editions. Of course, unlike sororities and quarterbacks (Section 3.7), this is not a phenomenon that is mostly restricted to the United States. The GSP affects the prominence sum calculations of all countries in the world.

Our analyses in this section also show that the increasing number of Wikipedia-based research projects and applications that use GSP-affected spatial operations such as containment (e.g. [64, 65]) must also consider sub-articles when addressing the GSP. We not only found that each country was tagged with a lat/lon coordinate, but we also saw that many of their sub-articles were also tagged with the same coordinate. For instance, the article “Geography of the United States” (English) is also geotagged with the geographic center of the United States. In order to fully remove the effect of the GSP, we had to successfully identify these sub-article relationships (Section 3.5.1.3) and use these relationships to connect sub-articles to their main articles’ geospatial footprint representation, a task that – as we show in Section 3.5.1.3 – is much more complex than simply associating an article with a polygon. Failing to do so in the case of the “Geography of the United States” (English), for instance, results in Kansas being the most prominent U.S. state according to PageRank and indegree, although the effect is less severe than is the case with “United States” (English).

---

49 <http://www.bigwell.org/>

After we had calculated all prominence sum metrics for all geographic concepts and had aggregated all these metrics at the country and first-order administrative district levels accounting for the Geoweb Scale Problem, we had the data necessary to test the hypotheses we presented at the beginning of this section. This is a process we begin below.

### **3.10.2 Results: Cultural Context in Multilingual Wikipedia**

Recall that the goal of the present study is to determine if each language edition depicts the corresponding language-defined community's home cultural regions as disproportionately important to encyclopedic world knowledge compared to other geographic regions. If this is the case, it would strongly suggest that the language editions of Wikipedia reflect the cultural contexts of their contributors. If, however, there is a consensus among the language editions as to the geographic areas most important to encyclopedic world knowledge, this would provide evidence to the opposite.

Once all of the geographic prominence sums were calculated as described above, determining the support for each of the above scenarios was a relatively straightforward process: we simply compared the sums for the home cultural regions for each language edition to the sums of other regions. The best approach to establishing the extent of a cultural region usually involves a combination of qualitative field work and in-depth quantitative analysis. Moreover, the result of this work is typically a continuous surface, not regions with discrete boundaries. For our present work, however, we take a satisficing heuristic approach to home cultural region identification and leave the application of more nuanced techniques to future work. Specifically, we define any country to be in a language-defined community's home cultural region if the language is an official or de facto official language of the country as determined by "List of

official languages” (English) [122]<sup>50</sup>, which is a relatively well-sourced Wikipedia page.

Additionally, a trained human geographer reviewed the list of home cultural regions and found no errors. As we will see, this heuristic approach was more than sufficient for the needs of this study.

Below, we present the results of the prominence sum comparisons between home cultural regions and other regions in two ways. First, we illustrate our findings using cartographic best practices. Next, we introduce and utilize our “self-focus bias index”, which provides a different, more summative understanding of the cultural contextualization in multilingual Wikipedia. We begin with prominence sum visualizations.

### **3.10.2.1 Visualizing Cultural Context in Multilingual Wikipedia**

Cartographic visualization is an ideal medium for initially introducing the results of this study for one reason: it is an excellent communicator of the study’s lopsided results. For almost every language edition and for almost every prominence sum metric, the home cultural region of each language edition was *the most prominent region in the entire world*. In other words, the language editions strongly reflect the cultural contexts of their contributors.

Figures 3.10-b through 3.10-h are maps of the PageRank score sums and indegree sums in a variety of different language editions. All of the maps use an equal interval classification scheme, which means that the range of each prominence score has been split up into equal bins, with each bin assigned a color. This classification scheme makes the extent of cultural contextualization in these language editions strikingly clear. In all the maps, no country other than those in the home cultural regions of the language-defined culture is in the top two bins.

---

50 The countries in each language-defined community’s home cultural regions can be found in Appendix G.

Moreover, in many of the maps the home cultural region countries are several bins beyond all other countries. Note also that despite the similarities between the Scandinavian language editions we have seen throughout this chapter, these language editions still reflect a great deal of the context of their corresponding language-defined culture.

In the context of the interpretations of PageRank score sums and indegree sums above, each of the maps in Figures 3.10-b through 3.10-h tell us that (1) home cultural regions are the most important regions in the entire world according to the encyclopedic world knowledge in each language edition, (2) home cultural regions are the most discussed regions in the entire world in each language edition and (3) these results reflect the encyclopedic world knowledge throughout each language edition, not just the information on a single page.

That said, in terms of the information on single pages, we found nearly identical results for the outdegree sum and concept count prominence metrics. For instance, as demonstrated in Figure 3.10-i, Turkey has by far the highest outdegree sum of any country in the world in the Turkish Wikipedia. The same is true of concept counts for China in the Chinese Wikipedia (Figure 3.10-j). These findings tell us that not only do the article graphs of entire language editions have a bias towards their home cultural regions, but this is true of individual articles as well. In other words, in multilingual Wikipedia, articles about spatial features in home cultural regions are much greater in number and have a much higher number of aggregate links than articles about other regions.

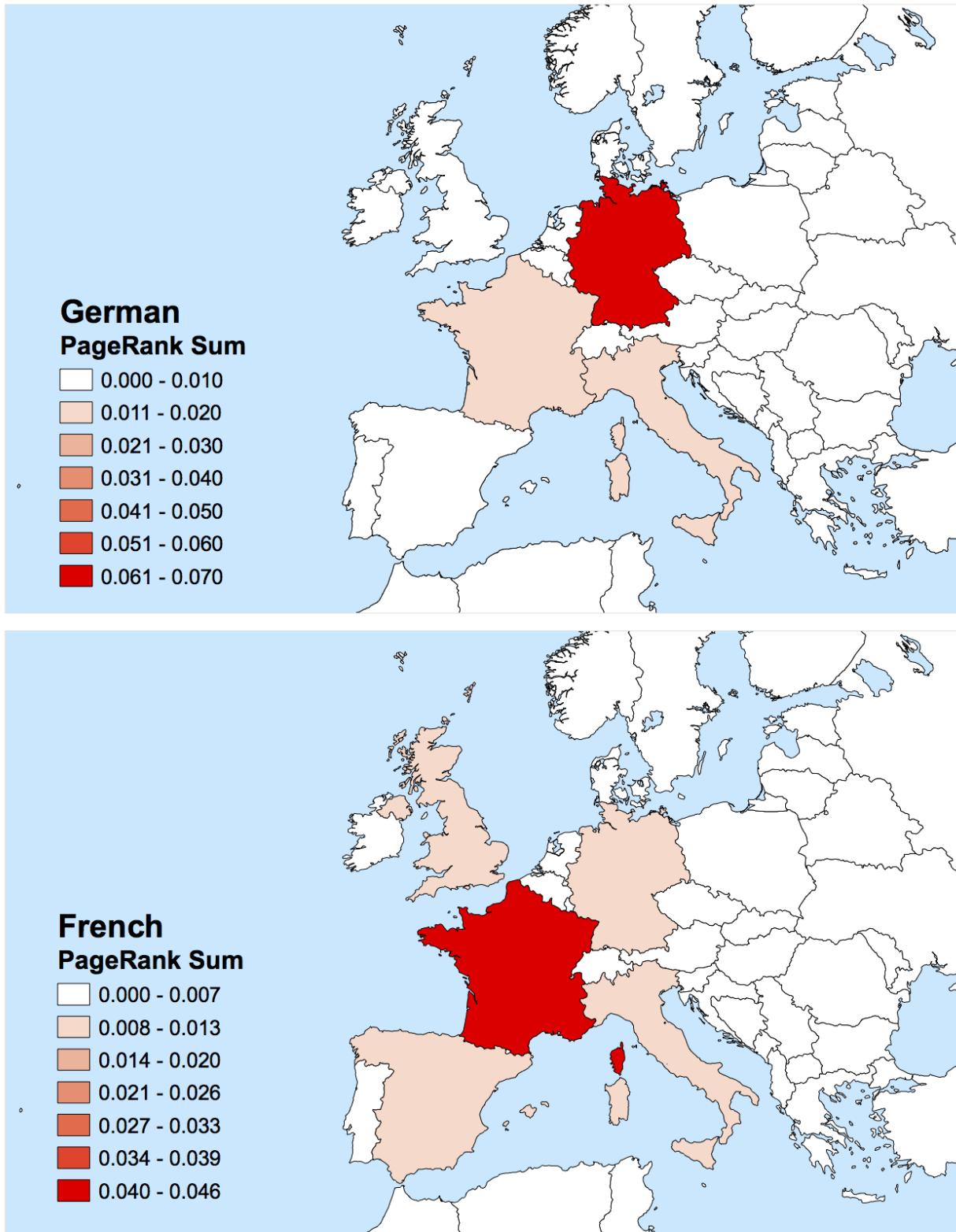


Figure 3.10-b: PageRank score sums for the German and French Wikipedias.

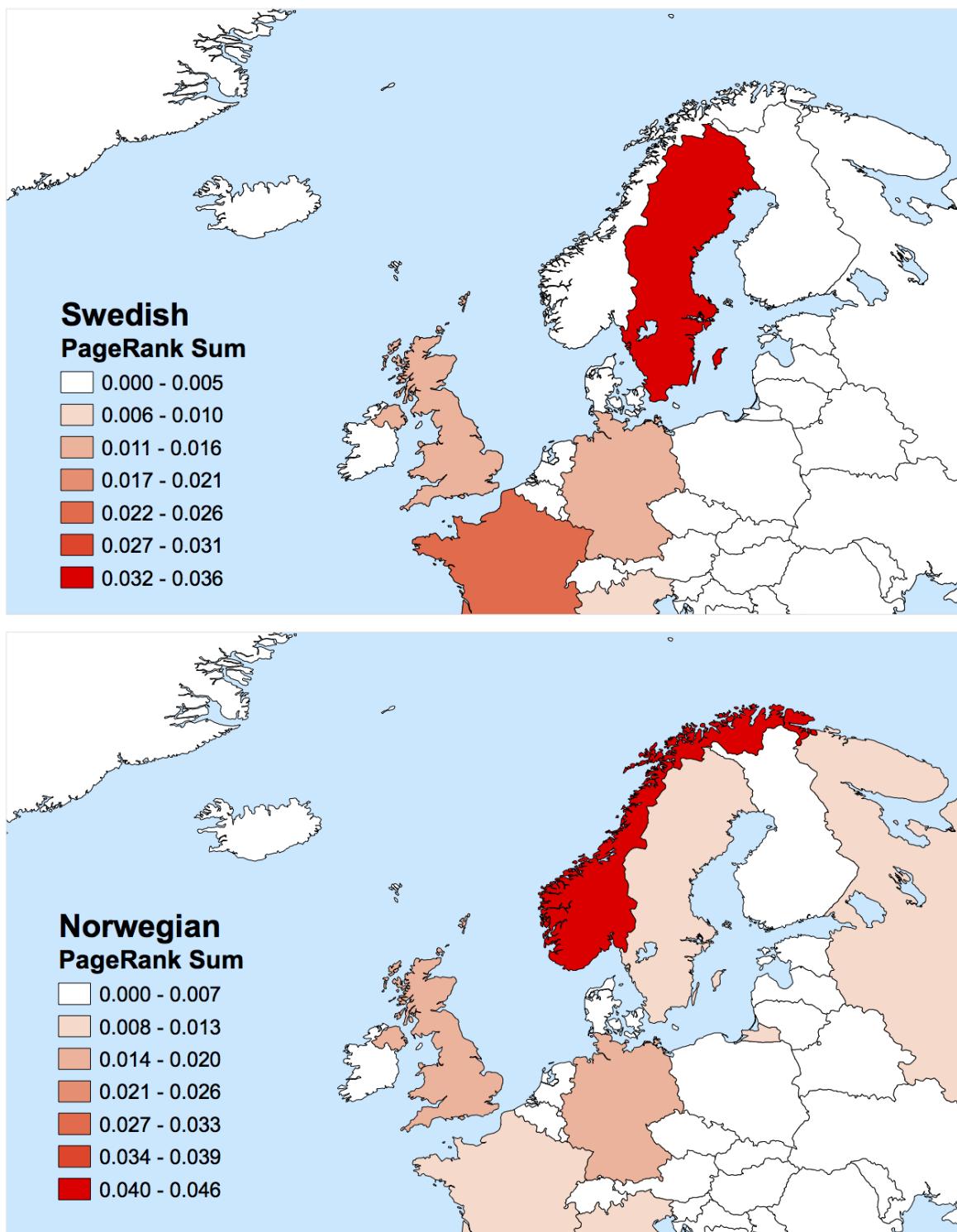


Figure 3.10-c: PageRank score sums for the Swedish and Norwegian Wikipedias.

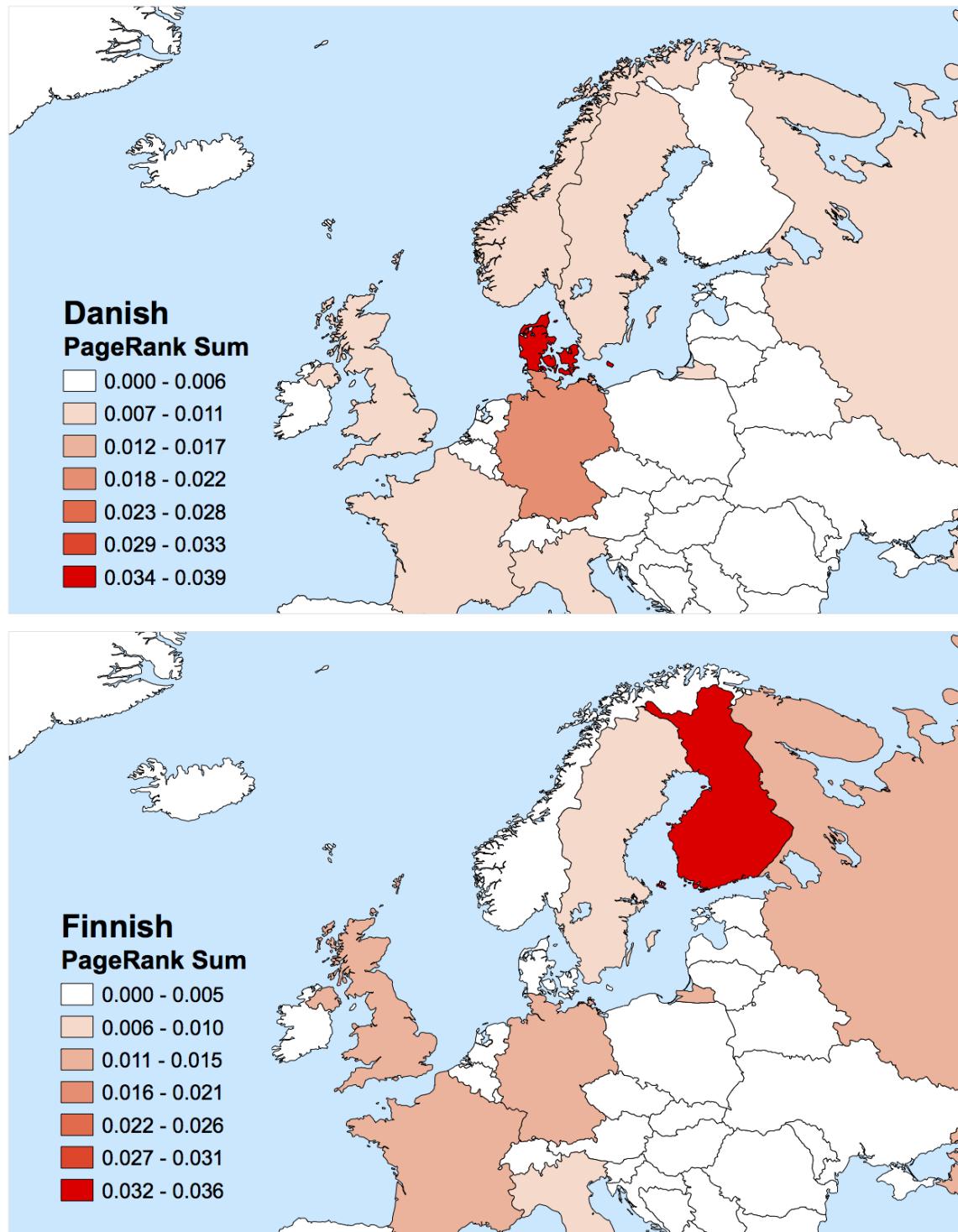


Figure 3.10-d: PageRank score sums for the Danish and Finnish Wikipedias.

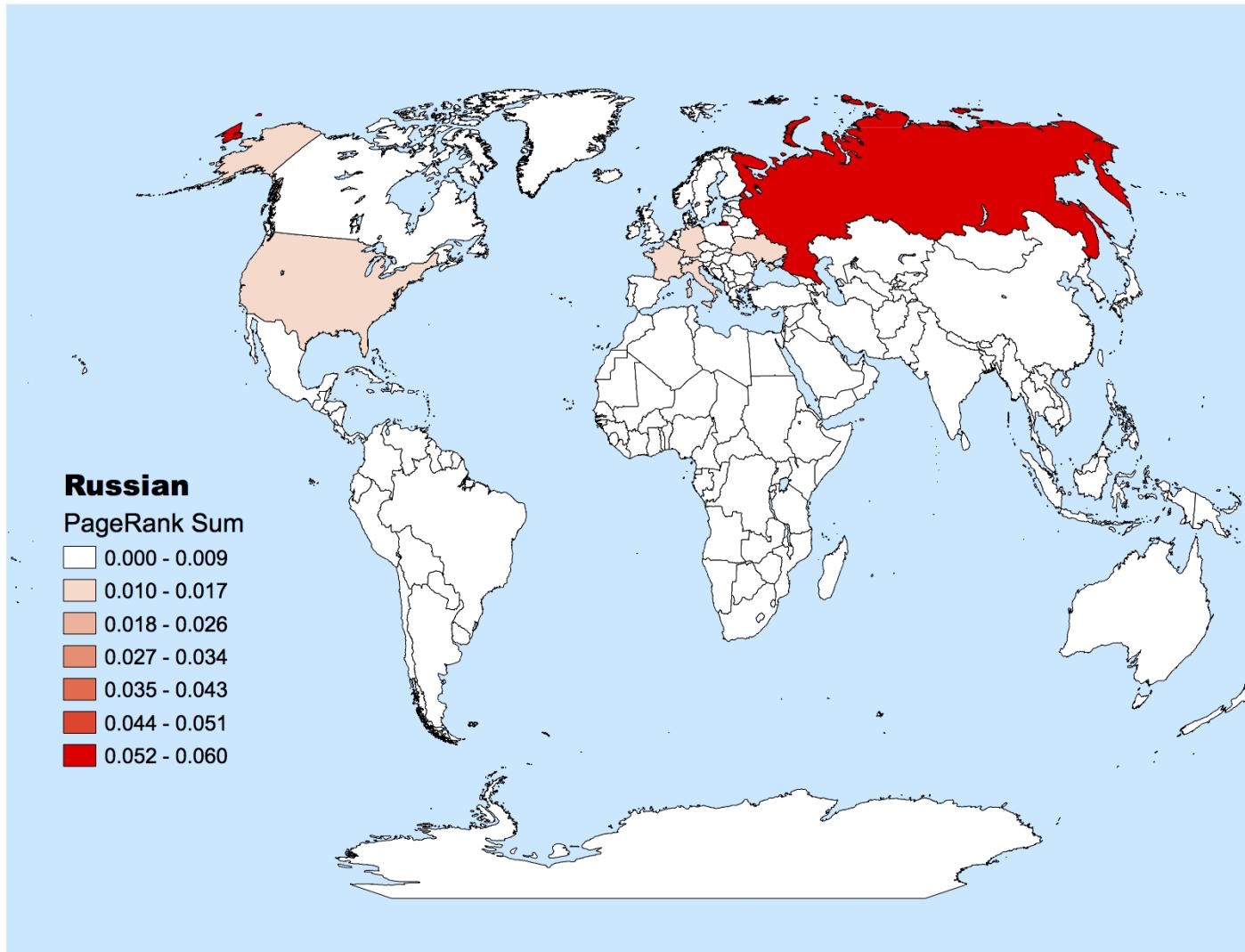


Figure 3.10-e: PageRank score sums for the Russian Wikipedia.

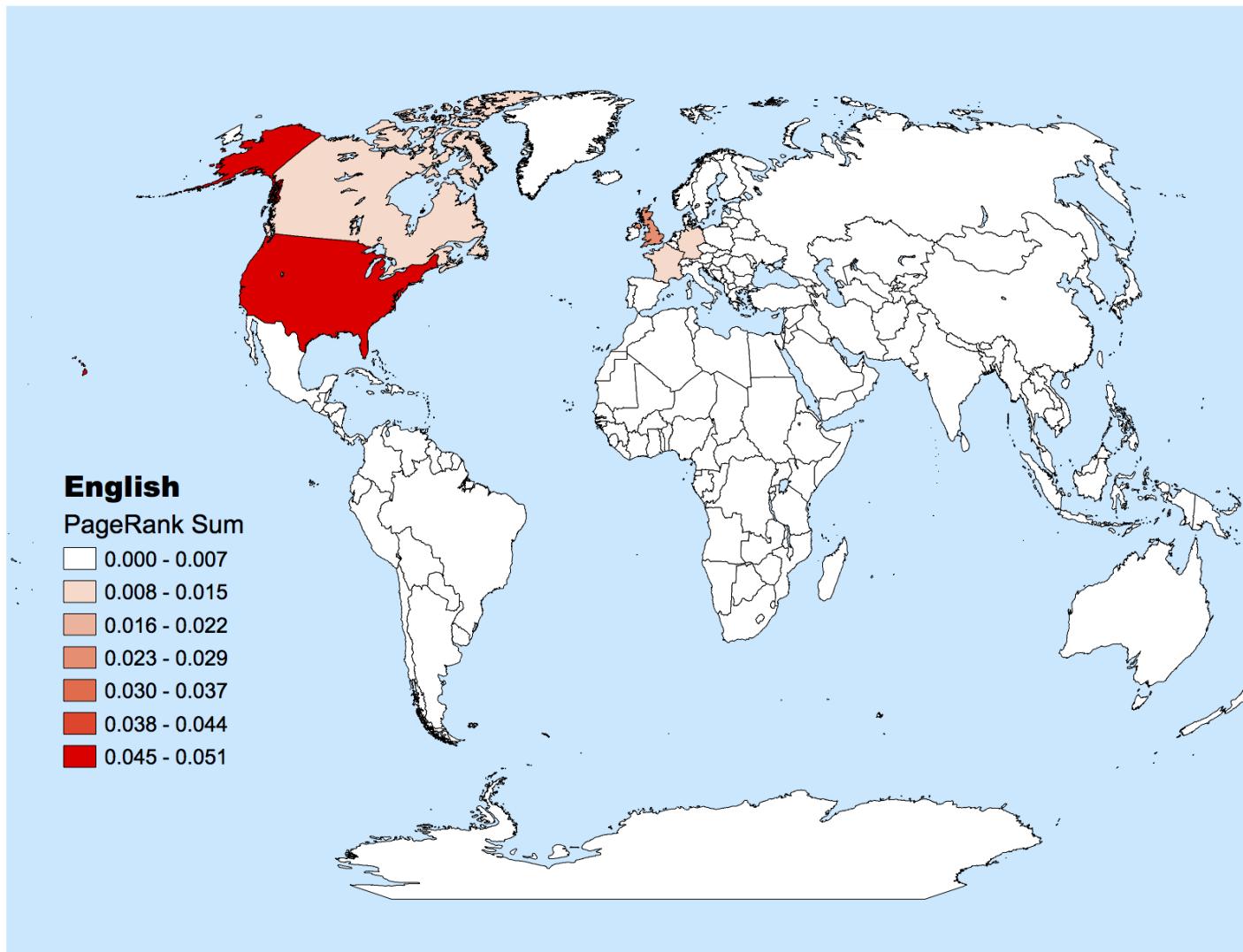


Figure 3.10-f: PageRank score sums for the English Wikipedia.

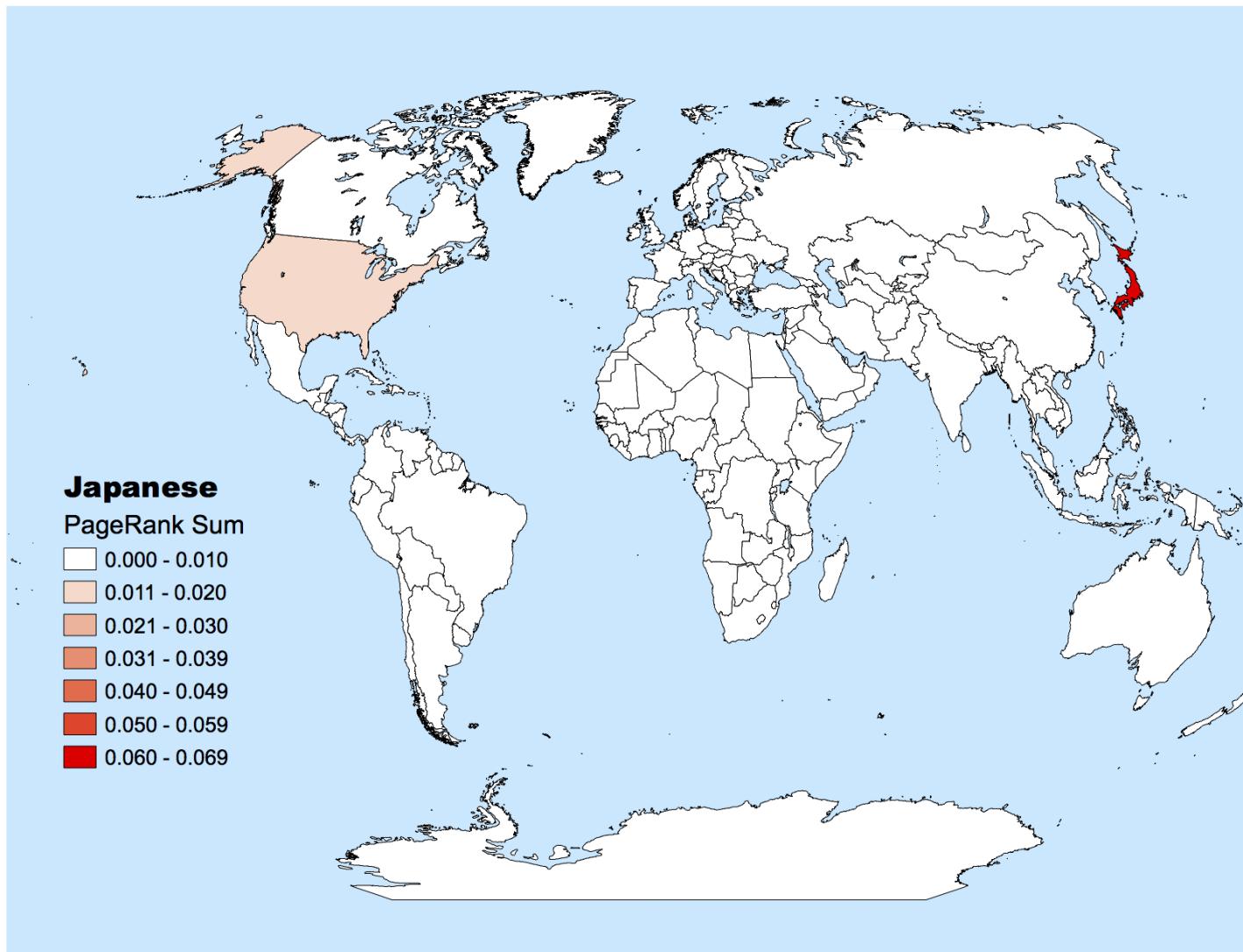


Figure 3.10-g: PageRank score sums for the Japanese Wikipedia.

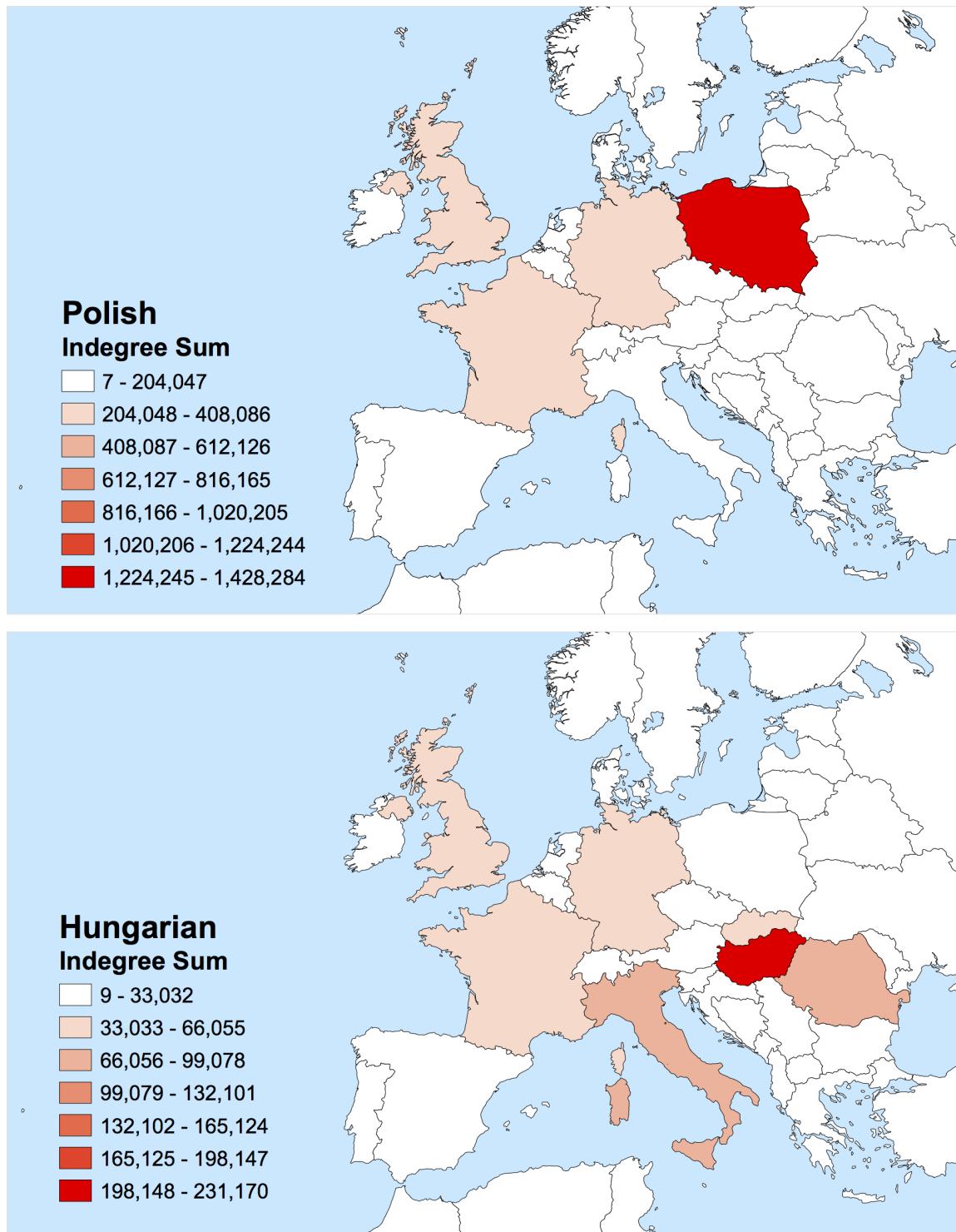


Figure 3.10-h: Indegree sums for the Polish and Hungarian Wikipedias.

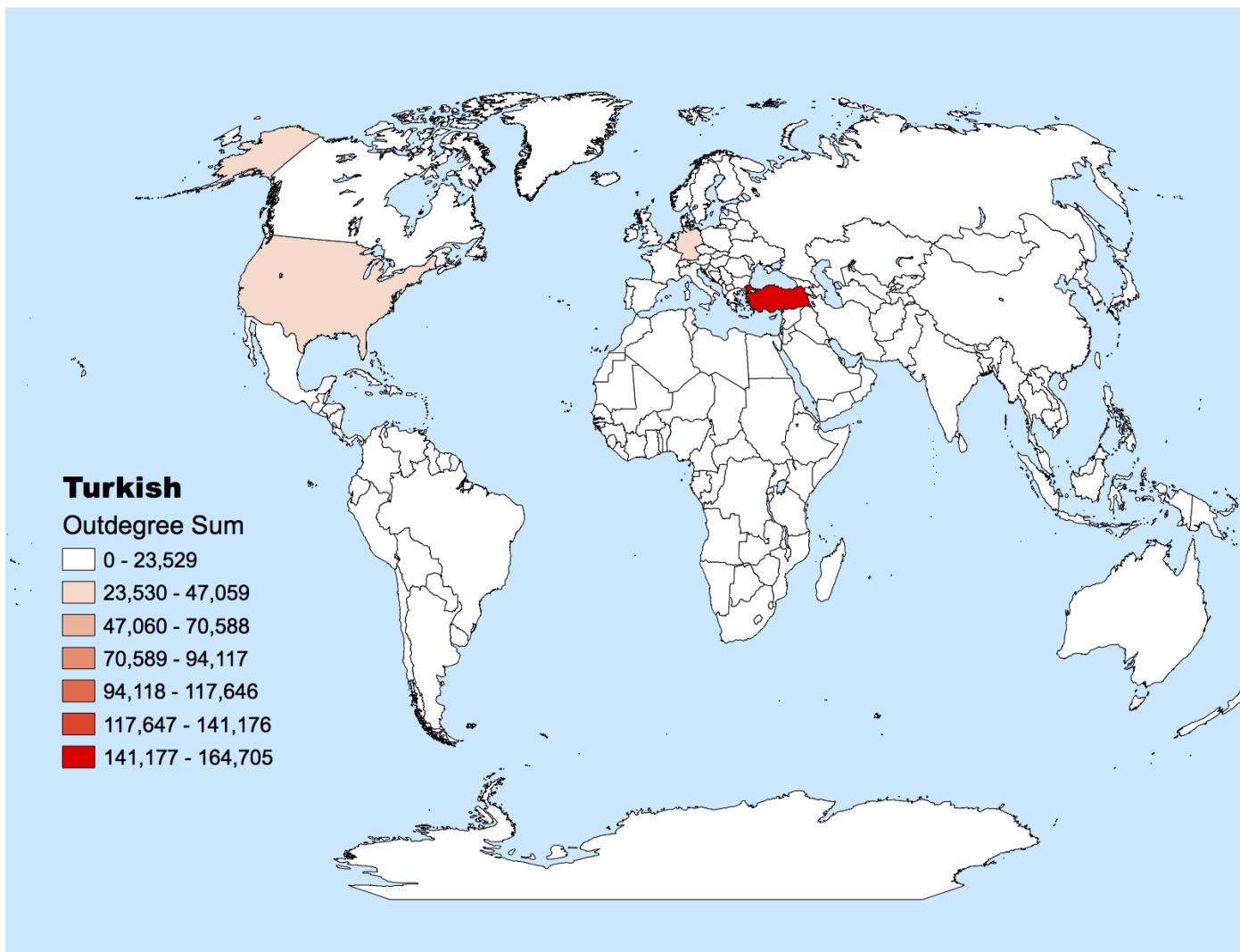
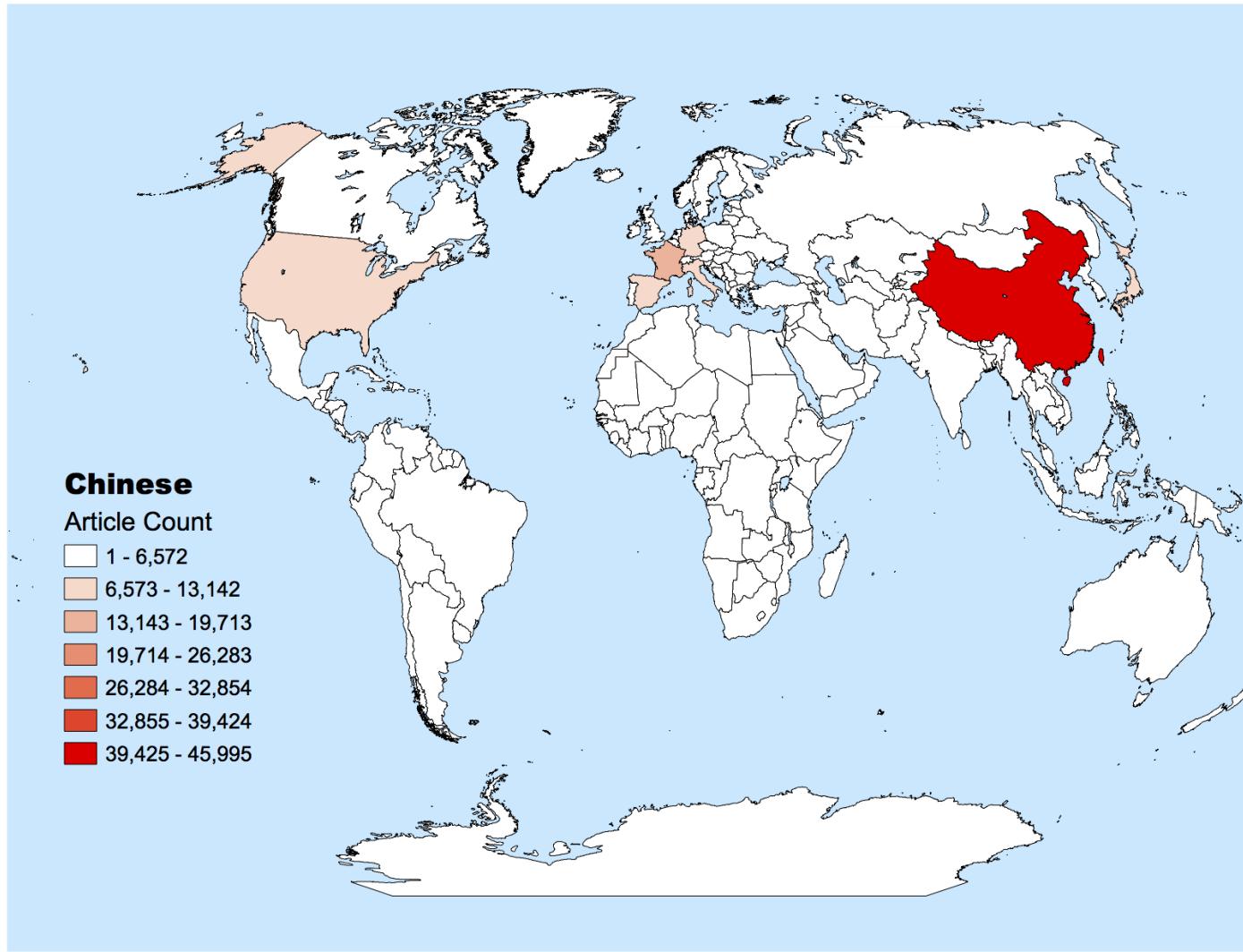


Figure 3.10-i: Outdegree sums in the Turkish Wikipedia.



*Figure 3.10-j: Article counts in the Chinese Wikipedia.*

There are a few exceptions to the general rule stated above that home cultural regions are the most prominent regions in each language edition of multilingual Wikipedia. Let us first consider the PageRank score sum and indegree sum metrics. In Portuguese and Spanish, no one home cultural region country has a higher value for these metrics than the most prominent country outside the home cultural regions (in both cases, the United States<sup>51</sup>). However, in both of these cases, there are multiple home cultural region countries, and summing the scores of these countries results in a prominence significantly greater than that of the United States.

More complicated is the case of Slovak, which is far more extreme and is the only other case where a home cultural region country is not the most prominent country in the world when it comes to PageRank score sums. As can be seen in Figure 3.10-k, although it is the second-ranked country<sup>52</sup>, Slovakia has approximately 40% of the PageRank score sum of the country with the highest PageRank score sum: France. Is it the case that a large number of Slovak editors have diligently worked to contribute content about France, both by adding articles about places in France and by integrating France-related material into the descriptions of other concepts? Investigating this situation, we found that the opposite is true: several bots have automatically created thousands of articles about France. A brief survey of these articles on the live version of the Slovak Wikipedia revealed zero manually-added content on these pages. While these types of automatically-created articles appear in all language editions including English [120] (e.g. the fencing example in Section 3.7) in no other case is the language edition-wide effect so significant. The implications of automatically created content and content automatically

---

51 There is a case to be made that the United States is close to being in the home cultural region of Spanish speakers. In the nomenclature of cultural geography, while the United States might not be in the “core” of the Spanish language-defined culture, it certainly is in the “domain.”

52 Not surprisingly, the Czech Republic is the third-ranked PageRank score sum country.

transferred from language edition to language edition is discussed below in Section 3.11.

The automatically created pages in Slovak also cause the article-level metrics to be heavily skewed towards France; France is the country with the highest outdegree sum and article count in the Slovak Wikipedia.

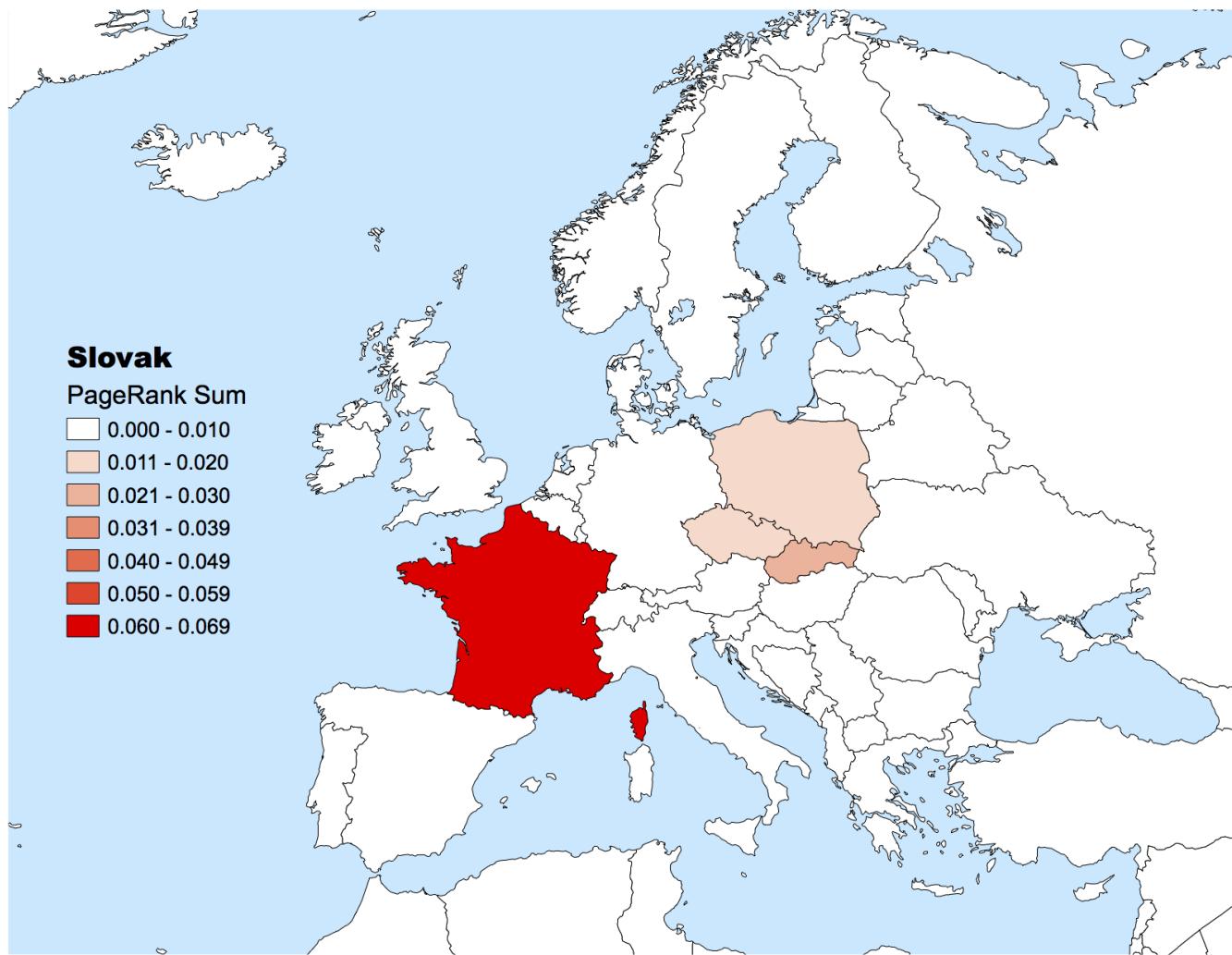


Figure 3.10-k: PageRank score sums in the Slovak Wikipedia.

In the same way that we can map prominence metrics at the country level, we can map these metrics aggregated over first-order administrative districts. At the scale of states, provinces, oblasts, and so on, there are many fewer opportunities to examine differences across language-defined cultures that do not exist at the country level as well. However, those opportunities that present themselves are significant as they disambiguate national (in the vernacular sense) culture effects from that of language-speaking communities.

Below, we show first-order administrative district indegree score sums in North America for the English Wikipedia (Figure 3.10-l) and the French Wikipedia (Figure 3.10-m). In English, the sums track population, with states and provinces such as California, New York, Florida, and Ontario having high prominence values. On the other hand, the equivalent map for the French Wikipedia breaks significantly from population trends, at least *general* population trends. When it comes to the French-speaking population, however, there is a strong correlation; Québec is by far the largest home cultural region for French speakers in North America and has close to the highest indegree score sum. In other words, Figure 3.10-m shows that according to the world knowledge in the French Wikipedia, Québec is one of the two most important administrative districts in North America, and one of two most mentioned districts across all articles. By a small margin, the highest value in Figure 3.10-m belongs to California, which is far from a French-speaking region and which rivals Québec in all the French Wikipedia prominence metrics in North America. It is important to note though that California has approximately four times the population of Québec, yet it has approximately the same indegree sum. California was quite prominent in many language editions.

Switzerland is another multilingual country whose language-defined cultures are largely divided by first-order administrative districts. The left half of Figure 3.10-n is a map that

categorizes all Swiss cantons by the language edition that has the highest relative<sup>53</sup> PageRank score sum. By comparing this map to the righthand side of Figure 3.10-n, which shows a map based on official Swiss government data, one can see that the prominence metrics in these language editions almost *exactly* reflect the home cultural regions of their corresponding language-defined communities. The only exceptions occur in cantons that are bilingual and whose language-defined culture boundaries occur at the sub-canton level. These boundaries would be undetectable at the current granularity of our analysis no matter how strong the signal of cultural context.

Figure 3.10-o shows a similar phenomenon occurring in Belgium, which is a country split between French speakers (in the South) and Dutch speakers (in the North). In this case, all provinces have a higher relative PageRank score sum in Dutch, but the extent to which they are higher varies nearly perfectly with the dominant language in each province. In other words, the Dutch Wikipedia considers all of Belgium to be more important than the French Wikipedia does, but the Dutch language edition places more importance on the Dutch-speaking regions than the French-speaking ones.

---

53 PageRank Sum scores were normalized within each language edition before comparing across language editions. This was done to remove the effect of the prominence of geographic concepts generally.

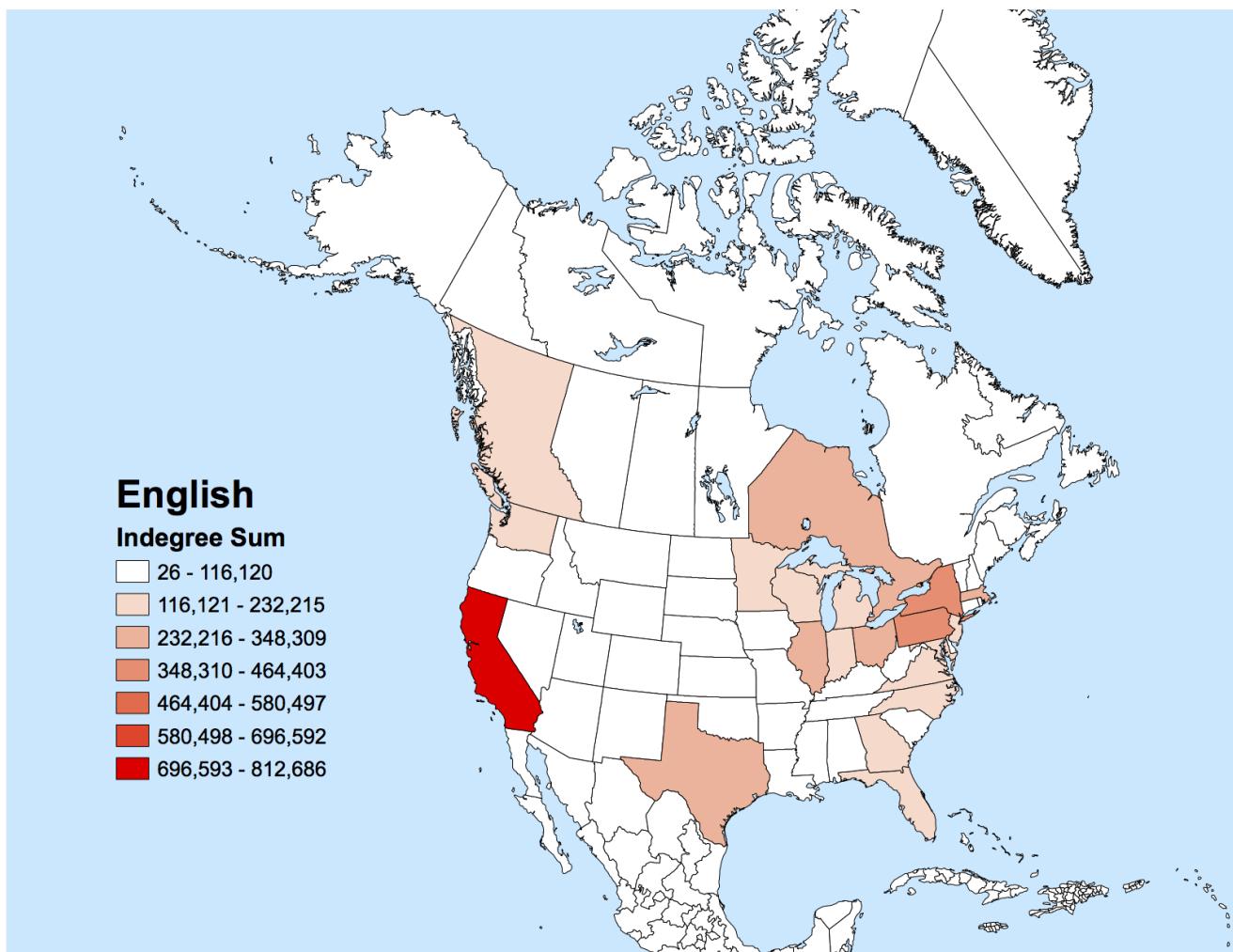


Figure 3.10-l: English Wikipedia indegree sums in North America.

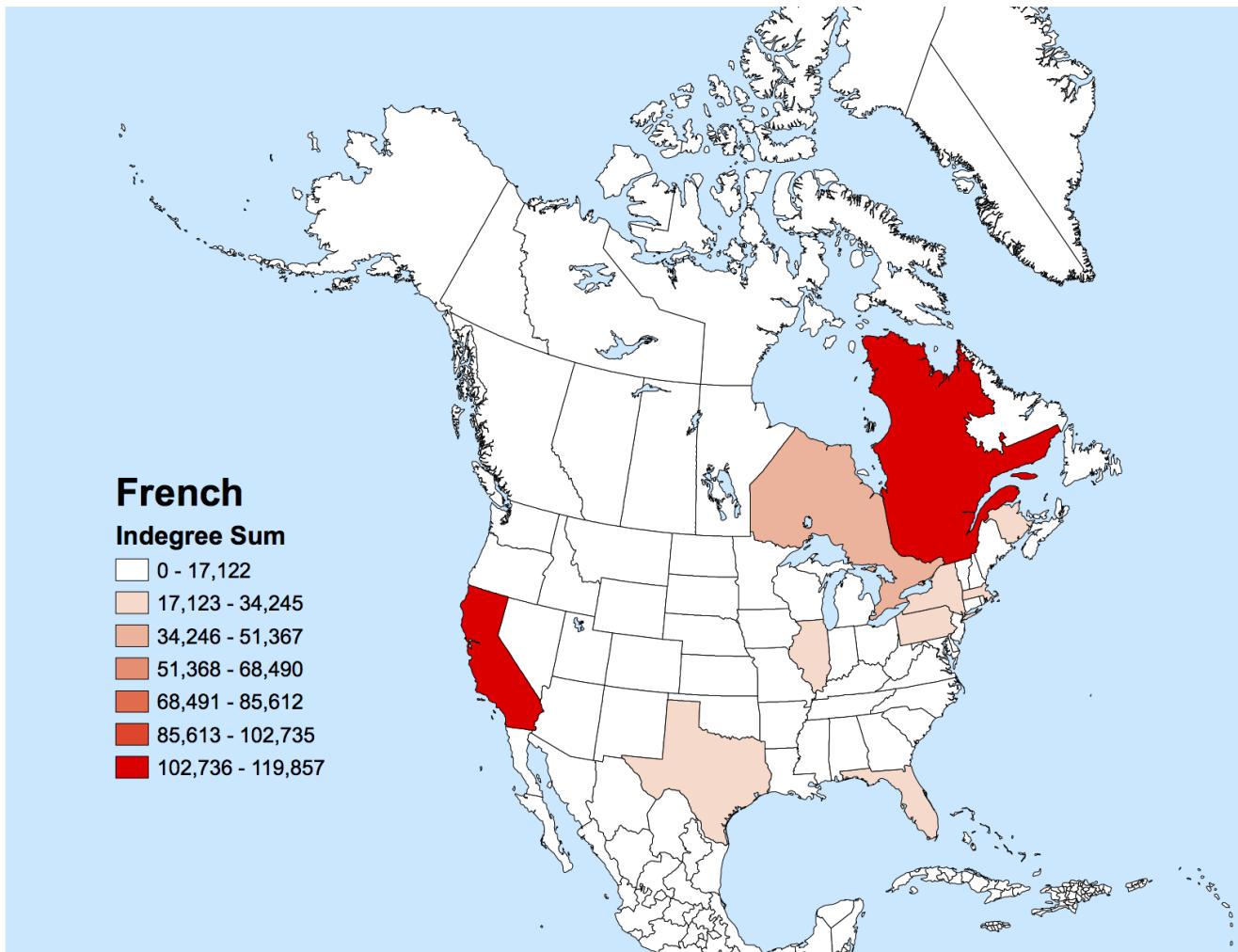
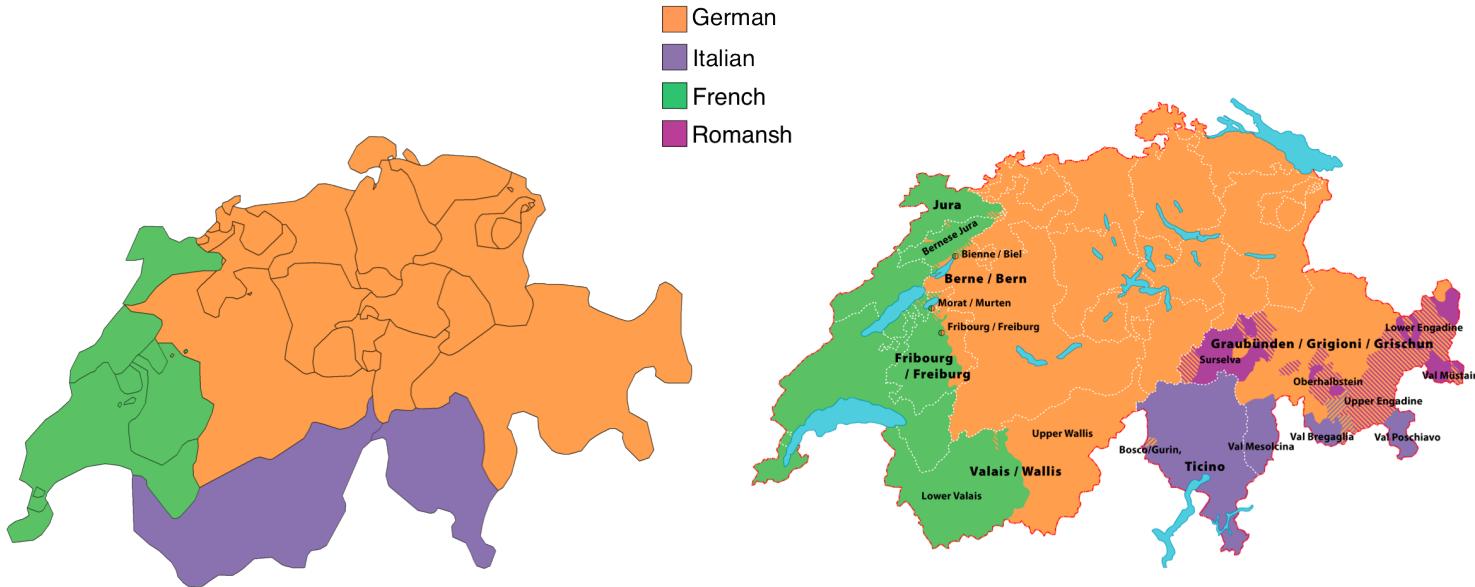


Figure 3.10-m: French Wikipedia indegree sums in North America.



**Map Based on Cultural Context in UGC**

Cantons categorized by the language edition in which they have the highest relative PageRank score sum

**Map Based on Official Government Data**

[http://upload.wikimedia.org/wikipedia/commons/9/9f/Map\\_Languages\\_CH.png](http://upload.wikimedia.org/wikipedia/commons/9/9f/Map_Languages_CH.png)

*Figure 3.10-n: The left side of the figure shows a map of Swiss cantons categorized by the language edition with the highest relative PageRank score sum. The right side of the figure is a map based on official data from the Swiss government. The maps are nearly identical, save one canton in the southwest. The Romansh Wikipedia is not considered in this study, so it cannot appear on the left side of this figure.*

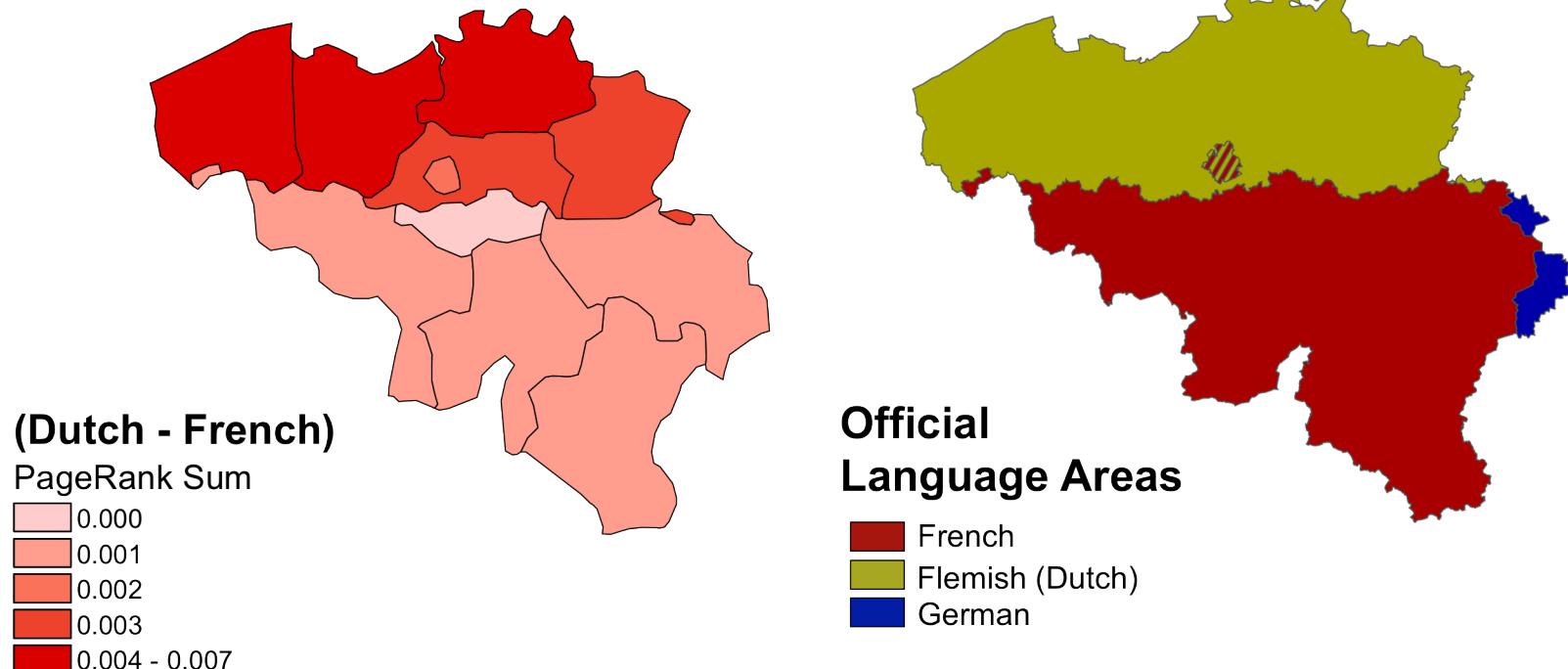


Figure 3.10-o: The left side of the figure shows the difference between the Dutch and French Wikipedia PageRank score sums in Belgium (natural breaks classification scheme). The right side of the figure is a map of the actual dominant languages in Belgium as determined by official Belgian “language areas.”\* While Dutch always has a higher relative PageRank score sum than French, the degree to which it is higher varies almost exactly with the dominant language.

\* Source: <http://en.wikipedia.org/wiki/File:BelgieGemeenschappenkaart.png>

### 3.10.2.2 The Self-Focus Bias Ratio

In this section, we leverage an index we developed called the *self-focus bias ratio* [80] to summarize the amount of cultural contextualization in each language edition in a single value, and to compare these values across language editions. In the framework of the research in this chapter, “self-focus” is merely synonym for cultural contextualization; we use the original terminology here for consistency with our published work.

The self-focus bias ratio (*SFBR*) is a simple metric defined as follows:

$$SFBR(W_l) = \frac{\max(C_{L=l})}{\max(C_{L \neq l})}$$

where  $W_l$  is the Wikipedia of language  $l$ ,  $\max(C_{L=l})$  is the maximum prominence score of a country inside a home cultural region of speakers of  $l$ ,  $\max(C_{L \neq l})$  is the maximum prominence score of a country outside of  $l$ 's home cultural regions. Put more simply, the self-focus bias ratio for a language edition is the ratio of the maximum prominence of a home cultural region country over the maximum prominence of a country not in a home cultural region.

Let us consider the case of the PageRank score sums in the Norwegian Wikipedia, depicted cartographically in Figure 3.10-c. Norway has a PageRank score sum of 0.04589 in the Norwegian Wikipedia, which as noted above is the Norwegian Wikipedia's highest PageRank score sum in the world. The maximum PageRank score sum of a country not in a home cultural region of Norwegian speakers (i.e. not in Norway) in the Norwegian Wikipedia is 0.02435, which belongs to the United States. In this case, we would say that the self-focus bias ratio for PageRank score sums for the Norwegian Wikipedia is  $0.04589 / 0.02435 = 1.88$ .

One of the primary benefits of self-focus bias ratios is that they are easy to interpret. For

instance, from the above example we can say that, according to the Norwegian Wikipedia, Norway is 1.88 times more important than any other country in the world. We can also say that Norway is discussed 1.88 times more often than any other country (in a PageRank-weighted sense) in the Norwegian Wikipedia.

The *SFBR* can be calculated for any of the prominence metrics used in this study. For instance, the indegree *SFBR* for the Norwegian Wikipedia, which is 1.91, provides additional confirmation of the above results. Norwegian's article count *SFBR* is 1.97, indicating that Norway has almost twice the number of articles as any other country in the world in the Norwegian Wikipedia. Finally, at 2.67, the outdegree sum *SFBR* for the Norwegian Wikipedia is somewhat higher, meaning that almost 2.7x as many links originate in articles about places in Norway than in those about places in any other country in the world.

Table 3.10-a shows all 25 language editions' self-focus bias ratios for all four prominence metrics. The dominant property of Table 3.10-a is that the vast majority of SFBRs are above one. This is an additional stark confirmation that each language edition reflects the cultural contexts of its contributors. As noted above, an SFBR greater than one indicates that at least one country in a home cultural region is the *most prominent country in the world* according to the corresponding prominence metric. Some of the most significant results in Table 3.10-a include:

- According to the Japanese Wikipedia and the indegree sum prominence metric, Japan is almost seven times more important to all of encyclopedic world knowledge than any other country in the world. Places in Japan are also mentioned almost 7x more often than places in any other country.
- The same is true in the English Wikipedia (with regard to the United States). The Indonesian Wikipedia has the third highest indegree *SFBR*, with Indonesia being represented as 4.8x more important than any other

country in the world.

- The Czech Republic, China, the United States, Germany, Indonesia, Japan, and Russia are more than three times more important than any other country in the world according to the respective language editions and the PageRank score *SFBR*.
- There are almost seven times as many articles about places in the Czech Republic than about any other country in the world in the Czech Wikipedia. The same is true of Germany (in the German Wikipedia) and Japan (in the Japanese Wikipedia) at rates of 6.3x and 5.8x respectively.
- The highest *SFBR* in the table is the Japanese *SFBR* for outdegree sums, which is 10.698. This means that there are 10.698 times more aggregate outlinks from articles about places in Japan than any other country in the world in the Japanese language edition.

Even in the 17% of cases when the *SFBR* is less than one, it is often in the very near vicinity of one, with a home cultural region country a close second in global prominence. For instance, although prominence in the Portuguese and Spanish Wikipedias is spread out over several home cultural region countries, the *SFBR* is frequently very close one, even in the centrality prominence sums. For instance, for indegree sums in Spanish, Spain has only slightly fewer inlinks than the United States. The same is true with regard to the Portuguese Wikipedia and PageRank score sums (with Brazil replacing Spain), and the indegree *SFBR* for Portuguese is greater than one.

There are a number of additional significant language-edition specific findings in Table 3.10-a. One of the most important is that the Japanese Wikipedia has the highest *SFBR* in three out of four of the prominence metrics, including both language-edition wide metrics (indegree

sum and PageRank sum). Throughout this chapter, we have seen that Japanese is one of the most unique language editions in our 25-language edition dataset. The results in Table 3.10-a show that the extensive cultural contextualization of the encyclopedic world knowledge in the Japanese Wikipedia is at least one cause of this trend.

With regard to the non-Spanish and non-Portuguese *SFBRs* that are less than one, the most significant case is that of the Slovak Wikipedia. Not surprisingly, due to the automated processes that created so much information about France in the Slovak Wikipedia, France has significantly more prominence than Slovakia in Slovakia's native language edition. This is a point we return to below.

The difference between the language edition-wide and article-level *SFBRs* in the Indonesian and Korean Wikipedias is also quite interesting, with the language-edition-wide *SFBRs* being significantly higher. The gap between the *SFBRs* tells us that a relatively small number of articles in the respective home cultural regions are receiving large numbers of inlinks from articles throughout each language edition, and in the case of PageRank scores, these are inlinks from important articles. These results also tell us that there is a mismatch between the number of articles in the home cultural regions and the importance of these regions according to the WAGs of the language editions. In the Korean Wikipedia, both China and Japan have more articles than South Korea. In the Indonesian Wikipedia, the top three article counts belong to Italy, France, and Germany, likely due to the same phenomenon as the Slovak/France relationship highlighted above.

Other *SFBR* trends can be seen by examining more closely the values visualized in the preceding section. One relatively significant trend is the tendency for the United States to be the second- or third-most prominent country according to all of the prominence sum metrics. Table

3.10-b shows the PageRank score sum rank of the United States in each language edition. The average rank is 2.52.

Thus far we have paid attention to the top of the *SFBR* range. However, there are also important results present at the bottom. Namely, in all language editions, Sub-Saharan Africa is one of the, if not the *the*, least prominent major region of the world. Table 3.10-c shows the most prominent country in Sub-Saharan Africa in each of the language editions according to PageRank score sums, as well as that country's PageRank score sums rank. The mean rank of the most prominent Sub-Saharan African country is only 62.8. This means that, on average, 62 countries are encoded as more important to world knowledge than all countries in Sub-Saharan Africa (and 62 countries are mentioned more often, in a weighted sense). This problem is exacerbated by the fact that no language-defined community that has a cultural core [40] in Sub-Saharan Africa has a language edition with more than about 30,000 articles, which is the number in the Yoruba Wikipedia<sup>54</sup>. The Yoruba Wikipedia is the 70<sup>th</sup> largest language edition, with language editions like Latin and Welsh having more articles.

The implications here are significant. In aggregate, it is clear that Sub-Saharan Africa is the most underrepresented major region of the world in multilingual Wikipedia. Africa, as in so many other domains, gets the short end of the stick here, likely due to both a dearth of links to articles that exist about Africa, as well as a limited number of such articles. Like is often the case with economics and politics, this study shows that Africa is unfortunately on the periphery of Wikipedia.

---

54 And these articles were created by a bot [209].

Language Ed.	Indegree SFBR	PageRank SFBR	Count SFBR	Outdegree SFBR
Catalan	1.119	2.170	1.689	1.163
Chinese	3.308	3.226	2.943	1.916
Czech	4.164	3.762	<b>6.757</b>	5.058
Danish	2.169	2.090	2.724	2.662
Dutch	1.077	1.094	0.885	1.140
English	<b>6.902</b>	<b>5.174</b>	3.739	6.618
Finnish	1.515	1.619	1.737	2.510
French	3.238	2.297	3.261	<b>6.953</b>
German	3.881	3.639	<b>6.356</b>	<b>6.836</b>
Hebrew	1.197*	1.259*	2.533	2.157
Hungarian	2.543	1.907	0.876	1.533
Indonesian	<b>4.831</b>	<b>4.241</b>	0.651	0.959
Italian	2.714	1.768	1.423	6.585
Japanese	<b>6.927</b>	<b>5.288</b>	<b>5.873</b>	<b>10.698</b>
Korean	1.440	1.330	0.354	0.613
Norwegian	1.906	1.884	1.978	2.668
Polish	3.434	2.537	3.761	6.732
Portuguese	1.162	0.993	0.378	0.929
Romanian	1.523	1.246	1.195	1.876
Russian	4.093	3.638	1.781	3.462
Slovak	<b>0.520</b>	<b>0.404</b>	0.392	<b>0.525</b>
Spanish	<b>0.975</b>	<b>0.796</b>	<b>0.386</b>	<b>0.582</b>
Swedish	1.639	1.418	0.996	1.566
Turkish	3.103	2.370	2.629	4.139
Ukrainian	<b>0.954</b>	1.291	1.961	1.445

Table 3.10-a: The four self-focus bias ratios for all 25 language editions. SFBRs greater than or equal to 1.0 are depicted in green, with those less than 1.0 are in red. The top and bottom three values for each SFBR are in bold.

\* The country dataset we use excludes the West Bank and Gaza Strip from Israel (not a choice we made). Adding these to Israel's total increases the SFBR by a somewhat significant margin. For instance, the indegree SFBR jumps to over 1.5, more than 0.3 higher than the original value.

Language Edition	US Rank	Higher-ranked Countries	Language Edition	US Rank	Higher-ranked Countries
Catalan	3	Spain, France	Korean	3	South Korea, Japan
Czech	2	Czech Republic	Dutch	3	Netherlands, France
Danish	3	Denmark, Germany	Norwegian	2	United States
German	2	Germany	Polish	3	Poland, France
English	1	-	Portuguese	1	-
Spanish	1	-	Romanian	4	Romania, France, Germany
Finnish	2	Finland	Russian	3	Russia, Ukraine
French	1	France	Slovak	4	France, Slovakia, Czech Republic
Hebrew	2	Israel	Swedish	3	Sweden, France
Hungarian	2	Hungary	Turkish	2	Turkey
Indonesian	2	Indonesia	Ukraine	6	Ukraine, France, Italy, Romania, Russia, United States
Italian	3	Italy, France	Chinese	3	China, France
Japanese	2	Japan	AVERAGE	2.52	-

Table 3.10-b : The PageRank score sum rank of the United States and the countries that rank higher than the United States in each language edition.

Language Edition	Top Sub-Saharan Rank	Country	Language Edition	Top Sub-Saharan Rank	Country
Catalan	61	Ethiopia	Korean	64	Ethiopia
Czech	75	Ethiopia	Dutch	75	Ethiopia
Danish	64	Chad	Norwegian	64	Ethiopia
German	61	Namibia	Polish	78	Ethiopia
English	67	Ethiopia	Portuguese	35	Angola
Spanish	77	Ethiopia	Romanian	60	Ethiopia
Finnish	64	Ethiopia	Russian	76	Ethiopia
French	39	Ethiopia	Slovak	62	Ethiopia
Hebrew	44	Ethiopia	Swedish	57	Kenya
Hungarian	55	Ethiopia	Turkish	67	Ethiopia
Indonesian	73	Nigeria	Ukraine	78	Ethiopia
Italian	63	Mali	Chinese	63	Ethiopia
Japanese	74	Ethiopia	AVERAGE	63.8	-

Table 3.10-c: The highest-ranking country in Sub-Saharan Africa according to PageRank Score sums, along with that country's rank.

### 3.10.3 Content vs. Consumption Self-Focus Bias

In the above section, we focused on four prominence metrics that were all derived from the *content* of multilingual Wikipedia. In this section, we switch gears and introduce a prominence metric based on the *consumption* of that content: *page view score sums*. Page view score sums work identically to indegree, PageRank, article count, and outdegree score sums, with the exception that they measure the number of views received by concepts in each country or first-order administrative district. As such, if a country (or district) has a page view sum in a given language edition that is 3x higher than that of another country (or district), the readers of that language edition have visited articles about the first country three times more than those about the second country. Following the discussion in Section 3.8, page view sums can thus be thought of as the importance of a given country or administrative district as determined by the behavior of readers of each language edition.

In this section, we consider all page views that occurred during the 2010 – 2012 data collection period. For instance, Figure 3.10-p shows the 2010 – 2012 page view score sums for the German and Swedish Wikipedias. Note that in Figure 3.10-p there is only one country in Europe that is not in the bottom bin in either of the maps (Germany in the Swedish Wikipedia map). This is quite different than was the case in the PageRank score sums maps for these language editions above. A similar pattern can be found in Figures 3.10-q and 3.10-r, which show page view sums for the Chinese and Japanese Wikipedias respectively.

Language Edition	Page View SFBR	% PageRank SFBR	Language Edition	Page View SFBR	% PageRank SFBR
Catalan	4.46	205%	Japanese	11.23	212%
Chinese	4.30	133%	Korean	1.54	116%
Czech	6.78	180%	Norwegian	3.31	176%
Danish	2.86	137%	Polish	6.70	264%
Dutch	2.96	270%	Portuguese	3.60	363%
English	9.27	179%	Romanian	3.81	306%
French	4.06	177%	Russian	4.47	123%
Finnish	3.29	203%	Slovak	4.50	1113%
German	5.08	177%	Spanish	1.55	194%
Hebrew	3.63	289%	Swedish	3.11	219%
Hungarian	3.47	182%	Turkish	4.38	184%
Indonesian	4.51	106%	Ukrainian	6.26	485%
Italian	5.05	286%	AVERAGE	4.57	250%

Table 3.10-d: The page view self-focus bias ratio for all 25 language editions and the multiple of this SFBR relative to the language edition's self-focus bias ratio.

This raises an important question: do the readers of Wikipedia reflect much more self-focus bias than the actual content of Wikipedia? That is, is the cultural contextualization in the page views dataset greater than that in the content of Wikipedia? Table 3.10-d reveals that, overwhelmingly, this is indeed the case. For no language edition was the page view SFBR less than the PageRank SFBR and on average, the page view SFBR was *2.5 times higher* than the PageRank SFBR. This means that the importance of home cultural regions as determined by Wikipedia readers is 250% the importance of home cultural regions as determined by the content of Wikipedia. The equivalent numbers for indegree SFBR, outdegree SFBR, and article count SFBR were 2.3x, 2.2x, and 3.2x respectively.

For one language edition in particular, the difference between the page view SFBR and the PageRank SFBR was particularly large: Slovak. Readers of the Slovak Wikipedia find Slovakia

to be over 11x as important as the content of the language edition does. This means that the automated process that created thousands of articles about French-speaking places in the Slovak Wikipedia created a drastic mismatch between the content of the encyclopedia and the interests of the readers of the encyclopedia. Figure 3.10-s shows that despite the fact that there are 3.9 times as many articles about France than Slovakia in the Slovak Wikipedia, Slovakia receives 4.5 times as many page views. In fact, France is not even the second-most viewed country in the Slovak Wikipedia; that rank goes, not surprisingly, to the Czech Wikipedia. We further discuss the tension between content and consumption in Section 3.11.

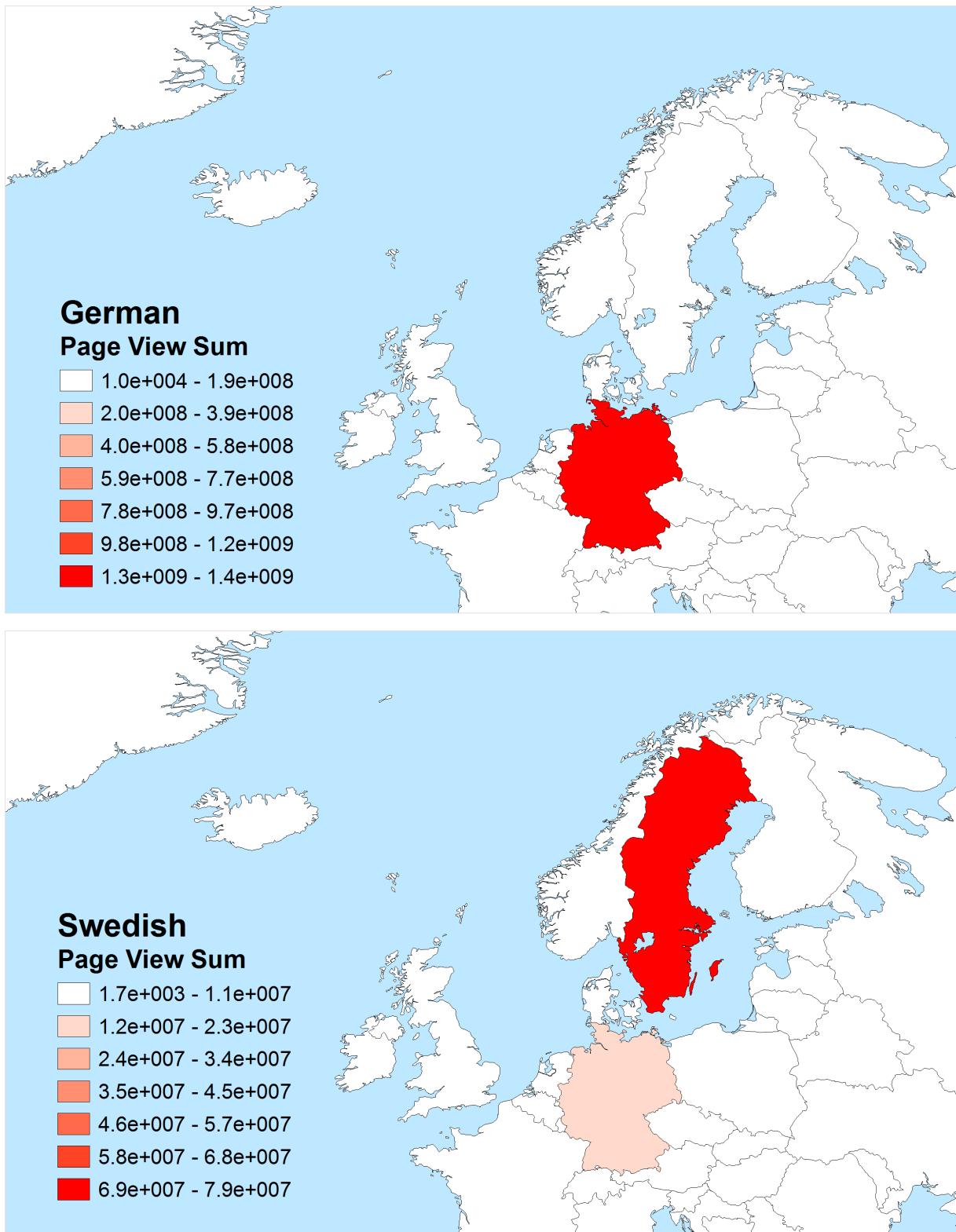


Figure 3.10-p: Page view score sums for the German and Swedish Wikipedia.

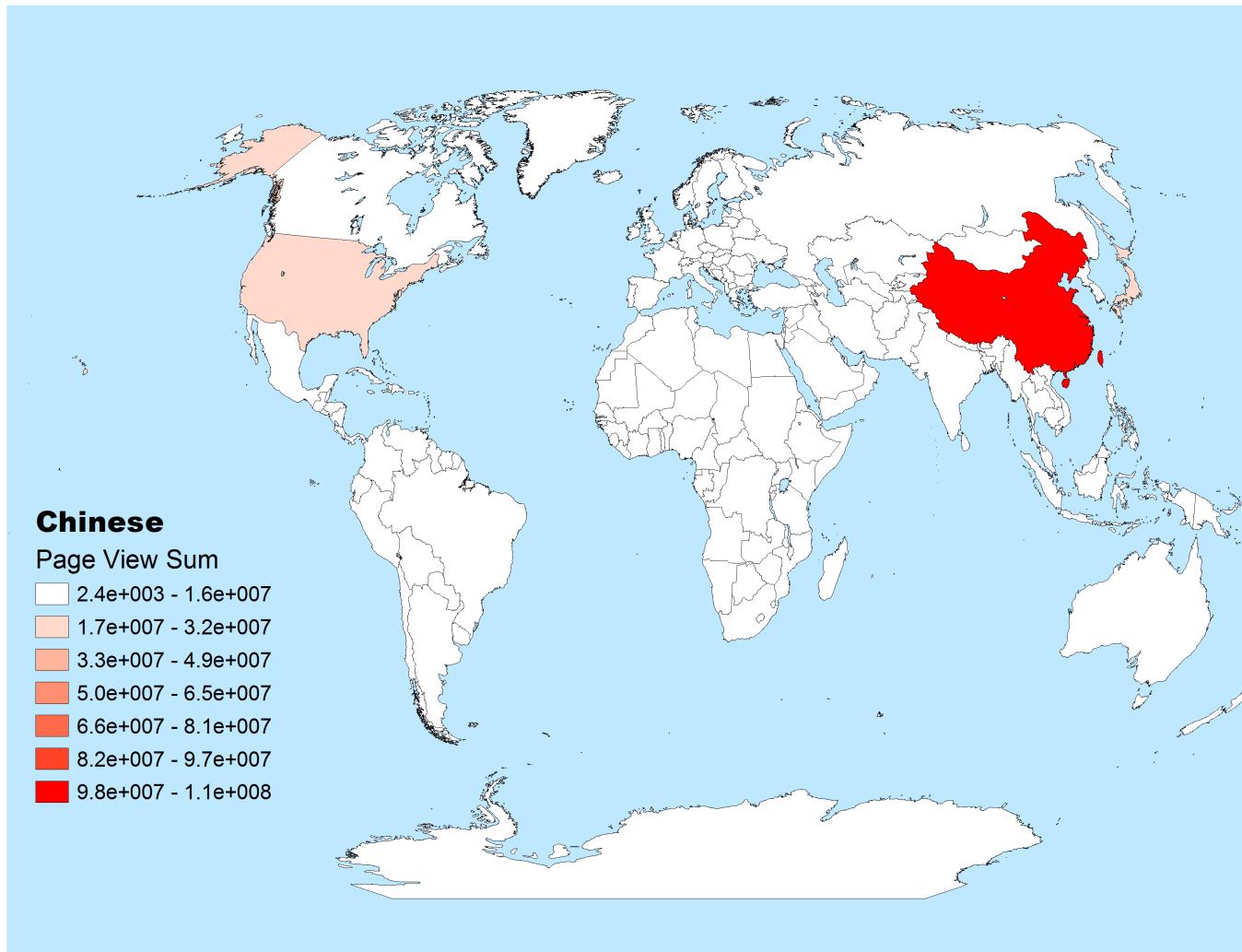


Figure 3.10-q: Page view sums for the Chinese Wikipedia.

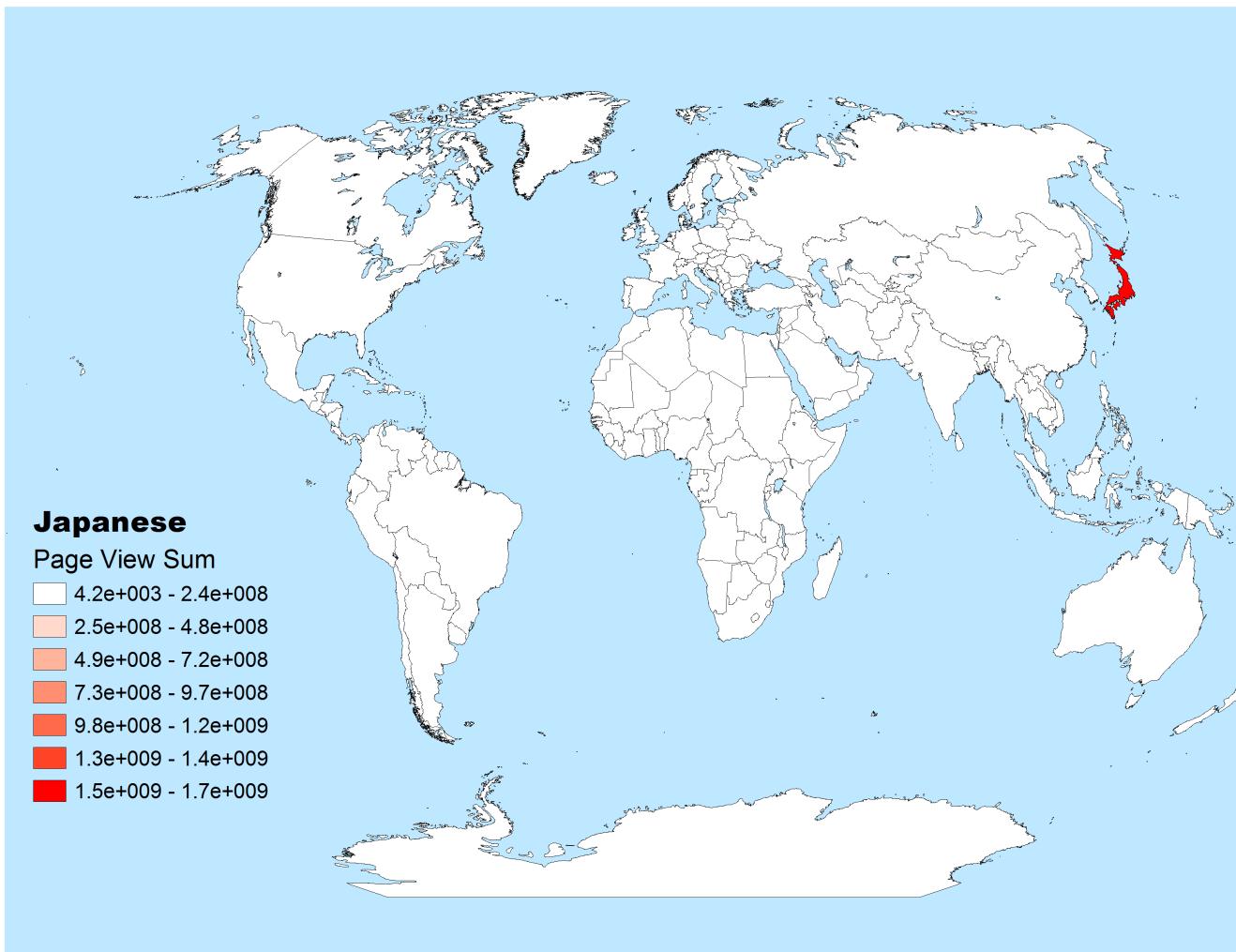


Figure 3.10-r: Page view sums for the Japanese Wikipedia.

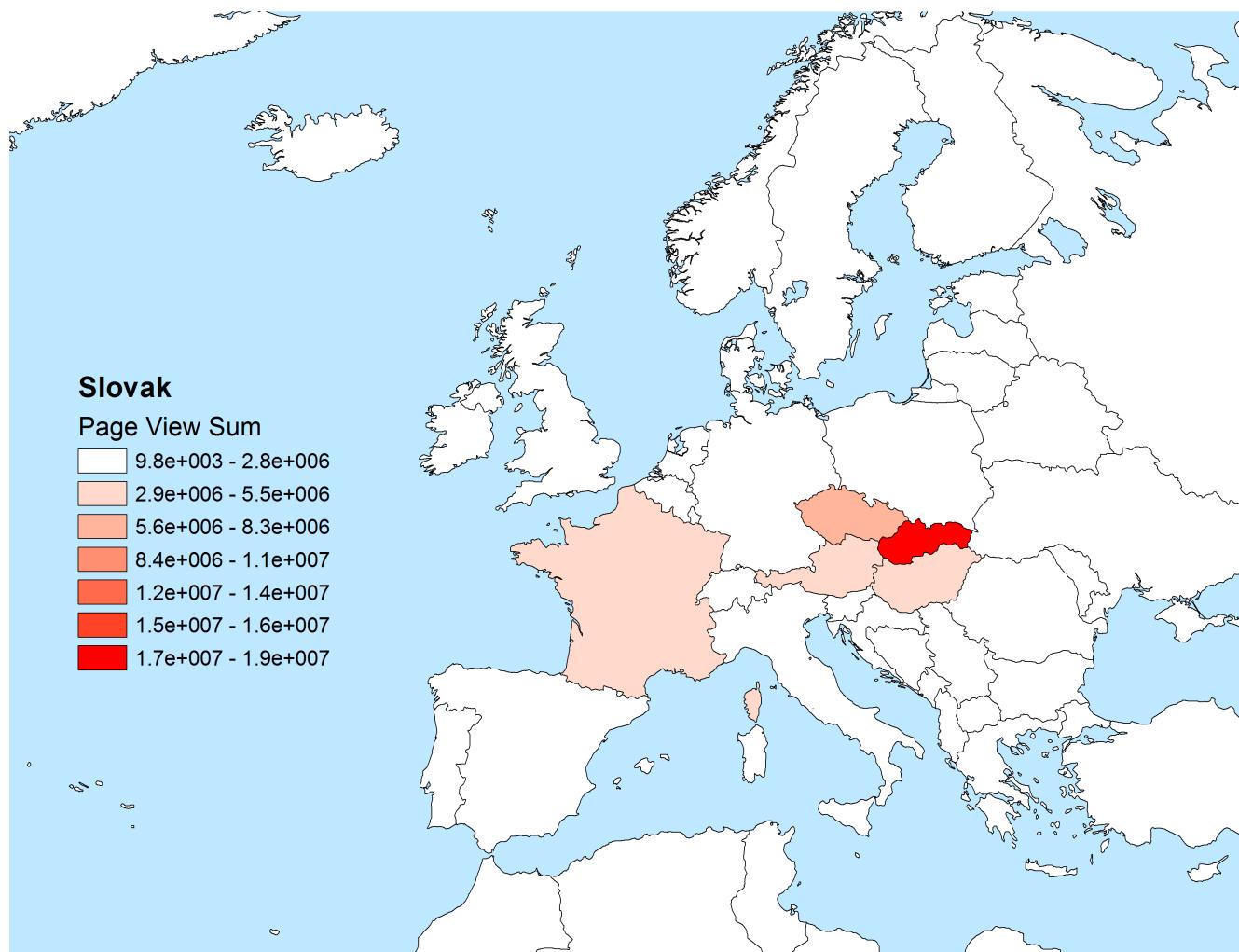


Figure 3.10-s: Page view sums in the Slovak Wikipedia.

In addition to examining the differences between page view cultural contextualization and content cultural contextualization at the scale of entire language editions, we can also do so on a country-by-country and administrative district-by-administrative district basis. Performing this analysis can reveal areas of the world that are the target of a great deal of reader interest in a given language edition but are not given a great deal of importance by the content of the language edition. For instance, Figure 3.10-t shows a map of the page view sum / PageRank sum ratio in the Spanish Wikipedia over the entire world. It is clear in the figure that there is a great deal more interest in Latin American topics than importance placed on these topics by the content of the Spanish Wikipedia. Recall that PageRank sums can also be interpreted as the extent to which central articles discuss each country/district. Under this interpretation, Figure 3.10-t reveals that reader interest in Latin American topics is not met with the same depth of content as it is for places like the United States and to a certain degree Spain. If we assume that page views imply demand for more content, then we can say that the area of the world with the least amount of content relative to demand in the Spanish Wikipedia is Latin America. Figure 3.10-t shows that the Spanish Wikipedia has something of a Spain bias in this respect, a bias that is not shared by the readers of the Spanish Wikipedia.

Figure 3.10-v is a map of the same ratio in the Slovak Wikipedia. Note that the importance placed upon Slovakia as determined by reader interest is far greater than the importance placed upon Slovakia by the content of the Wikipedia. In fact, the Slovak language edition page view sum / PageRank sum ratio is far greater in Slovakia than anywhere else in entire content of Europe. Note also that the reverse is true for France. The extensive amount of automated content about France in the Slovak language edition is decidedly unpopular relative to the amount of content that was created.

Another interesting trend we noticed when examining the spatial distribution of page view sum to PageRank sum ratios is a high demand for content about the Middle East relative to the amount of available information. This was a trend we saw in many language editions, including Spanish (see Figure 3.10-t). The most likely reason for the trend is the fact that the Arab Spring has occurred within our page view sampling window<sup>55</sup>.

The final major pattern we noticed in our analysis of page view sum / PageRank ratios is that in every language edition – even the ones with the most content self-focus – there was a relatively high ratio for home cultural region countries. We certainly saw this in Figure 3.10-t; even Spain, which has a lower ratio than most Latin American countries still had a relatively high ratio when compared to the rest of the world. Figures 3.10-u and 3.10-w, which show page view / PageRank score ratios in the Catalan and Russian Wikipedias, respectively, depict a similar phenomenon.

Before closing, it is important to note that while we have mostly focused on page view sums *in relation to* content prominence sums, the *absolute* page view sum values are also quite informative. For instance, in Table 3.10-d, we see that readers of the Japanese Wikipedia access information about Japan at a rate 11.23x greater than they access information about any other country in the world. The same is true of English and the United States at a rate of 9.27x. The language-defined culture that displays the least self-interest is Korean, whose members access information about Korea only 1.54x more often than they do about the next most-popular country (Japan). These differences in the page view *SFBR* represents a good candidate for further study.

---

<sup>55</sup> Note that this was almost certainly not the case for the United Arab Emirates (UAE). The UAE had one of the smallest concept counts beyond our 50-concept threshold, resulting in it being something of an outlier.

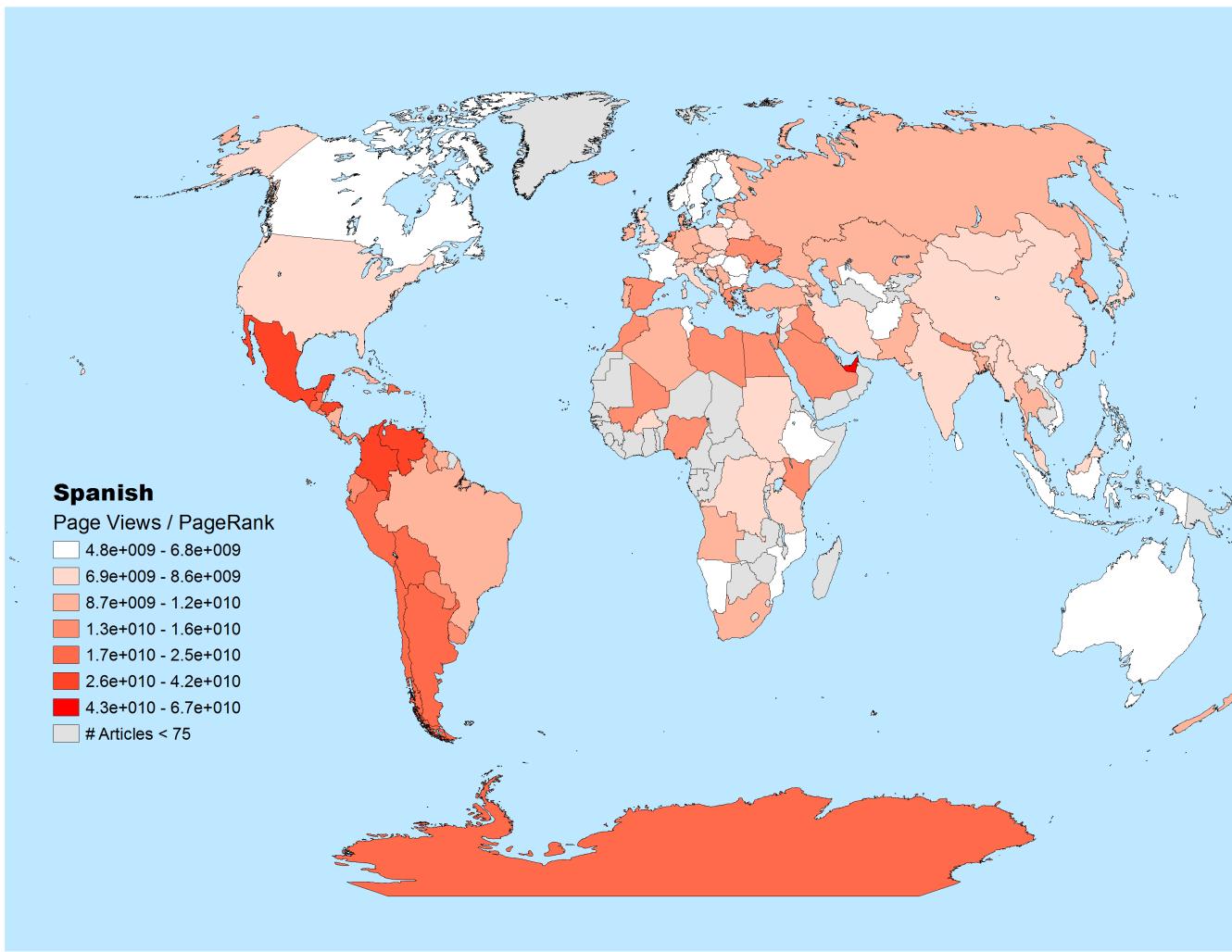


Figure 3.10-t: The ratio of page view sum to PageRank sum in the Spanish Wikipedia.

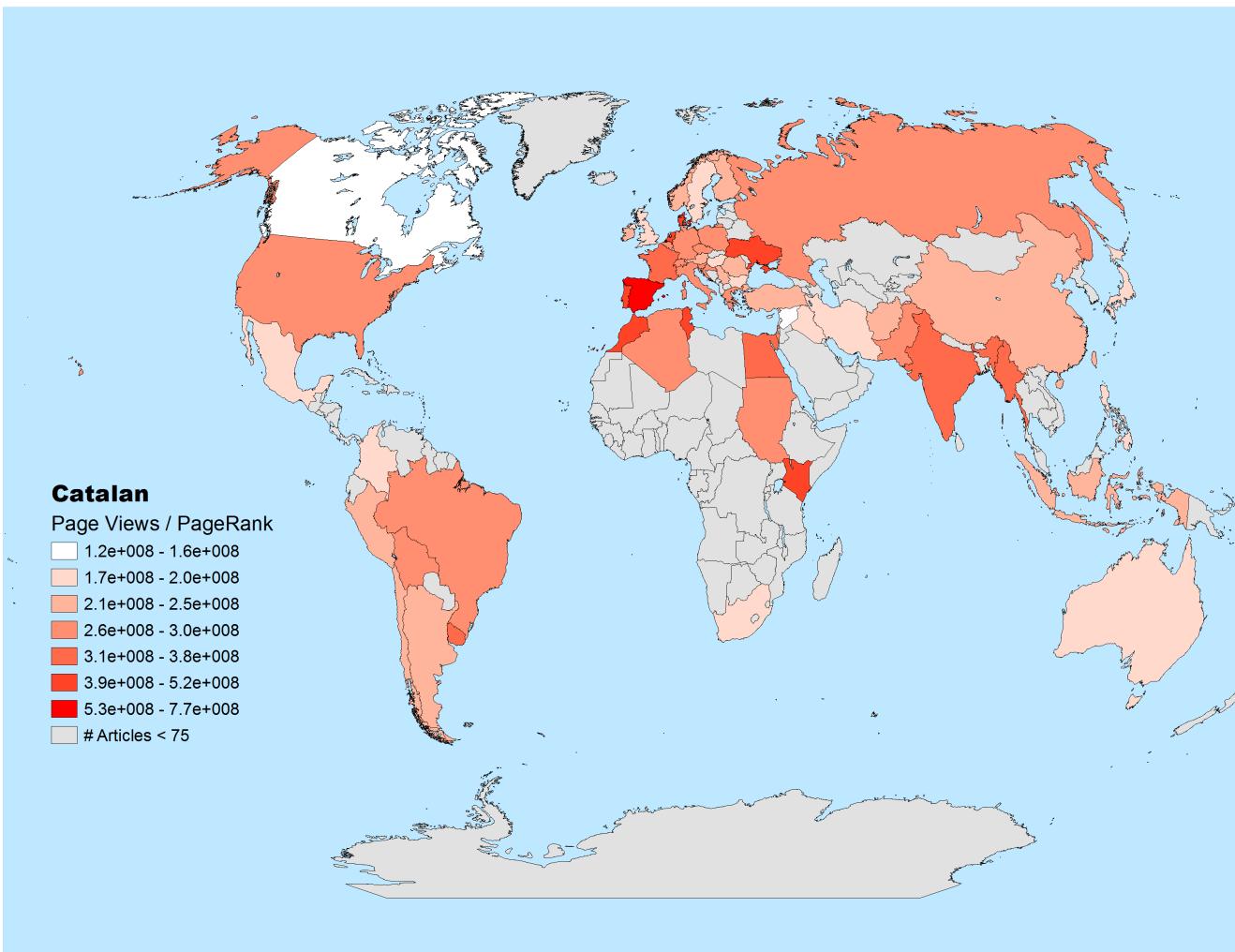


Figure 3.10-u: The ratio of page view sum to PageRank sum in the Catalan Wikipedia.

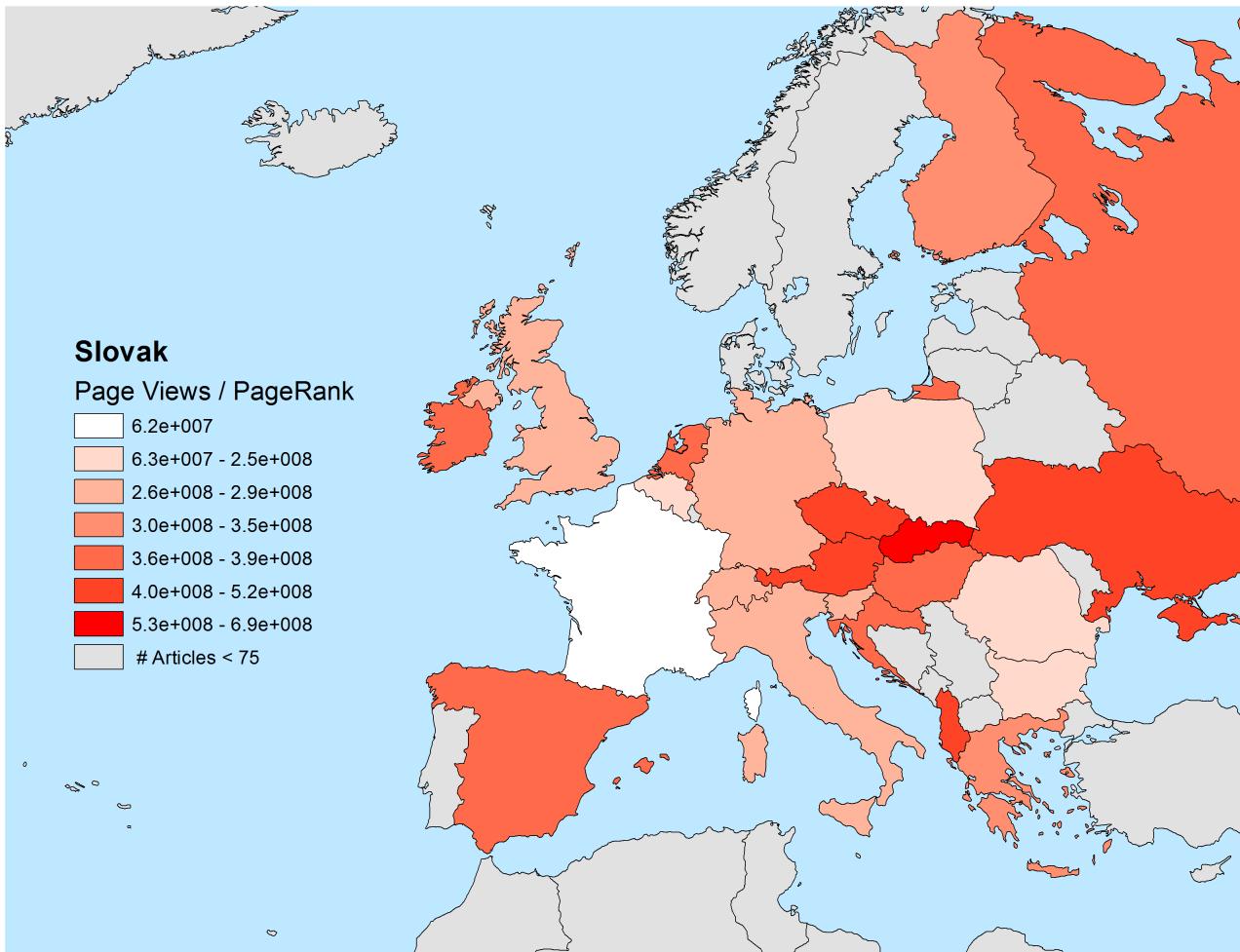


Figure 3.10-v: The ratio of page view sum to PageRank sum in the Slovak Wikipedia. Note that France, which contains a plethora of automatically created content in the Slovak Wikipedia receives very little reader interest relative to the importance placed upon it by the content of the language edition.

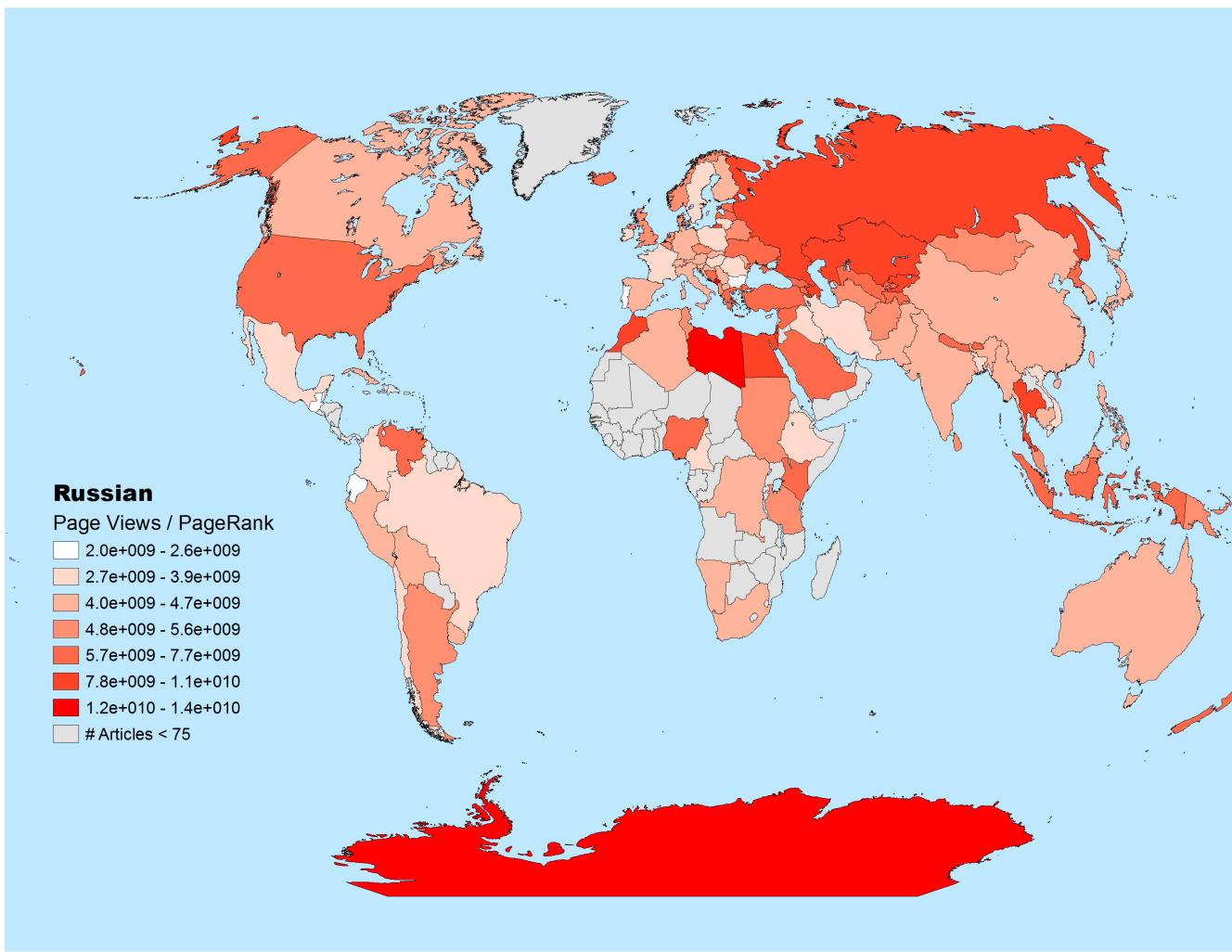


Figure 3.10-w: The ratio of page view sum to PageRank sum in the Russian Wikipedia.

### 3.11 Discussion

In this chapter, we have seen again and again from many different perspectives that there is far more support for the global diversity hypothesis than the global consensus and English-as-Superset hypotheses. We have also seen that a substantial portion of the diversity between the language editions of Wikipedia is due to each language edition contextualizing encyclopedic world knowledge for its own corresponding language-defined culture. Specifically, among other findings, we have demonstrated the following:

1. The set of concepts for which each language edition has articles is significantly different from language edition to language edition. Over 70% of concepts have articles in only a single language edition.
2. The English language edition covers no more than 76.5% of the concepts in any other of our 25 language edition dataset.
3. There are many examples of the concept-level diversity found in #1 and #2 being caused in part by the cultural contextualization.
4. The content of two articles about the same concept in different language editions only overlaps by at most 72% on average.
5. The longer of two same-concept articles is missing at least 11% of the content of the shorter article on average.
6. An English article about a concept that is covered in at least one other language edition is missing at least 29% (on average) of the information in multilingual Wikipedia about that concept.
7. Articles in the English Wikipedia cover a maximum of 93% (on average) of the content in another language edition's same-concept articles. The minimum amount was 75.5%, which is the average

percentage of content in Japanese Wikipedia articles covered by same-concept English articles.

8. There are many examples of the sub-concept-level diversity in #5 - #7 being caused in part by contributors to each language edition contextualizing the description of concepts for their own language-defined culture.
9. Concept-level diversity is greatest in the periphery than in the core of each language edition, with core and periphery defined by network centrality measures applied to the language editions' article graphs. However, the set of concepts that make up each language edition's core varies extensively from language edition to language edition.
10. Sub-concept-level diversity, on the other hand, is greater in the core of each language edition than in the periphery.
11. The amount of concept-level and sub-concept-level diversity varies significantly from topic to topic, and this variation is caused in part by cultural context. For instance, articles on topics most parochial to the English-speaking world like cricket and American football tend to be about single-language concepts. Even when another language edition covers these concepts, they cover them in far less depth than their English counterparts.
12. There is extensive diversity in the popularity of concepts across language editions, with language-defined culture being a major driver of this diversity.
13. The diversity in the importance of concepts as defined by Wikipedia readers' behavior is much greater than the diversity in the importance of concepts as determined by centrality measures.

14. Despite massive growth in multilingual Wikipedia since 2009, the amount of concept-level and sub-concept-level diversity has barely changed at all. This means that the language editions are not converging over time.
15. Using an experiment informed by theory from human geography, we were able to (1) prove in a quantitative fashion that each language edition reflects the cultural contexts of its contributors and (2) that this cultural contextualization of world knowledge is an important driver of the diversity between the language editions found in this chapter.
16. The cultural contextualization of world knowledge is so great in each language edition that we are able to exactly predict the geographic extent of certain language-speaking populations.
17. Using the same human geography theories, it was established in an even more robust and large-scale fashion that the cultural contextualization in the consumption of content in each language edition is much greater than that in the content itself. We also found evidence that automated content production processes severely exacerbate this tension between the content in multilingual Wikipedia and its consumption.

We have discussed in detail the meaning and implications of most of these findings in context in the above sections. However, we have not yet had an opportunity to discuss in detail this last contribution, that regarding the tendency for the content of each language edition to reflect significantly less cultural context than the patterns in its consumption. Doing so is the subject of the following subsection.

### 3.11.1 Culture, Content Consumption, and the Future of Wikipedia

Two Wikipedia-related developments are in the news as we finish this chapter. The first development, the impending launch of the Wikimedia Foundation-sponsored *Wikidata* project, is receiving by far the lion's share of the attention. Wikidata has even been called "Wikipedia 2.0" [21] and "Wikipedia's Next Big Thing" [154].

The Wikidata community defines its project as,

"...a free knowledge base about the world that can be read and edited by humans and machines alike. It will provide data in all the languages of the Wikimedia projects, and allow for the central access to data in a similar vein as Wikimedia Commons does for multimedia files."

The "data in all languages" component of this mission statement reflects one of Wikidata's most important goals: the instant transfer of new information from one language edition to *all* the other language editions of Wikipedia. The long-term end game for Wikidata is a Wikipedia with no language barriers at all.

The second development, one that made much smaller waves, is the release of a clip from the upcoming documentary, "*Web*" [105] featuring the co-founder of Wikipedia, Jimmy Wales. The clip highlights the Wikimedia Foundation's efforts to increase the amount of native-language information of interest to cultural groups that are underrepresented in Wikipedia. In the clip, an instructor is shown leading a class of students in creating a Spanish Wikipedia page about Palestina, Peru, their rural Peruvian village. The students, using computers from the One Laptop Per Child project, are surprised to find that there is no information about Palestina, and proceed to create the page. To readers of this thesis, the fact that there was no article about Palestina should come as less of a surprise; in Section 3.10 we showed that the Spanish Wikipedia is biased towards Spain and that there is extensive demand for additional content about Latin

America.

The two Wikipedia-related events represent divergent – but not mutually exclusive – visions of Wikipedia’s future. In the Wikidata vision, the best way to improve multilingual Wikipedia is to transfer information from one language edition to all other language editions. In many ways, this vision is not new. It is shared by Wikipedia translation projects (e.g. [35, 198, 208]), as well as by several research projects in the artificial intelligence and natural language processing communities (e.g. [2, 29, 188]). Wikidata, however, is in by far the best position to realize this transfer of information.

On the other hand, the vision of efforts like those to create a page about Palestina is that the best way to improve multilingual Wikipedia is to create *new information* that corresponds to the interests of each individual language-defined community. Our results suggest that the goals of the Palestina vision are equally important – if not more important – than those of Wikidata. While we believe that providing access to information in all language editions is useful, greater attention needs to be paid and more resources dedicated to replicating the Palestina example many times over and in many more language editions.

Consider for example our findings related to the Slovak Wikipedia. In Section 3.10, we saw that an automated process had created an enormous amount of content about France. However, it appears that few readers of the Slovak Wikipedia pay any attention to this bot-created content. Despite the fact that the number of articles about places in Slovakia is only 40% of the number France-related articles, the Slovakia articles received 4.5 times as many page views from 2010 to 2012. If we assume that the consumption of content represents demand for more content about the same topic, there is a much greater need in the Slovak Wikipedia for information about Slovakia than there is for information about places like France.

The Slovak Wikipedia example is instructive because the end result of Wikidata will closely resemble the actions of the bot that added so much information about France. If Wikidata is successful, Slovak readers may be able to access a great deal more information about topics well-covered in other language editions, but unlike this transferred information, their interests will continue to reflect the shared expertise of their own language-defined culture. That is, they will continue to demand more information about Slovakia and will pay relatively less attention to the information added from other language editions. As such, according to the behavior of its readers, the Slovak Wikipedia is more in need of Palestina-like content creation than Wikidata-like information transfer.

Section 3.10 also provided another key piece of evidence in support of Palestina-like content creation. In this section, we showed that for *every language edition*, the relative popularity of articles about home cultural regions was greater than the relative amount of content about those regions. We identified a similar phenomenon in Section 3.8, where we saw that the sets of most-visited articles in the language editions displayed significantly more diversity than the sets of the most-central or most-discussed articles. If we again assume that popularity implies demand for new information, every language edition could benefit from directed efforts to create new content about topics in the shared expertise of the corresponding language-defined cultures.

While our results advocate for significantly more attention to be paid to Palestina-like content creation, it is important to reiterate that this is not mutually exclusive with the goals of Wikidata. Indeed, as noted above, we fully agree with the need to provide access to all information in Wikipedia. However, our results indicate that Wikidata should be thought of as a long-tail project, whereas Palestina-like content creation forms the “short head.” With Wikidata currently the main target of the Wikipedia community’s energy and resources, a rebalancing

towards the creation of new, culturally-relevant content is likely in order.

Before closing, it is important to discuss one additional issue related to Wikidata and the cultural contextualization of Wikipedia. The mechanism by which Wikidata plans to transfer information across language edition boundaries is a language-neutral central repository of data. The creation of such a repository could have represented *enormous* risks for the ability of each language edition to contextualize world knowledge for its corresponding language-defined culture. In particular, it had the potential to wipe out at least a portion of the cultural context embedded in each language edition at the sub-concept level. However, a decision was made early in the Wikidata planning stages to allow for every “fact” (property-value pair) in the central repository to have multiple versions, with each language edition able to decide on its own which version to use. For instance, the Hebrew Wikipedia could display one number for the area of Israel, while the Arabic Wikipedia could display a different one [63]. In our conversations with the Wikidata team soon after the project launched, this was a design decision for which we advocated.

That said, important Wikidata-related risks for the cultural context in multilingual Wikipedia do still exist, even with the support of multiple versions of each fact. First, Wikidata does not allow multiple versions of interlanguage links. That is, as of this writing, the cultural nuance inherent to the conflicts in the interlanguage link network (Section 3.3) are completely ignored. Each language edition is currently only permitted to have a single article about each concept, and this is a problem that needs fixing. Doing so, however, is likely to be difficult. As of now, all of Wikidata is based on a one article-to-one concept assumption.

The second risk Wikidata poses to culturally contextualized encyclopedic information is that even though multiple versions of each fact may supported, this does not mean that diverse

cultural views will actually get represented. Readers of language editions with small editor communities like Slovak may simply have to deal with getting the point of view of editors from large language editions like English, German, and French.

That said, by enabling easier access to information that *is* in all relevant cultural contexts, it is possible that Wikidata could open the door to creating encyclopedias contextualized for cultures other than those defined by language. For instance, country music fans could create a much higher quality country music encyclopedia than is currently available by leveraging country music-related information from Wikidata while at the same time writing unique articles about country music topics that are not considered sufficiently notable for inclusion in an encyclopedia for all English speakers. If this flowering of new high-quality encyclopedias were to occur, it would greatly expand the utility of the methods and findings in this chapter. That is, everything we have done here with language-defined communities could be investigated in cultures defined by any number of other characteristics.

### 3.12 WikAPIdia

This chapter's last in-detail discussion is dedicated to this chapter's final major contribution: WikAPIdia, the Wikipedia software library we wrote and used to execute every one of the above studies. Despite the importance of WikAPIdia to our research, it has never before been described with any depth. Even without a formal write-up, previous versions of WikAPIdia have been used in at least one research project outside of our group [170]. We believe that the improvements in the latest version (v0.3) combined with the detailed description of its capabilities and functions below could potentially increase its adoption in the large community of researchers and practitioners that study or apply Wikipedia in their work. Upon publication of

this thesis, WikAPIdia 0.3 will be made available on WikAPIdia’s website<sup>56</sup> under an LGPL license, allowing for its use in both research and commercial settings.

The work in this chapter was executed almost exclusively using version 0.3 of WikAPIdia, a major upgrade from version 0.2, on which we based all of our previous Wikipedia work. Version 0.3, a near-total rewrite, contains many important technical improvements that make the large-data analyses in this chapter much more tractable. For example, the version of the first concept-level diversity study (Section 3.4) that appeared in our CHI 2010 paper [82] originally took approximately one week to complete. Using version 0.3 of WikAPIdia, as we did below, it took approximately minutes. Along the same lines, WikAPIdia 0.2 was only able to support proof-of-concept versions of Omnipedia and Atlasify due to inefficient use of memory and non-optimal implementations of important algorithms, a design choice that was appropriate given the early stage of those research projects. Version 0.3 enables research and implementations of Omnipedia and Atlasify to progress past what was possible with version 0.2.

WikAPIdia differs from the two most well-known Wikipedia APIs – Java Wikipedia Library<sup>57</sup> (JWPL) [220] and Wikipedia Miner<sup>58</sup> [135] in a number of respects. The five most significant of these differences are as follows:

1. WikAPIdia is *multilingual* by nature. It can be used to *simultaneously* access the data in any number of languages. This functionality enabled essentially every study in this chapter.
2. WikAPIdia is *spatially-referenced* by nature. WikAPIdia has extensive functionality for connecting the concepts in Wikipedia to any number of spatial reference systems. We leveraged the

---

56 [http://collablab.northwestern.edu/wikapedia\\_api/Wikapedia/Home.html](http://collablab.northwestern.edu/wikapedia_api/Wikapedia/Home.html)

57 <https://code.google.com/p/jwpl/>

58 <http://www.nzdl.org/wikification/index.html>

geographic subset of these reference systems in Section 3.10, but this feature is most fully utilized in our Atlasify project (Chapter 8).

3. WikAPIdia includes implementations of several *semantic relatedness* algorithms based on Wikipedia and integrates them closely into the API. These algorithms are used extensively in Chapter 6 and 8 and aided in our implementation of *Conceptualign* (Section 3.3).
4. WikAPIdia provides integrated API access to Wikipedia resources for which there are limited or no existing APIs. These resources include sub-articles, missing links, topics, page views, page rank calculations, and others.
5. WikAPIdia has several low-level features that make multilingual Wikipedia-based projects easier to implement.

### #1: Multilinguality

Both JWPL and Wikipedia Miner support non-English language editions of Wikipedia, but they do so in a monolingual fashion. That is, due to their lack of an algorithm that groups articles about the same concept in different language editions such as *Conceptualign*, they necessarily silo the data from each language edition away from that of the others. It is therefore impossible to simultaneously access information from multiple language editions about the same concept. WikAPIdia, on the other hand, natively integrates information from multiple language editions, while at the same time preserving language-by-language access if it is needed by developers. WikAPIdia currently supports 25 languages, but adding a new language is straightforward and does not require the reparsing of already-parsed editions.

This ability to access multilingual Wikipedia in a structured fashion at a concept-level opens

up a whole new category of Wikipedia-based research and applications. All the research and applications that use Wikipedia data in this thesis leverage this capability, and it is likely that there are many additional possibilities for employing this information.

### *#2: Spatial Referencing*

In the same way that multilingual information access is a first-class citizen in WikAPIdia, so is the spatial referencing of Wikipedia concepts. WikAPIdia uses a process we call *spatiotagging* (Chapter 8) to connect Wikipedia concepts (and their corresponding articles) to spatial representations such as latitude/longitude coordinates.

Additionally, through the integration of open-source libraries and spatial databases and their customized application to Wikipedia’s concepts, WikAPIdia effectively allows for any Geographic Information System (GIS) algorithm to be executed using Wikipedia information or a derivative of this information that is the output of a Wikipedia-based algorithm. This “built-in GIS” is used in all spatial and geospatial research and applications in this thesis, including the work in Section 3.10, and it is absolutely essential to Atlasify (Chapter 8).

### *#3: Semantic relatedness algorithms*

WikAPIdia has tightly integrated and successfully validated implementations of a number of Wikipedia-based semantic relatedness (SR) measures, including some of the most well-known in the literature and those we have developed ourselves in previous work. SR measures are fully described in Chapter 6. The close integration of these measures with the rest of the API provides the opportunity to apply these algorithms directly to multilingual and spatially-referenced Wikipedia data in a straightforward manner. This functionality is integral to the study of the effect of UGC diversity on UGC-based algorithms (Chapter 6), it helps Omnipedia users find related concepts (Chapter 7), and it is essential to the entire Atlasify project (Chapter 8). Moreover, because SR measures are an important component of many of the most well-known applications of Wikipedia data in artificial intelligence and natural language processing,

WikAPIdia's support for SR measures should enable researchers and practitioners to more easily innovate this domain. We did exactly this with our ensemble semantic relatedness measure, AtlasifySR+E (Chapter 8).

#### *#4: New data resources*

WikAPIdia provides integrated API access many Wikipedia resources that are not available in other Wikipedia APIs. These resources include:

- sub-articles
- missing links identified via wikification
- page views
- spatial references
- parseable/unparseable links
- PageRank scores
- YAGO2s topics

Adding support for these resources required brute force manual information gathering (e.g. potential sub-article relations), training and testing of learned models (e.g. sub-articles), aggregating diverse datasets from various providers (e.g. page views, spatial references, topics), and the efficient implementation of computationally expensive algorithms (e.g PageRank scores).

#### *#5: Lower-level contributions*

WikAPIdia also has a large number of lower-level advantages over other APIs, many of which were added in version 0.3. First and foremost, WikAPIdia 0.3 completely abstracts its API from the underlying data source. This means that all the algorithms based on Wikipedia data can use data from any store of information. This has two main advantages: (1) all aspects of the API, without any modification, can use data that is in-memory, in a database that is on disk, or in a key-value store in the cloud, and (2) all API methods can access data from the live Wikipedia (not the parsed version) through calls to Wikipedia's native API.

Both of these properties have significant benefits. With respect to the former, researchers and developers can customize the data structures that are loaded into memory through simple changes to an XML configuration file or using programmatic methods. This allows limited memory resources to be allocated in the most efficient fashion for a given algorithm or application. For instance, when running a Wikipedia Article Graph-based semantic relatedness measure, the graph (a very large data structure) can be loaded into memory. When running a SR measure that relies on the text of Wikipedia articles and uses the article graph only incrementally, the article graph can remain on disk and the index that maps articles to their corresponding documents in WikAPIdia's text index can be loaded into memory. We expect that other researchers and practitioners will find WikAPIdia much more useful as a result of the abstraction of the data source from the API.

Fully abstracted integration of data from the live version of Wikipedia is perhaps an even more significant feature than disk/memory/cloud abstraction. While accessing data directly from Wikipedia is orders of magnitude slower than using local, parsed versions of database dumps, doing so also has important advantages. First, using the live data allows access to the most current version of each article, category, and so on. While database dumps are provided on a regular basis, parsing every database dump update can be cumbersome. In addition, even the week or two between dumps limits their usefulness in research and applications that require or benefit from up-to-the-minute information. Obtaining data directly from Wikipedia also provides access to terabytes worth of article histories without the extensive time and resources needed to parse this information locally. In this thesis, we leverage this connection to the live API to validate the locally parsed and extracted information. We are also working on many other applications of this feature. For instance, we are hoping to integrate live data into Omnipedia.

The second and final lower-level contribution we will discuss is WikAPIdia's close integration with various well-known Java libraries. For example, for both the WAG and the WCG of all language editions, WikAPIdia implements the graph interfaces of the Java Universal Network/Graph Framework (JUNG) software library<sup>59</sup>. For the ILL graph, the JGraphT interface is implemented. These interfaces allow software that uses WikAPIdia to run any algorithm in either of these packages on the WCG and WAG (e.g. PageRank, betweenness centrality). Some of the experiments in this chapter utilize this functionality. For instance, *Conceptualign* is implemented on top of JGraphT's breadth-first search connected component identifier. In addition, both implementations are at a low level and utilize WikAPIdia's data source abstractions. This means that the WAGs, WCGs, ILL graph all can operate in memory, from disk, in the cloud, or directly from Wikipedia's API.

With its ability to facilitate access to entirely new types of Wikipedia information in any language, its inclusion of both novel and state-of-the-art Wikipedia-based algorithms, and its lower-level improvements to the access of data from Wikipedia, WikAPIdia is both the result of and the methodological foundation of our Wikipedia work. It is our hope that it can be as useful to other researchers and practitioners as it has been for us.

---

59 <http://jung.sourceforge.net/>

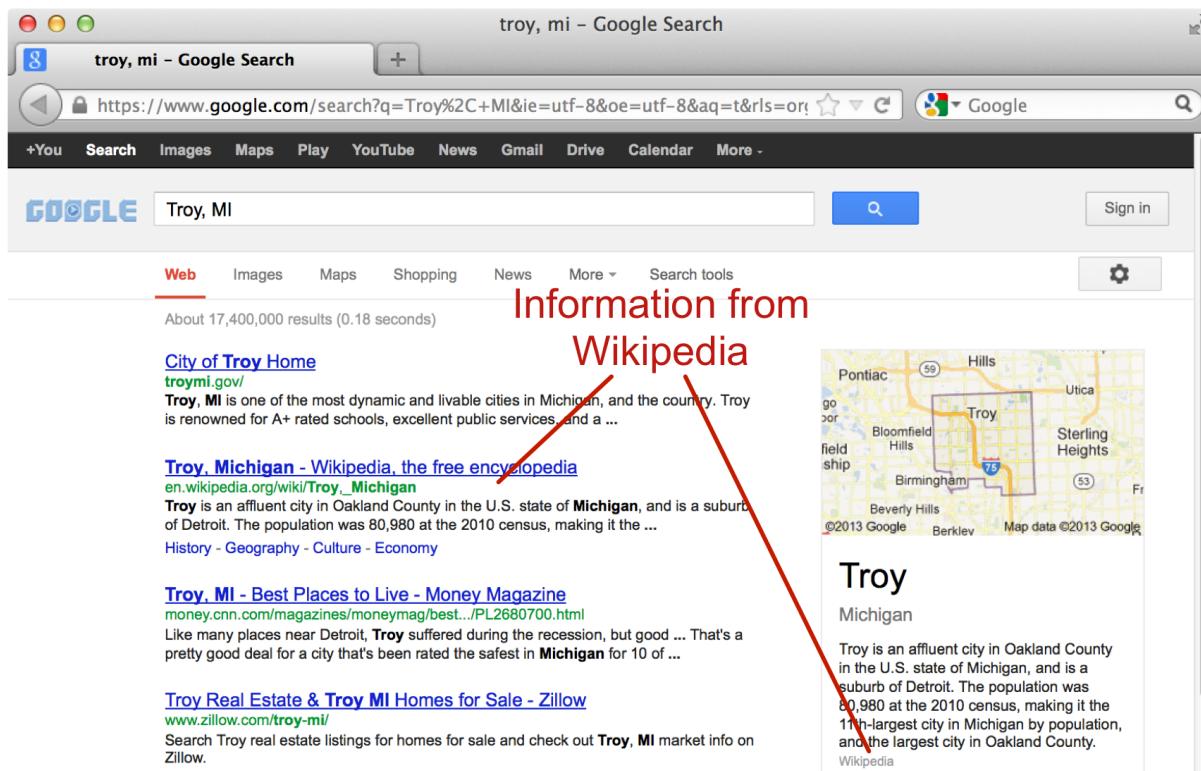
## 4 Geographic Localness Diversity in User-Generated Content

*Note: This work originally appeared in the Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW 2010) [81]. While much of the text here is original to this thesis, portions have been adapted from the original publication.*

In the previous chapter, we focused on the cultural contextualization of user-generated content across language-defined communities. In this chapter, we turn our attention to geographically-defined communities. Like was the case in Chapter 3, there is some explicit and implicit debate in the literature as to the degree to which well-known repositories of geographically-referenced UGC (“GeoUGC”) are culturally contextualized. A particularly important dimension of this debate revolves around whether GeoUGC is predominantly made up of information contributed by locals with their local knowledge or is largely a source of non-local information, for instance contributed by non-residents and/or tourists. This is a similar question to that which we addressed above in Chapter 3. If local information predominates, this is tantamount to the global diversity hypothesis being supported for GeoUGC: information about most regions will be culturally contextualized for those regions, just as information about Wikipedia concepts tends to be contextualized for each language-defined community. For instance, in the case of a photographic UGC repository like Flickr, we might see many photos of local parks, businesses, and other geographic features that reflect the local character of a given area. If, on the other hand, information contributed by non-residents and/or tourists predominates, we can expect more globally-consistent repositories. This would mean that Flickr,

for example, would be biased toward pictures of monuments, museums, and other geographic features deemed important by outsiders.

As mobile and location-based technologies have rapidly gained more and more importance, so too has the issue of the localness of user-generated content. In the bygone era of Web 1.0, a search for “Troy, Michigan” would have returned nearly guaranteed local knowledge such as Troy’s city homepage or a local newspaper. These days, however, GeoUGC from UGC sites such as Wikipedia has become a predominant source of information about scores of geographic concepts (e.g. cities, towns, national parks, landmarks, etc.). For instance, Figure 4-a shows Google’s unpersonalized search results for the query “Troy, Michigan.” The Wikipedia page for Troy is the second-ranked result, a phenomenon that is very typical for queries about geographic



*Figure 4-a: The unpersonalized search results for the query “Troy, MI” reveals how important geographic user-generated content like Wikipedia has become to the average user’s web experience.*

concepts. Moreover, as Google develops and utilizes the Google Knowledge Graph at an increasing rate, geographically-referenced UGC becomes even more significant; as can be seen in Figure 4-a, the Knowledge Graph contains a great deal of information that has been mined from Wikipedia.

Within the geography community, it has generally been assumed that geographically-referenced user-generated content – known as volunteered geographic information (VGI) in geography – does indeed represent local information. For instance, Michael Goodchild, a well-known geographer and coiner of the term “volunteered geographic information,” has written:

“...The most important value of [user-generated geographic information] may lie in what it can tell us about *local* activities...that go unnoticed by the world’s media, about life at the *local* level. It is in that area that [user-generated geographic information] may offer the most interesting, lasting and compelling value” [57] (emphases added).

On the other hand, certain areas of the computer science community have assumed implicitly that the GeoUGC in some repositories is predominantly *non-local*, with local information either having a minor presence or a small information value. Namely, a number of projects in the computer vision space adopt the perspective either algorithmically or in the framing of projects that Flickr and other photographic geographically-referenced user-generated content is predominantly made up of many photos of the same relatively small set of landmarks (e.g. [27, 75]). Moreover, some of these projects assume that these landmarks are representative of all photos in a certain area, an assumption that explicitly ignores any local knowledge in the repository.

In this chapter, we investigate the extent to which the information in five different large-scale UGC repositories is local. In addition, we introduce the idea of spatial content production models (SCPMs) to describe how the particular uses and features of UGC repositories might

influence the degree of “localness.” This allows us to characterize, for example, the differences between the “you have to be there” model of a UGC repository like Flickr with the more easily traversable “flat Earth” model of something like Wikipedia. Finally, theoretical and applied implications are summarized, and future work is discussed, including the possibility of adapting SCPMs to non-spatial contexts for the purposes of predicting cultural contextualization in non-geographic UGC.

## 4.1 Background and Related Work

GeoUGC has been shown to be incredibly useful as a resource for an enormous variety of technologies. In the human-computer interaction space, it has been leveraged in areas ranging from natural user interfaces (e.g. [179, 180]) to collaborative technologies (e.g. [151, 161]) to social network analysis (e.g. [118, 176]). As noted above, its applications outside of HCI have been even greater in number and diversity, with geographic user-generated content being used as input to Google’s Knowledge Graph, as training data for a large variety of Twitter-based algorithms (e.g. [22, 213]) and as a key component in many other research projects and technologies.

While its application has been prolific, there has been much less focus on the nature and properties of GeoUGC. The existing work in this area has largely been limited to Wikipedia-based GeoUGC. For instance, Hardy [70] showed that editors of Wikipedia follow a power law in their number of contributions of geographic UGC. Similarly, Lieberman and Lin [119] demonstrated that the convex hull of edited geographic articles in Wikipedia (specifically, the locations of the geographic entities they describe) is likely somewhat small for a large minority of registered English Wikipedia users. However, the degree of localness to the actual user is not

considered.

## 4.2 Data Preprocessing

In order to investigate the degree to which participation in UGC repositories is local, we draw upon data from five different UGC repositories: Flickr and four language editions of Wikipedia (English, Catalan, Norwegian, and Swedish). The following describes the processing done in order to prepare the data for analysis.

### 4.2.1 Flickr

As is evidenced by Flickr's own map interface to its photos<sup>60</sup>, a large portion of Flickr's dataset has been geotagged by its users, either automatically through a GPS-enabled camera (such as the iPhone) or manually. We used Yahoo!'s API access to Flickr to download approximately a year's worth of geotagged photo metadata beginning in May 2008, resulting in information about 10+ million photos.

However, for the purposes of the studies described below, we also needed data about the location of the Flickr users who took these photos. We again accessed the Flickr API to download photographer information using the photographer ID tags included in each of the 10 million photos' metadata. We were particularly interested in the photographer's self-specified location, an optional field in Flickr user profiles. While a small percentage of users did provide this information, it was text-based and often quite colloquial in nature (i.e. "Grand Rapids, U S & A", "Minneapolis-St. Paul, Twin Cities")<sup>61</sup>. This created a problem, as there is no formal gazetteer, to our knowledge, that is capable of handling this type of vernacular spatial data.

---

60 <http://www.flickr.com/map/>

61 This is a phenomenon we address in detail in our work on location fields in user profiles, which falls outside the scope of this thesis [83]

Fortunately, Wikipedia has a rich set of this vernacular data in the form of the redirects resource discussed in Section 3.2. Recall that redirects form a massive mapping table designed to “redirect” users who search for, say, “San Fran” or “San Francisco, USA” in the Wikipedia search bar, to the “San Francisco” article. As such, we were able to leverage these redirects to connect Flickr users’ self-specified locations to Wikipedia pages with geotags, which then gave us a latitude and longitude for the Flickr user. To supplement this process, we also performed a Wikipedia-only Yahoo! Search API query on each colloquial location, and if the first result was identified as a geotagged Wikipedia article, we applied the geotag to the user’s location. In the end, we were able to successfully geocode 14,295 photographers who took 185,871 geotagged photos.

#### **4.2.2 Wikipedia**

Identifying the spatial footprints of Wikipedia articles about geographic concepts simply involved using our existing database of these articles, the creation of which is outlined in Section 3.10. Identifying the location of Wikipedia contributors was a significantly more challenging process. Wikipedia contributors can be broadly split into two classes, anonymous and registered users. While we can mine the IP addresses of anonymous contributors and use these in IP geolocation, it is extremely difficult or even impossible to discover the position of large numbers of registered Wikipedia users, whose IP addresses are not recorded. As such, we omit them from our studies, admittedly a drawback given that they produce the lion’s share of the content that is read by Wikipedia consumers. Anonymous users are responsible for only about 26% of content read by visitors to the English Wikipedia [160]. However, given the largely unaddressed nature of questions surrounding the localness of user-generated content, anonymous users represent an

unprecedented opportunity to acquire large-scale data about UGC contributors' spatial contribution behaviors.

#### **4.2.3 The Problem of Scale**

Like the study in Section 3.10, our work here is affected by the Geoweb Scale Problem (GSP) [84]. Just as before, we avoid much of this problem by using simple name matching to connect first-order administrative districts and countries to their polygonal representations. In this study, however, instead of incorporating the polygonal representations into our analyses, we omit any data points where this is an issue (e.g. the Wikipedia article “United States” and Flickr users who specified their home location as “England, United Kingdom”). This was necessary due to our study’s use of simple distance as its main metric; measuring the distance between two polygonal representations in this context is not well-defined.

This study also faced another, more serious problem due to the GSP that we have not yet seen in this thesis. In our previous dealings with the GSP, we were operating at very non-local scales (e.g. global, continental). Here, we are dealing with much larger<sup>62</sup> scales, which means that a whole series of new geographic features – e.g. second-order administrative districts like counties, cities with a large areal extent – need to be attached to polygonal representations. Consider a situation in which a Flickr photographer from the neighborhood of Rogers Park, Chicago has specified her home location as “Chicago,” which we then connected to the Wikipedia-based geotag for Chicago. This latitude and longitude point in this geotag happens to fall in the city’s downtown, a full 20km from Rogers Park. If this photographer took a picture at her house, our system would register this photo as 20km from her home location. This is

---

<sup>62</sup> In this thesis we adopt the formal definitions of “small scale” and “large scale”, which refer to the size of the representative fraction (e.g. 1-inch-to-1-mile). A larger scale indicates a larger fraction (or fewer real-world units to every map unit) and a small scale indicates a smaller fraction (more real-world units to every map unit).

obviously incorrect, and if not accounted for, could have important deleterious effects on a study involving the concept of localness. Unfortunately, no spatial dataset that is publicly available allows us to take the same approach as we did with countries and first-order administrative districts in Section 3.10 with the same degree of accuracy. Compounding the issue when IP addresses are considered is the accuracy of IP geolocation. The IP geolocation software used in our study performs at 68 – 79% accuracy within 25 miles in most of the countries considered.

Our workaround to these two issues involves limiting the precision at which we report our results. The threshold we chose was 50 kilometers, the radius of a large city. This accounts for a significant portion of the GSP-related error and that introduced by IP geolocation. The main effect of this decision is that we do not report localness at any distance smaller than 50km. In other words, our hypothetical Chicago photographer would be correctly identified as 50km or less from the image she took at her home rather than incorrectly identified as 20km away.

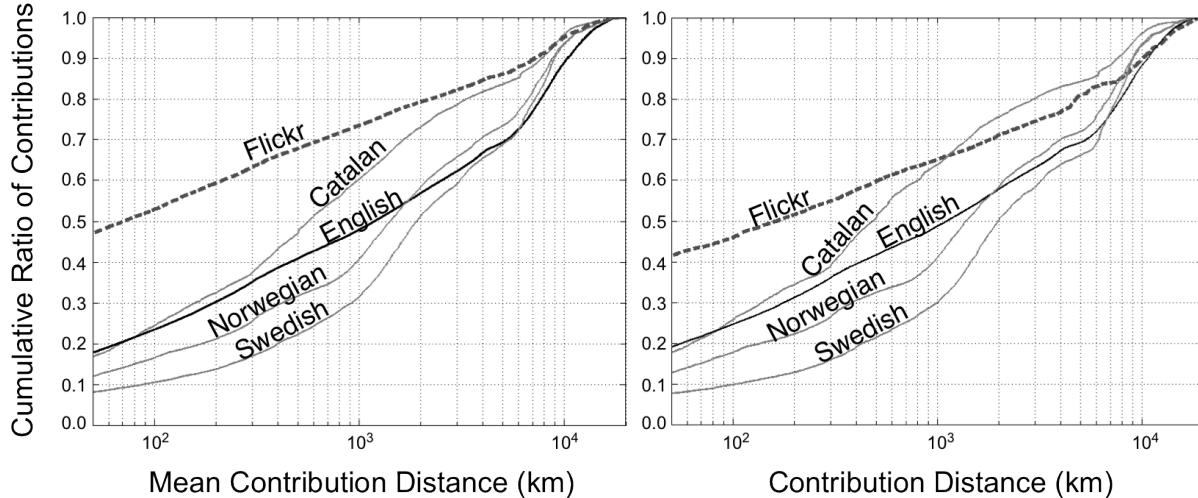
### 4.3 Study of Contributor Spatial Behavior

For our five repositories, we calculated for each contributor the *mean contribution distance* (MCD). A contributor's MCD is defined as:

$$MCD = \sum_{i=1}^n \frac{d(C, c_i)}{n}$$

where  $C$  is the specified location of the contributor, and the location of each of  $C$ 's  $n$  contributions is denoted by  $c_i$ . This metric has a large benefit over that used by Lieberman and Lin [119] in that each location is effectively weighted by the number of times a contribution is made, an important fact considering Lieberman and Lin's discovery that many Wikipedia users edit their "pet geopages" very frequently. Our distance function  $d$  is that of the great circle

## Localness of User-Generated Content Repositories



*Figure 4.3-a: LEFT: The empirical cumulative distribution of MCDs for each dataset examined, or  $\text{cdf}(\text{MCD})$ . RIGHT: Empirical cdfs of contributor-to-contribution distances for all contributions. In other words,  $\text{cdf}(d(C,c))$  for all  $(C,c)$  pairs in each repository. Note that in both cases, the x-axis is on a log scale.*

distance<sup>63</sup>.

The left side of Figure 4.3-a shows the empirical cumulative distribution function of contributors' MCDs. While approximately 47% of Flickr users contribute, on average, content that is 50km or less from their specified home location, this number drops quite a bit for Wikipedia users. The equivalent number for the English Wikipedia, for example, is around 17%. The Catalan Wikipedia displays a similar rate of localness to English with about 16% of contributors having an MCD of 50km or less. The Norwegian and Swedish Wikipedias have a lower percentage of users who contribute very locally; in both language editions, less than 15% of anonymous editors have an MCD of less than 50km.

Why does this difference between Wikipedia and Flickr exist? We hypothesize that the answer to this question lies in the spatial content production models (SCPMs) of each repository.

---

63 We use a spherical Earth assumption to speed the calculation of great circle distance. For the purposes of this chapter, the errors introduced by doing so are minimal.

In Wikipedia, “the encyclopedia anyone can edit,” contributors simply must possess the desire to add/edit/delete content. In the spatial domain, this means that there exists “total time-space compression” [73], and as such, we can categorize Wikipedia’s SCPM as a “flat Earth” model. In other words, it is just as easy for someone in Ann Arbor, MI to edit the “Ann Arbor, Michigan” page as it is for that person to edit the “Columbus, Ohio” page. This fact is reflected in the much smaller percentage of contributors who edit locally on average. Flickr’s “you have to be there” SCPM, on the other hand, more or less requires that contributors have visited the location about which they are contributing<sup>64</sup>. This creates a MCD pattern that begins to resemble offline spatial behavior models, and therefore creates a repository in which local participation is much greater.

While Wikipedia has less local participation compared to Flickr, it is important to note that distance still matters a great deal on Wikipedia’s “flat Earth.” Despite the fact that editors can edit a page about anywhere in the world, they still tend to edit pages about places closer to them at a disproportionate rate. The data on the left side of Figure 4.3-a elaborates on the findings of Lieberman and Lin [119], who found that the convex hull of edited spatial articles tends to be somewhat small for a large minority of English users.

We now turn our attention to the “localness” of participation across entire repositories, rather than individual contributors’ behaviors. In other words, we indirectly incorporate the power law found by Hardy [70] into our analyses. The right side of Figure 4.3-a is similar to the left side except that instead of showing the distribution of contributors’ MCDs, it shows the distribution of distances from *all* contributions to their respective contributors.

The right side of Figure 4.3-a demonstrates that the spatial contribution behavior is

---

<sup>64</sup> It is possible for users to (intentionally or unintentionally) mis-georeference images or manipulate their images’ EXIF data, but it is our assumption that this is a fairly rare phenomenon.

relatively independent of contribution rates. If it were not, we would see a significant difference between the left and right sides of Figure 4.3-a. For instance, around 42% of Flickr photos are taken within 100km from their photographer (righthand side), while we saw that around 47% of photographers take photos within 100km on average (lefthand side). There is one small exception, however: the Catalan Wikipedia line crosses the Flickr line at around 1000km on the right side of Figure 4.3-a, something it does not do for MCDs. Analyzing this phenomenon in detail is a subject for future work.

## 4.4 Discussion

The results shown above, combined with those from related work, have a number of applied and theoretical implications. First and foremost, the findings in Figure 4.3-a show that repositories of user-generated content are neither exclusively local or non-local, but rather contain substantial amounts of both types of information. Perhaps more importantly, our results suggest that the relative amount of non-local vs. local information can vary extensively across GeoUGC repositories, possibly due to the SCPM of each repository. If Goodchild is correct in his statement that the main benefit of GeoUGC is local knowledge, designers of GeoUGC communities will want to learn from the differences between Wikipedia and Flickr. One suggestion would be to adopt SCPMs that “decompress time-space” in content production, as is naturally done in the process of taking photographs. For instance, consider “geowikis” such as OpenStreetMap<sup>65</sup> and Cyclopath [161]. If these communities wish to ensure more local knowledge, they could require that users upload information from GPS units rather than allowing them to encode their knowledge using a web interface.

---

65 <http://www.openstreetmap.org/>

This brings us to an interesting two-part question: what is the equivalent of a SCPM for language-defined cultures like those considered in Chapter 3, and could these language-defined content production models affect the extent to which the language editions of Wikipedia contain culturally contextualized information versus information, for instance, translated from other language editions? One hypothesis here is that the availability of translation tools and related technologies form the (non-spatial) *content production model* (CPM) in this context. Currently, multilingual Wikipedia has something of a “you have to be there” CPM in that transferring information from one language edition to another involves a good amount of effort (even using translation tools such as WikiBhasha [208]). Analogously, using modern transportation technology, traveling long distances also requires effort. However, let us consider a situation in which technologies like Wikidata (Section 3.11) greatly reduce the cost of transferring information across language edition barriers. The effect would likely be the same as transitioning from a “you have to be there” SCPM to a “flat Earth” SCPM: the amount of culturally contextualized information, or “local information,” would be reduced. Investigating the potential of SCPMs as a more universal framework for understanding the cultural contextualization of UGC, not just GeoUGC, is a direction of our future work.

Another important theoretical direction that must be investigated involves the importance of UGC repositories as sources of *place* information [34]. The degree to which these repositories are defined by locals versus outsiders is an important question in this respect. While we have addressed this issue in the Flickr context, the dynamics of Wikipedia participation make this more difficult. Primarily, a deeper inspection of the content on Wikipedia pages is warranted. We were only able to measure participation, and it is well-known that participation and content have a complicated relationship in the Wikipedia context [160].

Of course, one additional important question in this area necessitates looking more deeply at contributors rather than simply classifying them as local or non-local (e.g. socioeconomic status). Doing so is a major direction of our current research, our approach to which is briefly summarized in Chapter 9.

## 5 Inferring Geographic Cultural Community Memberships from Tweets

*Note: This work originally appeared in the Proceedings of the 29th ACM Conference on Human Factors in Computing Systems (CHI 2011) [83]. While much of the text here is original to this thesis, portions have been adapted from the original publication.*

In the previous chapters, we have considered user-generated content to be the “dependent variable” of sorts and the cultural memberships of its contributors to be the “independent variable.” In other words, we have taken known cultural groups defined by geography and language and have examined whether or not we see a difference in the user-generated content contributed by members of these cultural groups. In this chapter, we flip this relationship around. We show that user-generated content is sufficiently culturally contextualized that we can predict the cultural memberships of its contributors simply by examining the content they contribute. Specifically, we show that we can predict using classification techniques the geographic cultural community (i.e. the home country and state) of a Twitter user solely by examining their tweets, and do so with decent accuracy.

Below, we first discuss how we collected Twitter data and processed it into a ground truth dataset for our geographic cultural membership classifier. Next, we cover the particulars of the classifier itself, highlighting our approach to feature selection. Third, we report our results, which show that we are able to predict the home country and state of a Twitter user at a rate

significantly better than random. Finally, we close with a discussion of the implications of this work.

## 5.1 Data Collection and Preprocessing

From April 18 to May 28, 2010, we collected over 62 million tweets from the Twitter Spritzer sample feed using the Twitter streaming API. The Spritzer sample represents a random selection of all public messages. Based on a report that Twitter produced 65 million tweets a day in June 2010 [178], we estimate that our dataset represents about 3-4% of public messages generated during the data collection period. From these 62 million tweets, we further identified the tweets that were in English using a two-step combination of LingPipe's text classifier<sup>66</sup> and Google's Language Detection API<sup>67</sup>. All together, we identified 31,952,964 English tweets from our 62 million tweets, representing 51% of our dataset.

Using these approximately 32 million tweets, we developed a ground truth dataset that would allow us to construct a learned model of the relationship between tweets and geographic cultural communities. Examining our tweets, we found that they were generated by over 5 million Twitter users. To identify the location of these users, we again used the Twitter API to retrieve the text the users had entered into the location fields of their public profiles. Twitter's profile location field is nearly identical to that which we encountered with Flickr in Chapter 4, and as such we were able to use a nearly identical approach to convert the text values in these fields to machine-readable latitude and longitude pairs<sup>68</sup>.

---

<sup>66</sup> <http://alias-i.com/lingpipe/>

<sup>67</sup> <https://developers.google.com/translate/>

<sup>68</sup> We did not use the embedded geotags in tweets as we found that they made up only 0.7% of the tweets in our dataset, a number that has not increased much in the three years since we executed our original study [191]. We were also concerned that Twitter users who turn on geotagging would not be representative of the overall Twitter population.

Using our Wikipedia-based technique of converting text to coordinates, we were able to identify valid lat/lon locations for 588,248 users. Next, we leveraged spatial data available from ESRI and the United States Census to calculate the country and state (if in the United States) of the users. This process is known as *reverse geocoding*. As noted above, the country and state are the “classes” (i.e. geographic cultural memberships) to which we attempted to automatically assign Twitter users.

In order to avoid problems associated with having a small number of tweets for a given user, we further restricted our ground truth data to those users who had contributed ten or more tweets to our dataset. In doing so, we removed 484,449 users from consideration. We also required that all users in our dataset have a consistent country and state throughout the sample period. A tiny minority of users manually changed their location information during the sample period. In addition, a larger minority of users had their location changed automatically by Twitter clients like UberSocial<sup>69</sup>. This temporal consistency filter pruned an additional 4,513 users from consideration.

In the end, our ground truth data consisted of 99,296 users for whom we had valid country and state information and 10 or more tweets. This ground truth data was the sampling frame for deriving our training and test sets for all the machine learning experiments below.

## 5.2 Classification model

To classify the country and state geographic cultural memberships of Twitter users based on the content of their tweets, we developed a Multinomial Naïve Bayes (MNB) model [129]. The model accepts input in the form of a term vector with each dimension in the vector representing a

---

<sup>69</sup> <http://www.ubersocial.com/>

term and the value of the dimension representing the term count in a user’s tweets. We also tried using more advanced topic models such as Explicit Semantic Analysis [47], which is discussed in more detail in Chapter 6. However, a pilot study revealed that the simple term frequency (TF) MNB model greatly outperformed the more complex models. For computational efficiency, we settled on using a fixed-length 10,000-term vector to represent each user in all cases. We tried two different methods for picking which 10,000 terms to use. The first was the standard frequency-based selection model in which we picked the 10,000 most-common terms in our corpus. We called this algorithm “COUNT,” for its reliance on term counts.

We also developed a less naïve heuristic algorithm designed to select terms that would help discriminate between users from different geographic cultural memberships. This simple diversity mining algorithm, which we call “CALGARI,” is based on the intuition that a classifier will perform better if the model includes terms that are more likely to be employed by users from a particular region than users from the general population. It is our assumption that these terms will help our classifier more than the those selected by the COUNT algorithm, which includes many terms that are common in all countries or states considered (e.g. “lol”). The CALGARI algorithm calculates a score for each term present in the corpus according to the following formula:

$$CALGARI(t) = \begin{cases} 0 & \text{if } users(t) < MinU \\ \frac{\max(P(t \mid c = C))}{P(t)} & \text{if } users(t) \geq MinU \end{cases}$$

where  $t$  is the input term,  $users$  is a function that calculates the number of users who have used  $t$  at least once,  $MinU$  is an input parameter to filter out individual idiosyncrasies and spam (set to

either 2 or 5 in our experiments), and  $C$  is a geographic class (i.e. a state or country). The max function simply selects the maximum conditional probability of the term given each of the classes being examined. Terms are then sorted in descending order according to their scores and the top 10,000 are selected for the model. Finally, each user’s Twitter feed was represented as a term vector using this list of 10,000 terms as dimensions, populated by the feed’s term frequencies for each dimension.

A good example of the differences between CALGARI and COUNT can be found in the average term vector for each algorithm for users in Canada. Among the terms with the highest weights for the CALGARI algorithm were “Canada,” “Calgari,” “Toronto,” and “Hab”. On the other hand, the top ten for COUNT included “im,” “lol,” “love,” and “don’t.” Note that the CALGARI algorithm picked terms that are much more “Canadian” than those generated by the COUNT algorithm. This includes the #2 word “Calgari” (stemmed “Calgary”), which is the algorithm’s namesake.

### **5.3 Training and test sets**

In each experiment, we used a specific subset (described below) of the ground truth data as training data. Since the CALGARI algorithm and the COUNT algorithm both involve “peeking” at the ground truth data to make decisions about which dimensions to include in the term vectors, the use of independent test sets is vital. In all experiments, we split off 33% of the training data into test sets. These test sets were used only to evaluate the final performance of each model.

In both our country-scale and state-scale experiments, we implemented two different sampling strategies to create the training data from the ground truth data. The first, which we label as “UNIFORM,” generated training and test sets that exhibited a uniform distribution

across classes, or countries and states in this context. The experiments based on the UNIFORM data demonstrate the ability of our machine learning methods to tease out geographic cultural membership information in the absence of the current demographic trends on Twitter.

The second sampling strategy, which we label “RANDOM,” involved randomly selecting users for our training and test datasets. When using “RANDOM” data, the classifier considers the information that, for example, a Twitter user is much more likely to be from the United States than from Australia given population statistics and Twitter adoption rates. In other words, prior probabilities of each class (country or state) are considered.

## **5.4 Experiments**

We conducted a total of four experiments, each on a differently-sampled training and test set. In each experiment, we tested both the CALGARI and COUNT algorithms, reporting the accuracy for both. For the country-prediction experiments, we first focused on the UNIFORM sampling strategy. From our ground truth data, 2,500 users located in the United States, the United Kingdom, Canada, and Australia were randomly selected, resulting in 10,000 users total. These four countries were considered because there are less than 2,500 users in each of the other English-speaking countries represented among the 99,296 ground truth users. As noted above, 33% of these users were then randomly chosen for our test set and removed from the training set. The remainder of the training set was passed to one of two feature selection algorithms: CALGARI and COUNT. We then trained our Multinomial Naïve Bayes classifier and evaluated on the test set removed earlier. Next, we performed the same exercise, replacing the UNIFORM with the RANDOM sampling strategy, which selected 20,000 different users from our ground truth data, all of whom lived in one of the four countries listed above. Our state-prediction

experiments were roughly the same as our country experiments, with the only major difference occurring in the development of the UNIFORM datasets. Since the U.S. states range in population from California's 36+ million people to Wyoming's 0.5+ million people, our dataset was skewed in a similar fashion. We only had very limited data for small-population states like Wyoming. In fact, out of all our 99,296 ground truth users, we only had 31 from Wyoming. As such, we only included the 18 states with 500 or more users in our UNIFORM dataset.

## 5.5 Results

### 5.5.1 Country-prediction experiments

As shown in Table 5.5-a, for the UNIFORM sampling strategy, the best performing algorithm was CALGARI. Using CALGARI, we were able to correctly predict the geographic cultural membership of a Twitter user at the country scale 72.7% of the time, simply by examining that user's tweets. Since we considered four different countries in this case, one could achieve 25% accuracy by simply randomly guessing. Therefore, we also report in Table 5.5-a the accuracy of our classifier relative to the random baselines, which in the best case here was 291%

Sampling Strategy	Model Selection	Accuracy	Baseline	% of Baseline
<b>Country-Uniform-2500</b>	<b>Calgari</b>	<b>72.71%</b>	25%	291%
Country-Uniform-2500	Count	68.44%	25%	274%
<b>Country-Random-20K</b>	<b>Calgari</b>	<b>88.86%</b>	<b>82.08%</b>	<b>108%</b>
Country-Random-20K	Count	72.78%	82.08%	89%
<b>State-Uniform-500</b>	<b>Calgari</b>	<b>30.28%</b>	<b>5.56%</b>	<b>545%</b>
State-Uniform-500	Count	20.15%	5.56%	363%
State-Random-20K	Calgari	24.83%	15.06%	165%
<b>State-Random-20K</b>	<b>Count</b>	<b>27.31%</b>	<b>15.06%</b>	<b>181%</b>

Table 5.5-a: A summary of results from the country-scale and state-scale experiments. The better performing model selection algorithm is bolded for each experiment. The CALGARI result reported is the best generated by  $\text{Min}U = 2$  or  $\text{Min}U = 5$ .

(or 2.91x). With the RANDOM sampling strategy, we needed to use a different baseline. Since 82.08% of sampled users were from the U.S., one could achieve 82.08% accuracy simply by guessing “United States” for every user. However, even with these relatively decisive prior probabilities, the CALGARI algorithm was capable of bringing the accuracy level approximately 1/3 of the way to perfect performance. This represents roughly an 8.1% improvement.

### **5.5.2 State-prediction experiments**

Our classifier performed even better in our state-prediction experiments. As can be seen in Table 5.5-a, the classifier’s best UNIFORM accuracy relative to the random baseline was a great deal higher than in the country experiment. The same is true for the RANDOM dataset, which included users from all 50 states (even if there were only a dozen or so users from some states). The baselines were lower in each of these experiments because we considered more states than we did countries. The UNIFORM dataset included 18 states (or classes), resulting in a baseline of 1/18 (5.56%). The RANDOM dataset included all 50 plus the District of Columbia, with New York having the maximum representation at 15.06% of users. A baseline classifier could thus achieve 15.06% accuracy simply by selecting New York in every case.

## 5.6 Discussion

Table 5.5-a shows that in every single instance, the classifier was able to predict a user's country and/or state from the user's tweets at accuracies better than random. In most cases, the accuracy was several times better than random. This indicates that users incorporate a very strong signal of their geographic cultural memberships in their tweets, especially given the fact that we used a basic classifier and did not attempt to identify the optimal machine learning technique for this context. Indeed, since the publication of our work in this area, a number of researchers have shown that it is possible to accurately predict even finer-grained geographic cultural memberships from tweets (e.g. [22, 37, 126, 213]). For instance, Wing and Baldridge [213] were able to predict the location of Twitter users with a median error of only 439km.

By "looking under the hood" of our classifier, we can gain a better understanding of how users reveal their geographic cultural memberships in their tweets. Table 5.6-a contains some of the terms that were most predictive of a given cultural membership. Each term is listed with the country/state of which it was predictive as rough measures as well its "predictiveness," which is a simple ratio of the maximum conditional probability divided by the average of the non-

Stemmed Word	Country	"Predictiveness"	Stemmed Word	State	"Predictiveness"
"calgari"	Canada	419.32	"colorado"	Colorado	90.74
"brisban"	Australia	137.29	"elk"	Colorado	41.18
"coolcanuck"	Canada	78.28	"redsox"	Mass.	39.24
"afl"	Australia	56.24	"biggb"	Michigan	24.26
"clegg"	UK	35.49	"gamecock"	S. Carolina	16.00
"cbc"	Canada	29.40	"crawfish"	Louisiana	14.87
"yelp"	USA	19.08	"mccain"	Arizona	10.51

*Table 5.6-a: Some of the most predictive terms in our corpus along with the geographic feature of which they were predictive and a rough metric of their predictive power (explained in the text).*

maximum conditional probabilities. This can be roughly interpreted as the number of times more likely a word is to occur given that a person is from a specific region than from the average of the other regions in the dataset. In other words, an Arizonan is 10.51 times more likely to use the term “mccain” than a person from other states on average.

There appear to be four general categories of words that are particularly indicative of one’s geographic cultural memberships. As has been known in the social sciences for centuries (e.g. the gravity model [40]) and seen in our work in Chapter 4, people tend to interact with nearby places. In the Twitter context, this means that mentioning place names that are close to one’s location is very predictive of one’s location. In other words, tweeting about what you did in “Boston” narrows down your location significantly.

Tweeting about sports tends to be another common way of signaling one’s geographic cultural memberships. For instance, our classifier found that a user from Canada was six times more likely to tweet the word “hockey” than a user from any other country in our study. Similarly, Table 5.6-a shows that two of the most predictive terms for the state-prediction experiment were “redsox” (a reference to the Boston Red Sox) and “gamecock” (a reference to the University of South Carolina Gamecocks).

A third major category of predictive terms involves current events with specific geographic footprints. During the period of our data collection, several major events were occurring whose footprints corresponded almost exactly with the scales of our analyses. The classifier easily identified that terms like “Cameron,” “Brown,” and “Clegg” were highly predictive of users who were in the United Kingdom. Similarly, using terms related to the 2010 NBA playoffs was highly indicative of a user from the United States. More generally speaking, a model could theoretically utilize any regionalized phenomenon. For example, a tweet about a flood at a certain time (cf.

[190, 200]) could be used to locate a user to a very local scale.

Finally, regional vernacular such as “hella” (California) and “xx” (U.K.) were predictive of certain cultural memberships. It is our hypothesis that this category of predictive words helped our term frequency models perform better than the more complex topic models. It seems that the more abstract the topic model, the more it smoothes out the differences in spelling or slang. As noted by Clark [24] and Kramsch [108], however, such syntactic features can be powerful predictors of cultural memberships.

In this chapter, we have focused on demonstrating that UGC is sufficiently culturally contextualized such that the cultural memberships of a user can be accurately identified simply by examining the content she contributes. However, our work here has several additional implications. In a related project, we showed that users are often hesitant to provide real location information in their Twitter profile location fields. However, the results above indicate that no matter how hesitant they may be, their location is likely predictable if they tweet on a semi-regular basis. This of course has privacy implications if users wish to not reveal their location to unknown parties. However, it may also have privacy implications with respect to the revelation of other types of cultural memberships. If we can predict location from tweets, it is possible we may also be able to predict gender, sexual orientation, and a variety of socioeconomic status indicators as well. Indeed, researchers have already been able to identify the gender of Flickr users from the photo tags they use [157]. Investigating the degree to which additional cultural memberships can be mined from tweets, understanding the implications of this capability, and helping users defend themselves against these “inference attacks” [121] are important research directions that should be explored. The first of these directions is already receiving much attention, making the latter two even more important.

## 6 Implications for Existing Technologies: Semantic Relatedness Measures

*Note: Small portions of the text in this chapter originally appeared in our paper in the Proceedings of SIGIR 2012 [78]. All results are new to this thesis, however.*

Over the past three chapters, we have built up a substantial body of evidence in support of the hypothesis that user-generated content reflects the diverse cultural memberships of its contributors. In this chapter, we turn our attention to the effect of this diversity on the thousands of systems and algorithms that leverage UGC as a source of world knowledge. The research literature and product space of technologies dependent on UGC is massive. For instance, Twitter has been used to estimate gross national happiness [107], predict stock prices [14], and provide earthquake warnings [174]. Similarly, researchers have leveraged Flickr to induce ontologies [177], infer place and event semantics [166], and “help make sense of the world” through the generation of representative tags over geographic space [101].

It is Wikipedia, however, that has likely had the largest impact on computer science, primarily due to its contributions towards eliminating the infamous “knowledge acquisition bottleneck” (e.g. [26, 50, 114, 115]). For instance, within artificial intelligence (AI) and natural language processing (NLP), Wikipedia has been a game-changer for semantic web research (e.g. [13, 201]), has established new directions in information extraction (e.g. [3, 39, 165, 204]), has improved text categorization (e.g. [48, 49]), and has formed a substantial portion of the Google Knowledge Graph, which Google describes as a “one of the key breakthroughs behind the future of search” [106, 185]. Within human-computer interaction, Wikipedia is nearly as prominent, with it playing a role in augmented reality systems (e.g. [86, 88, 180, 181]), natural user

interfaces (e.g. [180]), virtual globes, ([225]) and other applications.

Overall, it is not a stretch to say that Wikipedia has become the “brains” of many modern computing technologies. However, the results we have seen thus far in this thesis beg the question, “Whose brain are we getting?” That is, we know now that the English Wikipedia represents world knowledge quite differently than the German Wikipedia, which is different than the Chinese Wikipedia and so on. If a Wikipedia-based technology switched from using one language edition as its “brains” to another, would the output be the same in both cases? More importantly, as the vast majority of Wikipedia-based technologies utilize only the English Wikipedia, are the results of these technologies biased toward the world view of English speakers?

In this chapter, we seek to provide evidence to help answer these questions. To do so, we focus on one particularly important class of Wikipedia-based technologies: semantic relatedness (SR) measures. Broadly speaking, SR measures return a value (usually between 0 and 1) that summarizes the number and strength of relationships between a given pair of concepts [82]. For instance, a good SR measure might return, say, 0.90 for the relatedness between the Minnesota State Fair and deep-fried candy bars, but only 0.10 for the relatedness between the Minnesota State Fair and caviar. In every case – high or low – the value returned is intended to be as close to human relatedness judgements as possible, and SR measures are evaluated by determining the extent to which they match these judgments. SR estimates are a low-level input to a plethora of technologies in natural language processing, artificial intelligence, and information retrieval, and have been applied in tasks such as word sense disambiguation, text summarization, indexing, information extraction, and even general search [11, 18, 78, 156, 163, 219]. They have also been leveraged in a number of more user-facing applications, for instance in the visualization of

conversations [11] as well as in our Omnipedia (Chapter 7) and Atlasify (Chapter 8) projects.

Wikipedia has had an enormous impact on the semantic relatedness literature. Prior to the arrival of Wikipedia, nearly all SR measures were based on WordNet. However, since Strube and Ponzetto introduced their *WikiRelate* SR measure in 2006 [192], the vast majority of the focus has been on Wikipedia-based SR measures, and the vast majority of this focus has been on SR measures that use the English Wikipedia.

In general, it has been assumed that the output of Wikipedia-based SR measures should be the same, regardless of the language edition used [74, 143]. When differences have been observed, they have been attributed to the quality or size of each language edition rather than the cultural contextualization inherent to their content (which is largely surmised to not exist). In the language of Chapter 3, the assumption that SR output should be language-neutral is an applied instance of the global consensus hypothesis. The social science literature, on the other hand, suggests that the global diversity hypothesis is relevant in this context just as much as it is in the context of Chapter 3. In particular, it has long been known that different language-defined communities assess the semantic relatedness between certain concepts quite differently [108]. This suggests that Wikipedia-based SR measures, which are designed to match human judgements, should output variable values depending on the language edition being used. The goal of this chapter is, in summary, to determine for which of these two hypotheses there is more support.

## 6.1 Semantic Relatedness Measures

There are many SR measures in the literature. Even limiting our attention to Wikipedia-based SR measures, there are still quite a few to consider (e.g. [47, 136, 137, 163, 192]). We

focus on three of the most well-known Wikipedia-based SR measures – *WikiRelate* [192], *MilneWitten* [136, 137], and *Explicit Semantic Analysis* [47, 50].

Aside from their prominence in the literature, another benefit of this set of SR measures is that each uses a different Wikipedia “lexical resource” [220] and each of these resources detects a different type of relationship between concepts. *WikiRelate* uses Wikipedia Category Graphs (WCGs) (see Section 3.2.1) as its lexical resource. Due to their nature as categorization structures, WCGs contain mostly hypernymy/hyponymy relationships (i.e. “is-a” and “has-a” relations). These are the relationships upon which *WikiRelate* makes its assessments of the semantic relatedness between two concepts.

*MilneWitten* leverages Wikipedia Article Graphs (WAGs) as its lexical resource. WAG-based measures are more capable of discovering *non-classical* relations [18], such as *graduatedFrom* and *failedOutOf*. *MilneWitten*<sup>70</sup> works by comparing the set of articles in a given language edition that link to the first concept in a concept pair with the set of articles that link to the second concept. The intuition is that if these two sets overlap extensively, the two concepts should be given a high relatedness estimate (according to the language edition).

*Explicit Semantic Analysis* (ESA), the most well-known of the SR measures considered here, uses the Wikitext resource, or the actual plain text on Wikipedia pages. ESA models the two input concepts “in terms of Wikipedia-based [articles]” [47]. The measure is “explicit” because Wikipedia articles, which are understandable to humans, define this modeling space. ESA’s use of real concepts stands in stark contrast to the abstract concepts of methods like Latent Semantic Analysis (LSA). The types of relationships ESA considers when calculating an SR

---

70 Our implementation of *MilneWitten* is slightly simplified from Milne and Witten’s final measure; we only consider their “Google Distance-inspired” metric. They were able to gain modest but insignificant improvements by averaging in their “TFIDF-inspired” metric.

estimate have been called “distributional” relationships [18], although another way of describing them might be “co-occurrence” relationships.

In this chapter, we also use two additional and straightforward semantic relatedness measures that we introduced in our Atlasify work [78]: *OutlinkOverlap* and *WAGDirect*. These two measures are designed to capture relationship types built into the WAG of each language edition not considered by *MilneWitten*. *OutlinkOverlap* uses the principle that, broadly speaking, if two concepts share a significant number of outlinks in a given language edition, then the two concepts are quite related (in that language edition). In other words, *OutlinkOverlap* is the same as *MilneWitten*, but considers sets of outlinks rather than inlinks. *WAGDirect* captures the notion that if in a given language edition a concept links directly to another concept and/or vice versa, this link obviously represents a significant relationship between the concepts. *WAGDirect* also weights links that occur in the “gloss,” or the first paragraph of articles<sup>71</sup>, more than those that appear further down the page.

The fact that each of these five SR measures use and understand different types of relationships allows us to investigate whether the cultural contextualization of encyclopedic world knowledge affects semantic relatedness calculations regardless of the type of relationship considered. The specifics of how we conducted this investigation and the investigation’s end results are the topics of the two sections that follow.

## 6.2 Experiment

At a high level, the design of our semantic relatedness experiment is relatively straightforward. Given two concepts  $c_1$  and  $c_2$ , our goal is to determine if  $SR_{ENGLISH}(c_1, c_2) =$

---

<sup>71</sup> This feature is implemented using the location property of links in WikAPIdia.

$SR_{JAPANESE}(c_1, c_2) = SR_{GERMAN}(c_1, c_2)$ . In other words, we seek to establish the extent to which the same SR measure will output similar or different results when we vary the language edition it is using as world knowledge (i.e. its “brains”). For instance, does  $MilneWitten_{ENGLISH}(\text{Germany}, \text{Fascism}) = MilneWitten_{GERMAN}(\text{Germany}, \text{Fascism})$ , or does the diversity and cultural contextualization in these language editions’ content alter  $MilneWitten$ ’s estimate of the relatedness between the two concepts?

Below, we address several of the more important experimental design decisions we made, namely those with regard to concept sampling and comparison metrics. Finally, prior to discussing results, we demonstrate that our versions of the SR measures described above have been implemented correctly, a necessary precondition to investigating their output when using different language editions as world knowledge.

### 6.2.1 Concept Sampling

In their paper comparing SR measures based on Wikipedia to those based on WordNet, Zesch and Gurevych make the important observation that “the real distribution of (semantic) relatedness values is largely unknown.” [219]. Given this major open question, it is not a surprise that there is no consensus in the literature as to the types of concepts on which to focus when doing SR research. Some researchers have manually developed sets of concept pairs that were hypothesized to uniformly span the semantic relatedness spectrum (e.g. [43, 152]). Others have developed automated approaches to generating these uniform distributions, e.g. using term co-occurrence in various corpora (e.g. [163]). At the same time, many applications of semantic relatedness (e.g. Atlasify in Chapter 8) require an SR measure to be successful in contexts where the expected SR between two concept pairs is at or near zero. That is, these applications operate

in an environment in which SR distributions have a very strong positive skew. However, regardless of the distribution of samples, there *is* consensus that the concepts used for evaluation and development should be common rather than highly esoteric. In other words, concept pairs nearly always resemble pairs like (war, peace) versus, say, (Mister Philippines 2008, List of University of Michigan alumni) (Chapter 3).

In our experiment, we adopt an application-focused approach to selecting concept pairs. That is, our sampling procedure results in the inclusion of many concepts that were expected to have a low semantic relatedness. The idea here is to sample concept pairs that best allow us to evaluate the extent to which the cultural contextualization of user-generated content affects real systems, although we do shed some light on the more theoretical issue of its variable effect across the SR spectrum later in this chapter. Despite our sampling procedure's bias toward the low end of the SR spectrum, it was also carefully designed to afford a high level of "commonness" for all concepts considered, as we will describe below. Finally, our procedure also enables a very large number of concept pairs to be sampled under these parameters, an additional benefit given that SR evaluation datasets tend to be somewhat small.

We are able to achieve these three properties in our final sample by leveraging the concept-level diversity work from Chapter 3. First, to ensure a high level of commonness, we used the 10,853 concepts in the "global encyclopedic core" (Section 3.4) as our concept sampling frame. Recall that these are the concepts that have articles in all 25 language editions considered in this thesis. The fact that these concepts appear in so many languages guarantees a certain level of universality. For example, concepts in our sample include Deutsche Bahn, David Duchovny, and the Dow Jones Industrial Average.

To generate a set of concept pairs with an expected positive skew based on these global

concepts, we used a basic random sampling approach. Specifically, we first randomly selected a global concept  $c_1$ , and then again randomly selected 10 other concepts to serve as  $c_2s$ . We then repeated this process 5,000 times, resulting in 50,000 concept pairs.

We did have to place one important restriction on our sample: all pairs that included a temporal concept as either  $c_1$  or  $c_2$  (e.g. years, decades, months, days of the week)<sup>72</sup> were filtered out. As noted in Section 3.6, English and German do not have the same linking practices with regard to these concepts as the other language editions. Since three of our SR measures are WAG-based, including these concepts would have conflated the communities' decisions to not link to these entities with diversity in the content of each language edition. The removal of temporal concept pairs reduced our total sample size to 40,490 concept pairs.

### **6.2.2 Comparison Metrics**

Another important experimental design consideration is the method of comparing the output of language edition-varied SR measures. Here, the semantic relatedness literature provides a great deal of guidance. Almost without exception, when researchers seek to evaluate the performance of a new SR measure, they do so by calculating the correlation between their measure's output and the “human gold-standard” semantic relatedness judgements in several well-known benchmark datasets. We adopt the same approach here, except instead of comparing algorithmically-generated and human-generated SR estimates, we do the same with the output of a given Wikipedia-based SR measure using two different language editions as world knowledge.

While there is widespread consensus on the use of correlation for evaluation in the SR literature, there is some disagreement over the type of correlation that is most effective.

---

<sup>72</sup> We were able to do this by excluding all concepts that appear in the “Time” and “Julian Year” reference system in the Atlasify project (Chapter 8).

Primarily, some papers use Pearson's  $r$  while others use Spearman's  $\rho$ . Here, following Zesch and Gurevych [219], we focus on Spearman's for three reasons:

1. Pearson's  $r$  is very sensitive to outliers, and several of the SR measures considered are prone to the occasional generation of outliers.
2. Pearson's  $r$  assumes that SR estimates are measured on a linear scale, something for which there is some evidence against [18].
3. Pearson's  $r$  is a measure of linear association. We had no *a priori* assumption that the association between the outputs of a language edition-varied SR measure would be linear.

Spearman's correlation coefficient is robust against all of these concerns. However, it does have one important disadvantage in the context of this experiment: it does not handle ties well. For these reasons, we also provide Kendall's  $\tau_B$  results where appropriate. Kendall's  $\tau_B$  makes adjustments to better account for ties.

### 6.2.3 SR Measure Implementations

We implemented versions of *MilneWitten*, *Explicit Semantic Analysis*, *OutlinkOverlap*,

<b>Semantic Relatedness Measure</b>	<b>Thesis Implementation Using Oct/Nov 2012 data</b>	<b>Atlasify Implementation Using Jan 2011 data</b>
Explicit Semantic Analysis	0.71	0.71
MilneWitten	0.67	0.68
OutlinkOverlap	0.63	0.61
WAGDirect	0.60	0.60
WikiRelate	0.40	0.52

*Table 6.2-a: Performance of our current implementations of the five SR measures considered in this chapter versus those used in our Atlasify project on the Atlasify240 human gold standard dataset. Note that for four of the five measures, the performance is nearly identical despite the fact that almost two years of growth in the English Wikipedia had occurred between the database dumps of the two tests.*

*WAGDirect*, and *WikiRelate* in WikAPIdia (Section 3.12). These implementations are modifications of those used in our Atlasify system, which we robustly validated to ensure that their accuracy matched that in their original publications (where possible). However, because we made important modifications – primarily to take advantage of the new abstracted data source functionality in WikAPIdia (Section 3.12) – we compared the output of our new versions against those in the Atlasify system using our *Atlasify240* human gold standard SR dataset. The *Atlasify240* dataset is English-only – like most other SR datasets – and thus we only used the English Wikipedia for this validation experiment. *Atlasify240* is ideal for validation in this case as it explicitly measures the semantic relatedness between already-disambiguated English Wikipedia articles, removing issues related to term disambiguation that are not in the scope of this chapter.

Despite the fact that almost two years had passed between the dates of the database dumps used in the Atlasify project versus those used here, the performance against the *Atlasify240* dataset was nearly identical in four out of five cases (Table 6.2-a). Not only does this establish that our new implementations are correct, but it demonstrates that the extensive growth in the English Wikipedia over the past two years has had little to no effect on the output of SR algorithms that use the English Wikipedia as world knowledge. This is an informal confirmation of a finding by Zesch and Gurevych [218], who established that the same has been true in the German Wikipedia over a much longer time frame. This longitudinal within-language consistency provides important context for our cross-language results in the following section.

The one SR measure that did not perform as well as its Atlasify counterpart is our current implementation of WikiRelate. Here, we identified a decrease in performance of 0.12. This is roughly consistent with Zesch and Gurevych’s work, in which they found that of the three lexical

resources in Wikipedia that are commonly used to calculate SR, the category graphs provide for the most variable results, although they showed that this variation has been getting smaller over time. However, further exploring our *WikiRelate* performance, we established that its accuracy is better on other datasets. For instance, using the *WordSim353* dataset, our implementation of *WikiRelate* performed as well as the original published version.

## 6.3 Results and Discussion

Following the finalization of our experimental design, we began the experiment by inputting our 40,490 concept pairs into all five validated SR measure implementations set to use the English Wikipedia as world knowledge. Next, we repeated this process, swapping out the English Wikipedia for the German Wikipedia. We then did the same for each of the 23 other language editions of Wikipedia considered in Chapter 3 (see Section 3.2.3). Finally, we calculated the pairwise Spearman's  $\rho$  and Kendall's  $\tau_B$  of all within-SR measure semantic relatedness estimates. For instance, for *MilneWitten*, we calculated  $\rho$  and  $\tau_B$  for *MilneWitten<sub>ENGLISH</sub>* and *MilneWitten<sub>GERMAN</sub>*, the  $\rho$  and  $\tau_B$  for *MilneWitten<sub>ENGLISH</sub>* and *MilneWitten<sub>FRENCH</sub>*, and repeated for all other 298 language edition pairs<sup>73</sup>.

### 6.3.1 Basic Results

In every case – regardless of SR measure – we saw substantial differences between the estimates generated by the same SR measure when varying the language edition used as world knowledge. Moreover, the fluctuation of estimates was largely consistent with the cultural similarities and differences between the language editions observed in Chapter 3. In other words, we saw that the cultural context reflected in Wikipedia-based user-generated content has a major

---

<sup>73</sup> Since all SR measures considered here are symmetric, we only needed to calculate  $\rho$  and  $\tau_B$  for 300 language edition pairs rather than all 600 pairs that would have been necessary otherwise.

effect on the output of several important algorithms that leverage this user-generated content as their source of world knowledge.

Tables 6.3-a and 6.3-b show the minimum, maximum, and mean pairwise  $\rho$  and  $\tau_B$  for each SR measure. The maximum  $\rho$  for all language pairs and all SR measures was only 0.578, which is the correlation between  $ESA_{ITALIAN}$  and  $ESA_{SPANISH}$ . Moreover, from our  $\tau_B$  results it appears that our  $\rho$  results are inflated due to the substantial number of ties, specifically the large number of zeros that occur in the bulk of mass of the positively skewed distribution. For instance, 72.1% of SR estimates were zero for  $ESA_{GERMAN}$ , and the number goes up for the smaller language editions. When using the tie-corrected  $\tau_B$  correlations,  $WAGDirect$  based on two Scandinavian languages,  $WAGDirect_{NORWEGIAN}$  and  $WAGDirect_{DANISH}$ , are the most correlated pair ( $\tau_B = 0.540$ ). Note also that the average  $\tau_B$ s for the most commonly used SR measures – *MilneWitten*, *Explicit Semantic Analysis*, and *WikiRelate* – are substantially lower than their  $\rho$  counterparts. The average  $\tau_B$  correlation for *Explicit Semantic Analysis* – easily the most-applied of the five SR measures – is only 0.373.

Recall that SR measures are used in a variety of low-level and user-facing technologies. What these initial results suggest is that any SR-based research project or system that uses just one language edition as its source of world knowledge is not accurately representing world knowledge as it is understood globally. As we engage in further analyses of our results below, this conclusion will become more robust against a variety of factors.

<b>SR Measure</b>	<b>Min <math>\rho</math></b>	<b>Mean <math>\rho</math></b>	<b>Max <math>\rho</math></b>
Explicit Semantic Analysis*	0.182 English/Slovak	0.408	0.578 Italian/Spanish
MilneWitten**	0.318 English/Slovak	0.467	0.555 English/French
OutlinkOverlap**	0.250 English/Slovak	0.358	0.483 French/Spanish
WAGDirect**	0.304 Hebrew/Slovak	0.410	0.540 Norwegian/Danish
WikiRelate	0.123 Indonesian/Japanese	0.329	0.520 English/Japanese

Table 6.3-a: The min, mean, and max Spearman's for all 300 symmetric language pairs considered in our experiment.

\* Throughout this chapter, the Chinese Wikipedia was removed from ESA summary statistics as it had the lowest correlations across the board. This is likely due to an issue with our ESA implementation and the Chinese Wikipedia. We are working to resolve the issue.

\*\* All WAG-based measures in this chapter were calculated on the parseable WAG of each language edition. In every case, sub-articles were considered, an important concern given the prominence of sub-articles among global concepts.

<b>SR Measure</b>	<b>Min <math>\tau_B</math></b>	<b>Mean <math>\tau_B</math></b>	<b>Max <math>\tau_B</math></b>
Explicit Semantic Analysis*	0.185 Catalan/Slovak	0.373	0.512 Italian/Spanish
MilneWitten**	0.286 English/Slovak	0.441	0.512 Portuguese/Spanish
OutlinkOverlap**	0.207 English/Slovak	0.304	0.390 Czech/Slovak
WAGDirect**	0.305 Hebrew/Slovak	0.410	0.540 Norwegian/Danish
WikiRelate	0.139 Indonesian/German	0.382	0.502 English/Danish

Table 6.3-b: The min, mean, and max Kendall's  $\tau_B$  for all 300 symmetric language pairs considered in our experiment.

### 6.3.2 Cross-language SR versus Cross-time SR

One way to continue our exploration into the possible bias of English Wikipedia-only technologies is to put the above absolute correlation coefficients into additional context. To do so, we compared the correlations of language edition-varied SR measures with those of time-varied SR measures in which the language edition is held constant. Specifically, using concept pairs from the *Atlasify240* dataset, we calculated the  $\rho$  for the output of all  $SR_{ENGLISH}$  measures when  $t_1$  = January 2011 and  $t_2$  = November 2012. We then compared these  $\rho$  values to the maximum  $\rho$  values for language edition-varied SR measures from Table 6.3-a. As shown in Table 6.3-c, our results strongly demonstrate that SR measures are substantially more consistent over time than they are across language editions. All of the differences in Table 6.3-a are significant at at least the  $p < 0.01$  level, a result that is all the more remarkable when considering the fact that the English Wikipedia has grown by over 668,000 (19.3%) articles and over 24 million (23.1%) links during the time between  $t_1$  and  $t_2$ . This represents an increase larger than the size of the entire corresponding resources of many of the language editions considered here. This growth represents an increase in general encyclopedic world knowledge, it also represents new knowledge that has come into existence in the past two years. For instance, the *Atlasify240*'s

Semantic Relatedness Measure	Max Between-Language Spearman's $\rho$	Max Between-Snapshot Spearman's $\rho$
Explicit Semantic Analysis	0.578	<b>0.841***</b>
MilneWitten	0.570	<b>0.909***</b>
OutlinkOverlap	0.526	<b>0.853***</b>
WAGDirect	0.605	<b>0.907***</b>
WikiRelate	0.520	<b>0.644**</b>

Table 6.3-c: The maximum between-language Spearman's correlation coefficients compared to the maximum between-snapshot coefficients for the English Wikipedia. The two snapshots considered are from January 2011 and November 2012. \*\* indicates  $p < 0.01$ , \*\*\* indicates  $p < 0.0001$ .

concept pair (Apple, Inc., Google) has likely been affected by new information related to current events (e.g. maps). This new knowledge is likely responsible for some of the small differences between time periods.

As noted above, Zesch and Gurevych have established that, at least in the case of the German Wikipedia, a variety of Wikipedia-based SR measures have produced surprisingly consistent output over a long period of time. They saw, for instance, that an *ESA*-like SR measure produced roughly the same output using a 2005 version of the language edition as it did using a 2008 version, which had more than twice as many articles. This suggests that if we were to extend our time window for the English Wikipedia and/or do a similar analysis with other language editions, we would still see that SR measure output varies far more across language editions than it does within the same language edition over time.

### **6.3.3 Pairwise Comparisons**

Another means by which one can investigate the differences in the SR estimates across language editions is to examine the 25-language-edition-by-25-language-edition symmetric matrix of pairwise correlations. Tables 6.3-d and 6.3-e show the Spearman's correlation matrices for *Explicit Semantic Analysis* and *MilneWitten*, respectively. A dominant trend in these tables is that language editions from similar language-defined cultures tend to have higher correlation coefficients. Examples of this phenomenon are numerous:

- *ESA*: The highest  $\rho$  for Catalan is Spanish ( $p < 0.0001$ ).
- *ESA*: The two highest  $\rho$ s for Danish are Swedish and Norwegian. The difference between these  $\rho$ s and all others for Danish is almost 0.05 and is significant ( $p < 0.0001$ ).
- *ESA*: The two highest  $\rho$ s for Norwegian are Swedish and Danish ( $p < 0.0001$ ).
- *ESA*: By far the highest  $\rho$  for Ukrainian is Russian. It is a full 0.10 higher than that for

any other language edition ( $p < 0.0001$ ).

- *MilneWitten*: The highest  $\rho$  for Danish is Norwegian ( $p < 0.0001$ ).
- *MilneWitten*: The highest  $\rho$  for Korean is Chinese ( $p < 0.0001$ ).
- *MilneWitten*: The highest  $\rho$  for Ukrainian is Russian ( $p < 0.0001$ )

## Explicit Semantic Analysis

	Cata	Chin	Czec	Dani	Dutc	Eng	Fin	Fren	Ger	Heb	Hun	Indo	Ital	Japa	Kor	Nor	Pol	Port	Rom	Rus	Slvk	Spa	Swe	Turk	Ukr	Min	Avg	Max	
<b>Catalan</b>		.079	.392	.362	.387	.434	.356	.461	.403	.335	.387	.372	.457	.366	.336	.416	.392	.459	.391	.412	.204	.495	.397	.367	.335	.079	.375	.495	
<b>Chinese</b>	.079		.128	.116	.141	.104	.134	.116	.148	.109	.140	.126	.117	.130	.150	.125	.137	.116	.124	.120	.118	.127	.131	.113	.117	.079	.124	.150	
<b>Czech</b>	.392	.128		.419	.443	.429	.419	.442	.446	.372	.442	.443	.468	.424	.378	.451	.475	.457	.425	.435	.301	.458	.452	.420	.388	.128	.413	.475	
<b>Danish</b>	.362	.116	.419		.425	.372	.420	.395	.422	.368	.399	.415	.409	.373	.365	.493	.414	.406	.406	.379	.302	.401	.471	.393	.361	.116	.387	.493	
<b>Dutch</b>	.387	.141	.443	.425		.458	.408	.482	.486	.373	.426	.440	.476	.418	.376	.465	.457	.457	.416	.430	.275	.471	.462	.404	.357	.141	.414	.486	
<b>English</b>	.434	.104	.429	.372	.458		.395	.566	.522	.350	.407	.427	.530	.484	.357	.424	.442	.518	.424	.470	.182	.545	.438	.403	.351	.104	.418	.566	
<b>Finnish</b>	.356	.134	.419	.420	.408	.395		.397	.429	.379	.413	.411	.421	.399	.377	.438	.420	.415	.380	.395	.287	.405	.454	.403	.370	.134	.389	.454	
<b>French</b>	.461	.116	.442	.395	.482	.566	.397		.516	.363	.418	.428	.544	.456	.365	.441	.458	.520	.441	.479	.243	.550	.442	.407	.377	.116	.429	.566	
<b>German</b>	.403	.148	.446	.422	.486	.522	.429	.516		.365	.431	.423	.507	.449	.368	.471	.463	.472	.416	.458	.256	.498	.488	.391	.369	.148	.425	.522	
<b>Hebrew</b>	.335	.109	.372	.368	.373	.350	.379	.363	.365		.347	.385	.385	.367	.360	.381	.375	.366	.350	.365	.261	.376	.377	.368	.343	.109	.351	.385	
<b>Hungarian</b>	.387	.140	.442	.399	.426	.407	.413	.418	.431	.347		.421	.436	.402	.370	.425	.442	.428	.408	.416	.298	.434	.420	.405	.362	.140	.395	.442	
<b>Indonesian</b>	.372	.126	.443	.415	.440	.427	.411	.428	.423	.385	.421		.444	.434	.401	.429	.436	.450	.428	.404	.307	.444	.434	.440	.367	.126	.405	.450	
<b>Italian</b>	.457	.117	.468	.409	.476	.530	.421	.544	.507	.385	.436	.444		.456	.389	.464	.475	.532	.442	.476	.246	.578	.462	.420	.375	.117	.438	.578	
<b>Japanese</b>	.366	.130	.424	.373	.418	.484	.399	.456	.449	.367	.402	.434	.456		.402	.398	.431	.444	.401	.425	.229	.457	.418	.405	.361	.130	.397	.484	
<b>Korean</b>	.336	.150	.378	.365	.376	.357	.377	.365	.368	.360	.370	.401	.389	.402		.383	.379	.377	.353	.377	.268	.385	.380	.361	.344	.150	.358	.402	
<b>Norwegian</b>	.416	.125	.451	.493	.465	.424	.438	.441	.471	.381	.425	.429	.464	.398	.383		.450	.448	.411	.421	.295	.469	.516	.417	.367	.125	.416	.516	
<b>Polish</b>	.392	.137	.475	.414	.457	.442	.420	.458	.463	.375	.442	.436	.475	.431	.379	.450		.454	.423	.456	.302	.471	.444	.413	.380	.137	.416	.475	
<b>Portuguese</b>	.459	.116	.457	.406	.457	.518	.415	.520	.472	.366	.428	.450	.532	.444	.377	.448		.454	.450	.446	.257	.567	.439	.422	.370	.116	.428	.567	
<b>Romanian</b>	.391	.124	.425	.406	.416	.424	.380	.441	.416	.350	.408	.428	.442	.401	.353	.411	.423	.450	.397	.277	.446	.415	.407	.364	.124	.392	.450		
<b>Russian</b>	.412	.120	.435	.379	.430	.470	.395	.479	.458	.365	.416	.404	.476	.425	.377	.421		.456	.446	.397	.230	.478	.423	.382	.479	.120	.406	.479	
<b>Slovak</b>	.204	.118	.301	.302	.275	.182	.287	.243	.256	.261	.298	.307	.246	.229	.268	.295		.302	.257	.277	.230		.253	.275	.279	.253	.118	.258	.307
<b>Spanish</b>	.495	.127	.458	.401	.471	.545	.405	.550	.498	.376	.434	.444	.578	.457	.385	.469	.471	.567	.446	.478	.253		.450	.422	.381	.127	.440	.578	
<b>Swedish</b>	.397	.131	.452	.471	.462	.438	.454	.442	.488	.377	.420	.434	.462	.418	.380	.516		.444	.439	.415	.423	.275	.450		.404	.369	.131	.415	.516
<b>Turkish</b>	.367	.113	.420	.393	.404	.403	.403	.407	.391	.368	.405	.440	.420	.405	.361	.417		.413	.422	.382	.279	.422	.404		.372	.113	.384	.440	
<b>Ukrainian</b>	.335	.117	.388	.361	.357	.351	.370	.377	.369	.343	.362	.367	.375	.361	.344	.367		.380	.370	.364	.479	.253	.381	.369	.372		.117	.355	.479
<b>Minimum</b>	.079	.079	.128	.116	.141	.104	.134	.116	.148	.109	.140	.126	.117	.130	.160	.125	.137	.116	.124	.120	.118	.127	.131	.113	.117	.079	.122	.150	
<b>Average</b>	.375	.124	.413	.387	.414	.418	.389	.429	.425	.351	.395	.405	.438	.397	.358	.416	.416	.428	.392	.406	.258	.440	.415	.384	.355	.122	.385	.470	
<b>Maximum</b>	.495	.150	.475	.493	.486	.566	.454	.566	.522	.385	.442	.450	.578	.484	.402	.516		.475	.567	.450	.479	.307	.578	.516	.440	.479	.150	.470	.578

Example: The language edition whose ESA output most agreed with Danish's ESA output was Norwegian ( $\rho = 0.493$ )

Example: ESA based on Norwegian and Danish agreed with the output of Swedish-based ESA at a rate significantly greater than ESA based on any other language edition.

Table 6.3-d: Pairwise correlations between the output of Explicit Semantic Analysis using the the 25 language editions considered in this thesis as world knowledge. The bold and red cells are those mentioned in the text.

## MilneWitten

	Cata	Chin	Czec	Dani	Dutc	Eng	Fin	Fren	Ger	Heb	Hun	Indo	Ital	Japa	Kor	Nor	Pol	Port	Rom	Rus	Slvk	Spa	Swe	Turk	Ukr	Min	Avg	Max	
<b>Catalan</b>		.502	.506	.465	.508	.448	.479	.494	.466	.494	.502	.471	.515	.486	.470	.496	.490	.530	.478	.497	.426	.535	.498	.471	.494	.426	.488	.535	
<b>Chinese</b>	.502		.481	.443	.489	.446	.477	.475	.469	.479	.483	.476	.494	.512	.496	.472	.481	.514	.467	.496	.407	.481	.486	.471	.470	.407	.478	.514	
<b>Czech</b>	.506	.481		.476	.495	.415	.480	.466	.465	.482	.502	.464	.486	.473	.465	.465	.502	.487	.495	.486	.484	.496	.462	.493	.458	.484	.415	.479	.502
<b>Danish</b>	.465	.443	.476		.449	.355	.475	.397	.399	.450	.452	.449	.428	.416	.460	.499	.438	.435	.464	.427	.440	.403	.474	.447	.446	.355	.441	.499	
<b>Dutch</b>	.508	.489	.495	.449		.467	.484	.508	.501	.479	.496	.445	.513	.499	.453	.497	.507	.513	.467	.505	.412	.501	.499	.461	.465	.412	.484	.513	
<b>English</b>	.448	.446	.415	.355	.467		.409	.556	.533	.426	.431	.363	.528	.514	.365	.413	.469	.504	.386	.511	.318	.550	.422	.380	.394	.318	.442	.556	
<b>Finnish</b>	.479	.477	.480	.475	.484	.409		.452	.450	.463	.489	.461	.467	.467	.471	.488	.492	.481	.475	.469	.437	.449	.512	.473	.409	.470	.512		
<b>French</b>	.494	.475	.466	.397	.508	.556		.452	.548	.472	.474	.402	.552	.521	.408	.452	.507	.525	.423	.528	.355	.547	.465	.418	.444	.355	.475	.556	
<b>German</b>	.466	.469	.465	.399	.501	.533	.450	.548		.453	.473	.386	.522	.510	.398	.461	.498	.499	.419	.506	.352	.521	.467	.405	.435	.352	.464	.548	
<b>Hebrew</b>	.494	.479	.482	.450	.479	.426	.463	.472	.453		.469	.440	.487	.478	.423	.467	.465	.500	.438	.476	.400	.401	.451	.427	.467	.400	.460	.500	
<b>Hungarian</b>	.502	.483	.502	.452	.496	.481	.489	.474	.473	.469		.452	.491	.481	.469	.488	.494	.499	.500	.495	.437	.482	.490	.485	.480	.431	.480	.502	
<b>Indonesian</b>	.471	.476	.464	.449	.445	.363	.461	.402	.386	.440	.452		.432	.431	.478	.479	.431	.458	.457	.431	.418	.413	.464	.481	.448	.363	.443	.481	
<b>Italian</b>	.515	.494	.486	.428	.513	.528	.467	.552	.522	.487	.491	.432		.517	.433	.489	.512	.543	.449	.521	.385	.548	.476	.442	.469	.385	.487	.552	
<b>Japanese</b>	.486	.512	.473	.416	.499	.514	.467	.521	.510	.478	.481	.431	.517		.445	.470	.490	.516	.441	.520	.384	.519	.475	.438	.458	.384	.478	.521	
<b>Korean</b>	.470	.496	.465	.460	.453	.365	.471	.408	.398	.423	.469	.478	.433	.445		.468	.437	.455	.474	.435	.444	.414	.467	.475	.460	.365	.448	.496	
<b>Norwegian</b>	.496	.472	.502	.499	.497	.413	.488	.452	.461	.467	.488	.479	.489	.470	.468		.478	.491	.485	.478	.437	.458	.506	.475	.413	.476	.506		
<b>Polish</b>	.490	.481	.487	.438	.507	.469	.492	.507	.498	.465	.494	.431	.512	.490	.437	.478		.502	.460	.518	.406	.502	.489	.452	.463	.406	.478	.518	
<b>Portuguese</b>	.530	.514	.495	.435	.513	.504	.481	.525	.499	.500	.499	.458	.543	.516	.455	.491	.502		.460	.522	.402	.548	.490	.462	.471	.402	.492	.548	
<b>Romanian</b>	.478	.467	.486	.464	.467	.386	.475	.423	.419	.438	.500	.457	.449	.441	.474	.485	.460	.460		.446	.452	.432	.472	.504	.468	.386	.458	.504	
<b>Russian</b>	.497	.496	.484	.427	.505	.511	.469	.528	.506	.476	.495	.431	.521	.520	.435	.478	.518	.522	.446		.390	.521	.488	.452	.521	.390	.485	.528	
<b>Slovak</b>	.426	.407	.496	.440	.412	.318	.437	.355	.352	.400	.437	.418	.385	.384	.444	.437	.406	.402	.452	.390		.360	.433	.439	.415	.318	.410	.496	
<b>Spanish</b>	.535	.481	.462	.403	.501	.550	.449	.547	.521	.461	.482	.413	.548	.519	.414	.458	.502	.548	.432	.521		.360	.466	.427	.438	.360	.477	.550	
<b>Swedish</b>	.498	.486	.493	.474	.499	.422	.512	.465	.467	.451	.490	.464	.476	.475	.467	.506	.489	.490	.472	.483		.433	.466		.470	.468	.422	.476	.512
<b>Turkish</b>	.471	.471	.458	.447	.461	.380	.473	.418	.405	.427	.485	.481	.442	.438	.475	.475	.452	.462	.504	.452		.439	.427	.470		.457	.380	.453	.504
<b>Ukrainian</b>	.494	.470	.484	.446	.465	.394	.473	.444	.435	.467	.480	.448	.469	.458	.460	.475	.463	.471	.468	.521		.415	.438	.468	.457		.394	.461	.521
<b>Minimum</b>	.426	.407	.415	.355	.412	.318	.409	.355	.352	.400	.431	.363	.385	.384	.365	.413	.406	.402	.386	.390		.318	.300	.422	.380	.394	.318	.386	.481
<b>Average</b>	.488	.478	.479	.441	.484	.442	.470	.475	.464	.460	.480	.443	.487	.478	.448	.476	.478	.492	.458	.485		.410	.477	.476	.453	.461	.386	.467	.519
<b>Maximum</b>	.535	.514	.506	.499	.515	.556	.512	.556	.548	.500	.502	.481	.552	.521	.496	.506	.518	.548	.504	.528		.496	.550	.512	.504	.521	.481	.519	.556

Example:  $\rho$  was highest between the output English-based and French-based MilneWitten.

Example: For Swedish-based MilneWitten,  $\rho$  was highest when comparing against MilneWitten based on other Scandinavian languages.

Table 6.3-e: Pairwise correlations between the output of MilneWitten using the the 25 language editions considered in this thesis as world knowledge. The bold and red cells are those mentioned in the text.

Another pattern in the tables is that the shared cultural context of similar language-defined cultures, manifest as higher  $\rho_s$ , is not the only cause of the variation in the correlation coefficients. It appears that (1) language edition size and (2) language edition quality also play a role. This is especially true in the *MilneWitten* table, where the cultural signal is less prominent than in the *Explicit Semantic Analysis* table. For instance, with *MilneWitten*, some of the lowest  $\rho_s$  correspond to cases where there is a large mismatch in language edition size, with the highest  $\rho_s$  corresponding to the opposite situation. Similarly, the highest *MilneWitten*  $\rho$  of all language pairs belongs to English and French<sup>74</sup>, two of the largest and most highly-regarded language editions. Although this effect is weaker with semantic relatedness measures other than *MilneWitten*, here we have at least some evidence in support of the global consensus hypothesis as it appears in the SR literature. That is, some aspects of the tables above advocate for the prevailing assumption that the only reason there are differences between the language editions is that the smaller language editions have yet to catch up with the larger ones in terms of amount of content.

There are, however, several important points to consider with regard to this evidence. First and foremost, the striking cultural signal in the *Explicit Semantic Analysis* table and to a lesser extent in the *MilneWitten* table seriously problematizes the strong form of the SR literature's global consensus hypothesis, which implies that cultural context should have no effect on SR output. This strong form of the hypothesis is not a theoretical construct; it has been adopted in prominent several research projects (e.g. [74, 143]). We saw in the list above that there are, in fact, dozens of examples of culture completely trumping size and/or quality. For instance,

---

<sup>74</sup> This effect goes away when considering  $\tau_B$ . In this case, Czech and Slovak – two language editions whose corresponding language-defined cultures are quite similar – have the highest *MilneWitten* scores.

Russian and Ukrainian are quite asymmetric on both accounts, yet Russian is the most similar to Ukrainian of all other language editions for *every single* SR measure other than *WikiRelate*, including language editions whose size is much more comparable to that of Ukrainian. The same goes for Danish with regard to Norwegian and Swedish. Size- and quality-wise, Danish should have a maximum correlation with a language edition like Hebrew. Instead, however, it is it with Scandinavian language editions that Danish has the highest  $\rho$  for nearly all SR measures.

There are also important counter-examples of the global consensus hypothesis that are less obviously associated with culture. Most prominently, the German and French Wikipedias are the most highly-regarded non-English language editions, are the second- and third-largest language editions, and happen to be roughly the same size. However, never do the correlation coefficients for these language editions go above 0.548, which is that between  $MilneWitten_{GERMAN}$  and  $MilneWitten_{FRENCH}$ . The *ESA* coefficient between these language editions is only 0.516. There are also dozens and dozens of additional examples of similar phenomena. For instance, the smaller and moderate-sized language editions' *ESA*  $\rho$ s do not have significant Spearman's correlations with language edition size. The larger language editions' *ESA*  $\rho$ s do tend to have relatively large and significant correlations, however.

#### **6.3.4 Concept Pair-by-Concept Pair Analysis**

Another means by which we explored the role of cultural context in SR measure output was by examining the estimates generated during our experiment on a concept pair-by-concept pair basis. At this lower level, establishing the possible cultural causes of differing SR output can be difficult because many of the relationships considered by the SR measures are extracted from entire language editions rather than from single pages. However, some cultural signals are

readily apparent. For instance, for the concept pair (Germany, Fascism), the German Wikipedia is an significant outlier on the low side relative to all other language editions for *MilneWitten*. This is interesting, as we have observed along with some of our participants in the Omnipedia study (Section 7.3) that the German Wikipedia tends to cover topics related to Judaism exceedingly well. However, it could be the case of course that there is variable coverage of World War II-related topics.

To more easily investigate the possible lower-level cultural causes of variation in SR, we ran a small experiment similar to that described above, but doing the concept sampling in a pairwise fashion across the language editions. That is, for each pair of language editions and each SR measure considered, we sampled from the set of concepts that existed in both language editions. In addition, we sampled  $c_2$  from the set of outlinks or inlinks of  $c_1$  in order to ensure a more related distribution to more easily identify differences between the language editions. We used outlinks for SR measures that use inlink-oriented relationships and inlinks for the others.

This experiment revealed, at least in an anecdotal sense, that the effect of cultural context on SR measure output is strong. Consider the concept pair (Rabin cryptosystem, Feige–Fiat–Shamir identification scheme). Both concepts in this pair are cryptography technologies developed by Israeli scientists. *MilneWitten<sub>HEBREW</sub>* gives this pair a relatedness of 0.99, or almost perfectly related. The *MilneWitten<sub>ENGLISH</sub>* gives it a zero. A similar phenomenon occurs with the concept pair (Psagot Investment House, Apex Partners), the largest pension fund manager in Israel and a private equity firm with an office in Tel Aviv, and many other concept pairs in our *MilneWitten<sub>ENGLISH</sub>* / *MilneWitten<sub>HEBREW</sub>* sample.

Returning to the German Wikipedia, we found that the *MilneWitten<sub>GERMAN</sub>* SR for the concept pair (Michael Häupl, Falter) was 0.67 while for the English Wikipedia it was 0.0.

Exploring this example further, we saw that the German Wikipedia has a number of *MilneWitten*-identifiable relationships between Michael Häupl (the mayor of Vienna) and Falter (a Viennese weekly magazine) that are not in the English Wikipedia. For instance, Michael Häupl is the president of the Jewish Welcome Service Vienna, which publishes a magazine with Falter. The article “Jewish Welcome Service Vienna” (German) does not even exist in the English Wikipedia<sup>75</sup>. The article “Wien” (German) also mentions both of these concepts, whereas “Vienna” (English) only discusses Michael Häupl.

As a final example, let us consider the concept pair (The Fever (1999 film), Yurika Hino) with  $WAGDirect_{ENGLISH}$  and  $WAGDirect_{JAPANESE}$ . Yurika Hino is the voiceover actress for Teri Hatcher in Japan, a relationship that encoded on the Japanese Wikipedia’s articles about both Yurika Hino and the movie, in which Terri Hatcher stars. This relationship is on neither page in the English Wikipedia, resulting in  $WAGDirect_{JAPANESE}(\text{The Fever (1999 film), Yurika Hino}) = 0.5$ , but  $WAGDirect_{ENGLISH}(\text{The Fever (1999 film), Yurika Hino}) = 0.0$ .

### 6.3.5 Other SR Distributions

As noted above, our primary SR experiment uses an application-focused SR distribution in which the bulk of concept pairs result in an SR estimate of zero, regardless of semantic relatedness measure. To understand whether our results are robust against changes in this distribution, we compared our output to that of Hassan and Mihalcea [74], who as a side result of an otherwise global consensus hypothesis-focused research project, calculated the pairwise *ESA* relatedness for translated versions of the *WordSim353* human gold standard dataset. The average SR estimates of *WordSim353* annotators have a roughly normal distribution, which is substantially different than the SR distribution we have used thus far. Hassan and Mihalcea

---

<sup>75</sup> This is even more surprising given that the name of the service is in English.

considered four language editions: English, Arabic, Spanish, and Romanian. Despite the divergent nature of the two SR distributions, our results for the pairs of these languages supported by our work are very similar to those reported by Hassan and Mihalcea. Remarkably, the two English/Spanish  $\rho$ s are *exactly* the same (0.55). Moreover, the English/Romanian results are only 0.04 apart (0.38 vs. 0.42). There was a larger distance between the Romanian/Spanish, however. Hassan and Mihalcea report that the  $\rho$  between  $ESA_{SPANISH}$  and  $ESA_{ROMANIAN}$  was 0.30, while we found it to be 0.45. Regardless, the fact that two of the language pairs were so similar despite the different distributions, different ESA implementations, and *entirely different set of concept pairs* suggests that our high-level results are robust against SR distribution. That said, the Spanish/Romanian result will need further study.

## 6.4 Discussion

In this chapter, we have seen that the diversity between the language editions of Wikipedia causes the output of semantic relatedness measures to vary substantially when using different language editions as their source of world knowledge. We also demonstrated that the cultural context embedded in this diversity is behind at least some of this variation. While these result have the high-level implication that UGC-based technologies can adopt the cultural viewpoints of their world knowledge they also have, as noted above, a more specific and more troubling implication. Namely, our results suggest that the large number of English Wikipedia-based technologies adopt the viewpoint of a single cultural group's view of world knowledge, specifically that of the English-speaking language-defined community. This means that these technologies adopt the perspective, for instance, that Germany and Fascism are quite related, something that is less true of a technology that uses the German Wikipedia as world knowledge.

Conversely, these technologies will be unaware of the relationships related to, for instance, Jewish civic institutions in Austria, of which they would be aware if they used the German language edition.

While semantic relatedness measures are a particularly important UGC-based technology in artificial intelligence and related fields, the extent to which we can generalize our findings to all UGC-based technologies is limited given the massive number of these technologies and their heterogeneous methods and applications. In addition, we do not know the extent to which our results apply to the cultural contextualization of user-generated content when it comes to non-language-defined cultural groups, for instance geographically-defined cultures. However, the findings in this chapter suggest that designers of UGC-based technologies need to at least be aware of this issue and take caution when applying the knowledge in UGC derived from one cultural group in a technology with a diverse base of users.

# 7 Omnipedia

*Note: This work originally appeared in the Proceedings of the 30th ACM Conference on Human Factors in Computing Systems (CHI 2012) [9]. While much of the text here is original to this thesis, portions have been adapted from the original publication, of which my colleague Patti Bao and I were primary co-authors.*

In the previous chapter, we discussed some of the risks associated with ignoring the cultural context in user-generated content. In this chapter and the one that follows, we explore the *benefits* of culturally contextualized UGC. Namely, we demonstrate through two novel UGC-based applications that by embracing the cultural information embedded in UGC instead of ignoring it, a whole new class of UGC-based technologies is made possible. This chapter is dedicated to one of these new technologies, an application we built called Omnipedia.

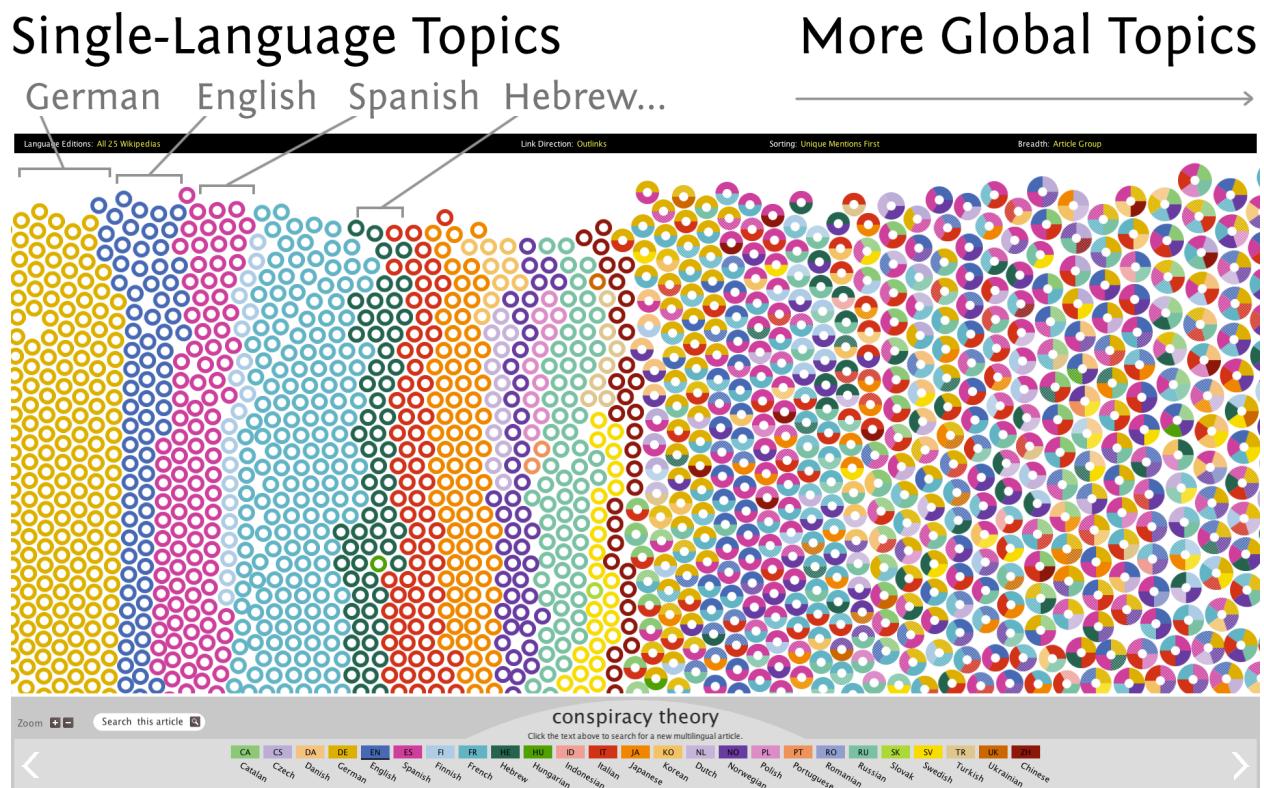
## 7.1 Introduction

In Chapter 3, we demonstrated from many angles that each language edition of Wikipedia is different from the others, and that some of these differences reflect diversity in the underlying world views of the corresponding language-defined cultures. In Chapter 6, we discussed the implications of encyclopedic world knowledge diversity on Wikipedia-based technologies. However, we have yet to seriously address the implications for Wikipedia *readers* in an in-depth fashion.

Put simply, the language-induced splintering of information in Wikipedia poses both an opportunity and a challenge for the 484 million monthly visitors to Wikipedia [210]. On the one hand, as we saw in great detail in Chapter 3, Wikipedia embodies an unprecedented repository of world knowledge diversity in which each language edition contains its own cultural viewpoints on a large number of topics. On the other hand, the language barrier serves to silo knowledge,

preventing the average Wikipedia reader from accessing most of the information on the site.

Omnipedia is a system that attempts to remedy this situation at a large scale. It reduces the silo effect by providing users with structured access in their native language to all 8.67 million concepts in our 25-language edition database. At the same time, it highlights similarities and differences between each of the language editions, allowing users to see the diversity of the represented knowledge. To achieve this goal, Omnipedia extracts the topics discussed in each language edition's coverage of a given concept using the BOL-based and wikification approaches identified in Section 3.5, then loads them into an interactive visualization that shows which language editions mention which topics and how those topics are discussed.



*Figure 7.1-a: A screenshot of Omnipedia visualizing the multilingual article “Conspiracy theory” in zoomed-out mode. “Multilingual article” is a user-friendly term for the idea of a concept as introduced in Chapter 3. Small circles on the left indicate topics that are discussed in only a single language edition’s coverage of the concept. Bigger circles on the right indicate topics that are discussed in multiple language editions’ coverage of “Conspiracy theory”.*

Consider, for example, the English Wikipedia article “Conspiracy theory”. This article discusses many topics, from “Moon landing” to “Kennedy assassination.” However, many other language editions also contain articles on this concept, such as “Verschwörungstheorie” (German) and “Teoría conspirativa” (Spanish). In fact, conspiracy theory is a global concept as defined in Section 3.4. Omnipedia visualizes this global concept as a single “multilingual article” (Figure 7.1-a). The small circles on the left of Figure 7.1-a represent topics discussed in only one language edition: yellow for German, dark blue for English, and so on. The left side of Figure 7.1-a helps users understand what we saw again and again in Chapter 3: reading just a single language edition—even English—means missing out on large amounts of content available in other language editions (at least 29% on average, see Section 3.5). Moving toward the right half of Figure 7.1-a, one begins to see larger, multi-colored circles that represent topics that are discussed in multiple language editions’ coverage of the conspiracy theory concept.

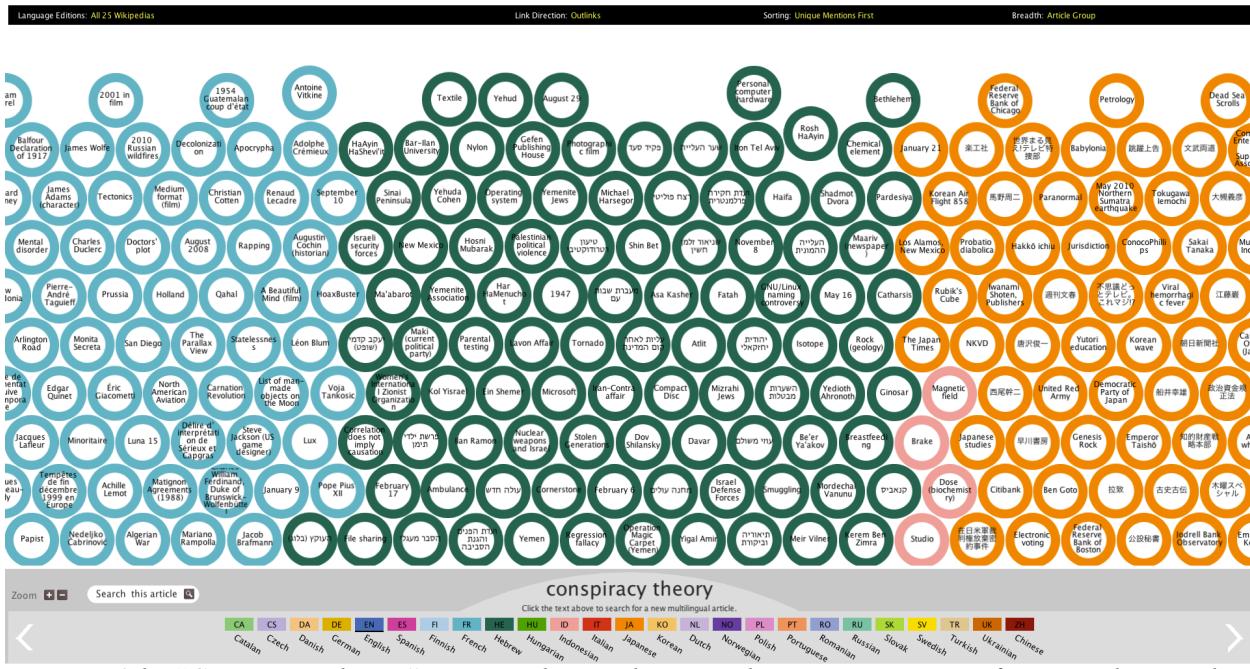


Figure 7.1-b: “Conspiracy theory” in zoomed-in mode. Here, the user can see specific topics discussed in each language edition’s article. Because the user has panned over to the single language linked topics, s/he can see that the Hebrew Wikipedia (dark green) discusses “Palestinian political violence” while the French Wikipedia (cyan) discusses “Algerian War.” Clicking on one of the circles calls up a snippet (Figure 7.1-c) from the corresponding Wikipedia article(s) that explains the discussion of each topic in detail.

Zooming in (Figure 7.1-b) allows users to explore content in more detail. For instance,

Figure 7.1-b shows that the Hebrew Wikipedia (dark green) has a great deal of exclusive content about Israel-related conspiracy theories. The French Wikipedia (cyan) also has unique content, both pertaining to French history, as indicated by “Algerian War,” and of more general interest, such as “Pope Pius XII.”

Panning right, users begin to find topics that are discussed in more than one language edition. Figure 7.1-c shows the most commonly discussed topics in the “Conspiracy theory” multilingual article, which include “Freemasonry,” “United States,” and “Central Intelligence Agency.” We also see that Judaism is discussed in many language editions’ coverage of conspiracy theories, demonstrating that this form of anti-Semitism is unfortunately widespread. To discover precisely how these topics are discussed in various language editions, users can click on a topic circle. This returns a snippet mined from each language edition that puts the topic in context, with snippets translated into a user-defined language using machine translation (Figure 7.1-c). Omnipedia’s snippet selection approach leverages the algorithms developed for Atlasify’s explanations, which are discussed in detail in Chapter 8.

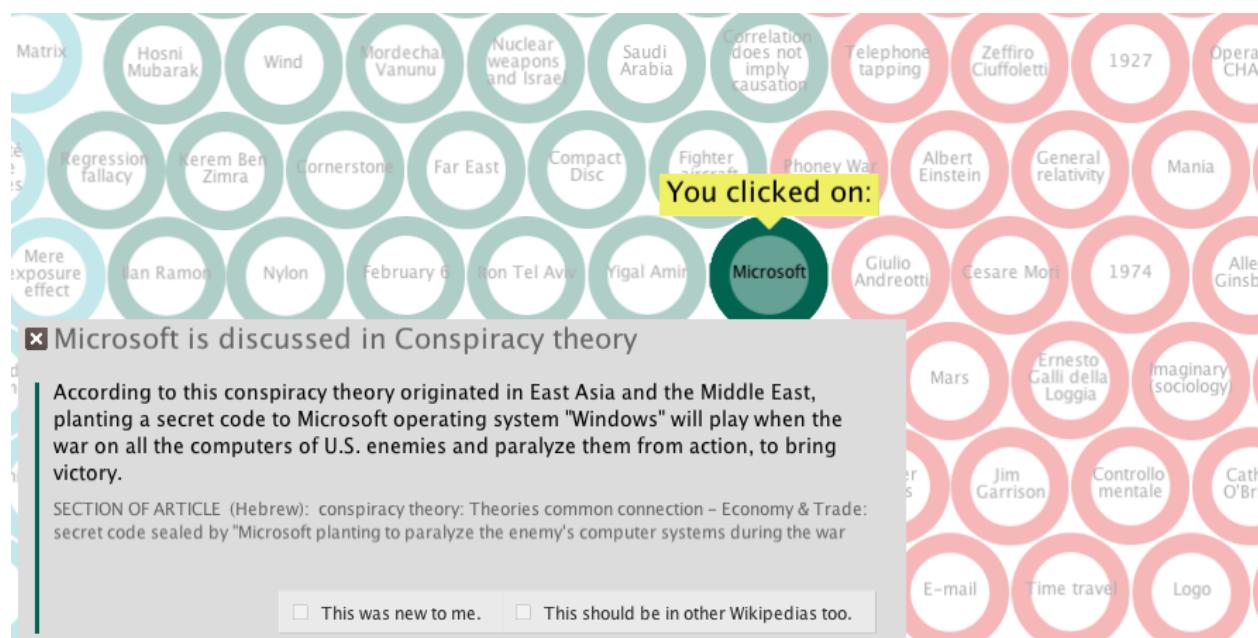


Figure 7.1-c: The snippet explaining how Microsoft is discussed in the Hebrew Wikipedia article on conspiracy theories. This is the only part of Omnipedia that relies on live machine translation.

By now it should be clear that Omnipedia is a way of providing simple concept-by-concept access for the average computer user to many of the repository-level findings from Chapter 3. As noted above, a multilingual article is nothing more than a user-friendly term for a concept. Moreover, each Omnipedia visualization is nothing more than a depiction of the sub-concept-level diversity (Section 3.5) of a given concept. While we use the term “topic” to describe the content in Omnipedia’s visualizations, a “topic” is simply a more accessible way of describing a link (missing or not) under the bag-of-links assumption (Section 3.5.1.2). In fact, Omnipedia leverages the exact same wikification algorithms we developed for our sub-concept-level diversity studies. Omnipedia currently uses the *<WikipediaTitle, GoogleTranslateNone>* wikification strategy, but we are working to incorporate the Google Translation information as well. Many other approaches and findings from Chapter 3 (and portions of Chapters 6 and 8) are utilized in Omnipedia, the most important of which are summarized in Table 7.1-a. As the chapter proceeds, we will highlight in more detail how our lower-level “back-end” Wikipedia-

<b>Omnipedia Feature / Construction</b>	<b>Lower-level Contribution</b>	<b>Main Section</b>
Multilingual articles	Equivalent to concepts identified via our <i>Conceptualign</i> algorithm.	Section 3.3
“Mentioned topics” in multilingual articles	Extracted using our BOL-based approach and multilingual wikification techniques.	Sections 3.5.1.1 and 3.5.1.2
Inclusion of sub-articles in multilingual articles	Enabled by our sub-article identification algorithm.	Section 3.5.1.3
Support for gloss-only visualizations	Uses WikAPIdia’s link location attributes.	Section 3.2.2
Highlighting of related topics	Leverages the <i>MilneWitten</i> semantic relatedness algorithm.	Section 6.1
Mining of snippets relevant to clicked topics	Uses the explanation generation engine from Atlasify.	Section 8.3.1

Table 7.1-a: A selection of Omnipedia features and our lower-level contributions that made them possible.

related contributions are given a “face” in Omnipedia.

## 7.2 The Omnipedia System

Users typically begin their interaction with Omnipedia by typing in a concept of interest, for instance “Conspiracy theory.” Omnipedia will then look up the corresponding multilingual article and display the types of visualizations seen in Figure 7.1-a - 7.1-c using circles of different sizes and colors to indicate the topics that are discussed in various language editions’ coverage of the entered concept. Each circle denotes a topic that is mentioned in at least one language edition of Wikipedia. As described in Section 3.5, topics are concepts/multilingual articles themselves, and by double-clicking on a circle, the user can browse through related topics, just as they can follow hyperlinks in the normal version of each language edition.

A central design premise for Omnipedia is that it be language neutral. As such, users are able to switch the interface language to any of the 25 supported languages. If a user switches the

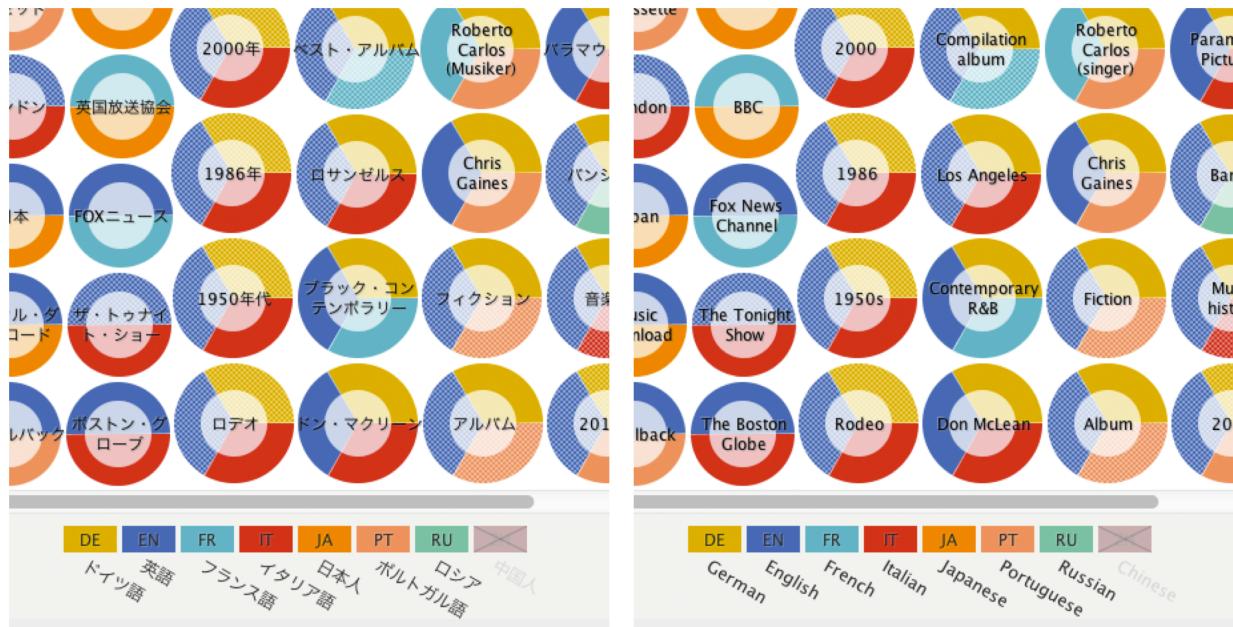


Figure 7.2-a: The multilingual article “Johnny Cash” with the interface language set to Japanese (left) and English (right). Titles that do not appear in Japanese on the left represent concepts that do not have a Japanese article.

interface language to Japanese, for example, she is able to look up multilingual articles by their Japanese titles and sees topic titles in Japanese (Figure 3.10-a). Because this process relies on the article metadata and redirects resources in Wikipedia (Section 3.2), it involves no machine translation—an essential requirement for the system as machine translation on this scale would be too slow and, in the common case where thousands of topic titles are visualized, would excessively tax common machine translation APIs. When a topic that is discussed does not have a corresponding article in the interface language, we use a back-off strategy that, for instance, displays a single-language linked topic title in its native language (e.g. Chris Gaines in Figure 3.10-a).

Omnipedia is also language neutral in that it allows users to include or exclude any of the 25 supported languages, creating custom language sets at will. Omnipedia makes several built-in language sets available to the user, including “Languages of the Top Ten Economies,” “NATO Languages,” “EU Languages,” etc. Once a language set is changed, all visualization and algorithmic systems are updated accordingly. For instance, in Figure 3.10-a, the user has selected the top ten economies language set.

When designing Omnipedia, we opted for a visualization strategy over a text-based approach largely because text alignment across just two language editions is known to be a difficult problem [1, 7, 146], let alone text alignment across 25 languages. In doing so we lose certain advantages of text, like the grouping of related topics into a single, cohesive discussion. We mitigate this situation by using the *MilneWitten* semantic relatedness algorithm [136, 137], which is discussed in detail in Chapter 6. When a user clicks on a topic circle to view an explanation snippet, Omnipedia flags highly related topics with a “See Also” tag. For instance, when browsing the multilingual article “Barbecue,” a user might be curious as to why “Kashrut”

is discussed in the corresponding German article. When she clicks on the “Kashrut” circle, the topics “Israeli cuisine,” “Modern Hebrew,” and “Pastrami” are highlighted to guide further exploration. As it turns out, there is an Orthodox Jewish barbecue festival in Memphis every year, a fact only discussed in the German Wikipedia.

Omnipedia allows users to adjust a breadth setting that determines what is treated as an “article.” In doing so, it takes full advantage of the various Wikipedia Article Graph edge properties and sub-article (Section 3.5.1.3) functionality made available in WikAPIdia, on which Omnipedia is built. At the broadest breadth setting, Omnipedia visualizes topics from the parent articles and all sub-articles that exist in the user-determined language set. This breadth setting is called an “Article Group” in Omnipedia. The medium breadth setting, labeled simply “Article,” shows only what is discussed in the main articles. Finally, the narrowest breadth setting considers topics mentioned in only the first section of each article, thus showing diversity in how concepts are summarized. We are also incorporating the even narrower setting that only looks at topics in the first paragraph, a trivial task thanks to WikAPIdia’s identification of location properties for each link.

Before moving on, it is important to note that WikAPIdia currently only visualizes parseable links. Including unparseable links would involve changing less than one line of code. However, it was our determination that unparseable links add too much computational<sup>76</sup> and visual complexity for their information value.

---

<sup>76</sup> Omnipedia’s approach to visualization, which relies on a physics engine, is somewhat computationally expensive.

### 7.3 Study

*Note: The large majority of the work in this section was completed by my co-author Patti Bao. Although I include information about our study for completeness, it should be considered research in which I only played a supporting role.*

Omnipedia provides an excellent opportunity to examine how people interact with the diverse information in multilingual Wikipedia and culturally contextualized user-generated content in general. To gain a better understanding of this interaction, we conducted an exploratory study that involved putting Omnipedia in front of a number of real users. Twenty-seven people (14 female, 13 male, ranging from 18-62 years old) participated in the study, all of whom had accessed Wikipedia at least once in the past 30 days. Participants consisted of 20 native English speakers, four native Chinese (Mandarin) speakers, one native Russian speaker, and two native speakers of English, and one other language not supported by Omnipedia. Sixteen users were fluent or proficient in at least one other language besides their native language. These additional languages were Spanish (10 users), English (5), French (1), Japanese, (1) Korean (1), and Telugu (1, not supported by Omnipedia). On average, participants had used the English Wikipedia for six years (self-reported, SD = 1.84). Ten participants had previously seen a non-English Wikipedia, but only three considered themselves frequent users.

Participants arrived at the laboratory and were taken to a private room with a desktop machine and 23" display. The experimenter then provided a 10-minute demonstration of Omnipedia's main features and afterwards proceeded to a separate observation room. Participants were given 30 minutes to freely explore multilingual articles in Omnipedia using any of the languages they wished. Afterwards, the experimenter returned to the room to lead a structured interview asking participants to reflect on their experience. Participants were

prompted with specific instances drawn from their interaction logs. We based our analysis on 58,900 words of transcribed interviews, 41,704 logged events (including mouse hovers, clicks, queries, and changes to view settings), and 30 pages of observation notes. We then used this data to characterize user exploration of multilingual Wikipedia through Omnipedia and describe some of the insights they shared with us.

### **7.3.1 Results**

Just over half (14) of our users had never seen a non-English Wikipedia prior to the study. When using Omnipedia, all users took the opportunity to access information from a non-English Wikipedia, the five most popular being French, Italian, Russian, German, and Spanish. Twenty-two users switched to one of Omnipedia's built-in language sets at least once during the study ( $M = 3.1$  switches,  $SD = 3.49$ ). Twelve of these users also created a custom language set. Users tailored these sets based on their own language proficiency, relevance to the concept of interest, or curiosity about a never-before seen Wikipedia. On average, users looked up 15 multilingual articles ( $SD = 6.25$ ). They clicked on 60 topics on average ( $SD = 34.7$ ) to load the type of snippets seen in Figure 7.1-c. Of all the topics users clicked on, 26.7% had been highlighted as related topics by the semantic relatedness algorithm.

### **7.3.2 Exploring Similarities and Differences**

After seeing the diversity of linked topics among language editions, many users concentrated on the most common topics (or biggest circles), typically citing reasons involving the perceived importance of these topics (e.g. "if it was in all four languages, it must be important" (P7)). Viewing these topics often required users to pan all the way to the opposite end of the visualization (as seen in Figure 7.1-c), which they took the effort to do in order to gain

insight into what was “well known” worldwide (P18). Many of these users were satisfied with clicking on these topics and reading just one of its multiple snippets. However, other users read all of the snippets from more globally-discussed topics in detail to see if there were “cultural nuances” (P23). Users who engaged in this type of behavior realized that just because multiple language editions shared a link to a topic did not necessarily mean that they agreed on how the topic was related. For example, P4 recalled looking up the multilingual article “Boeing 767” and discovering that the English snippet on a plane crash in Egypt included “different perspectives on what happened” while other language editions “just summarized one sentence.” Similarly, P12 expected that German coverage of “Siemens” would gloss over the company’s support of the Nazi movement during World War II. He was surprised to find that the German Wikipedia’s snippet was “the most descriptive about that fact.” Other users spent little time investigating the most common topics, regarding them as “pretty obvious” (P25) or “basic things” (P10) that would not yield the most interesting insights. Instead, they searched for differences in topic coverage by looking at single-language topics (or smallest circles) across language editions. One approach was to examine relative proportions of single-language topics in a given multilingual article. For instance, P17 inferred that American basketball player Dwight Howard was “definitely more famous in the English version than in any other language” based on the considerable number of topics discussed only in the English Wikipedia, a finding that echoes our work on topic-defined diversity in Section 3.7. Likewise, P24 was not surprised to find that a minor tennis player only had coverage in English while “Rafael Nadal had more single-language links from Spanish and other languages because he’s a worldwide figure.” She, like others, interpreted these differences as a measure of the concept’s (e.g. Rafael Nadal’s) global “reach” or “impact.” Users who took this approach also discovered distributions of single-language topics

that belied their expectations. P10 was surprised to find that “even Italian and Spanish had something to say” about Jainism, an Indian religion. P6 compared several music genres and was not surprised to find that hip-hop had “more in English” but was surprised to discover that reggae had “a lot in Japanese”.

Another approach for finding differences among language editions involved targeting concepts that might be more likely to reveal differences in perspective. A subset of users actively sought out what they considered to be “globally polarized” (P20) or “heavily charged” (P23) concepts like “Climate skepticism” and “War on Terror.” They intentionally included language editions that they thought would reveal “different sides” (e.g. P13, who looked up “The Holocaust” in German, Hebrew, and Polish). In most cases, however, users did not find the extreme differences they anticipated, leading them to reconsider their own expectations regarding the cultural contextualization of the language editions of Wikipedia.

### **7.3.3 Discovering New Knowledge**

Users actively sought out knowledge not available in their own language editions, effectively advocating for the benefit of surfacing user-generated content that is contextualized for a variety of different cultures. For example, the 11 “monolingual” users who were fluent in English but had no more than rudimentary knowledge of another language clicked on topics that, on average, were mentioned in 2.79 language editions ( $SD = 2.06$ ). Thirty-six percent of all topics clicked by these users were not discussed at all in English. Users often reported clicking on topics discussed in one language because they might have “interesting facts that I hadn’t heard of” (P26). For certain multilingual articles, users paid attention to unique topics in a single language edition where they expected a close tie to the language’s culture hearth. For example,

P6 “focused on the Italian side [of the visualization] just because Sardinia’s in Italy.” He also looked at Chinese-only topics discussed in “Google” because he thought they might reference Google’s search restrictions in China. Similarly, P16 thought “the Japanese-only information will be more authentic since Ayumi Hamasaki is from Japan.” Conversely, one user decided to exclude a language (Chinese) from the interface because it “wasn’t giving me much” in terms of unique information.

In other cases, users investigated single-language topics from many language editions. For instance, P10 wanted to see “if maybe one culture viewed a certain aspect of ‘Beauty’ that [she] didn’t know.” After discovering a number of Japanese-only topics that seemed to emphasize “character,” she went on to examine English-only topics and observed that they discussed “beauty in the eye of the beholder” as well as “physical” attributes.

Finally, it is worth noting that the sheer amount of single-language topics was a revelation to the majority of users. Reflecting on their use of Omnipedia, a few users who initially focused on the more global topics wished they had more time to explore the single-language topics, as those may have yielded different insights. P1 even told us in hindsight, “If I had bothered to take my time and go through all the single [language] ones, I think I would have learned more about what the differences were.”

### **7.3.4 Study Summary**

In sum, four key insights emerged from users’ interactions with Omnipedia. First, users took advantage of the fact that using the lens into culturally contextualized Wikipedia provided by Omnipedia, they could identify the most commonly and globally discussed aspects of a concept’s Wikipedia definition. Second, they were able to discover both similarities and differences in how

these topics were discussed among language editions. Third, access to single-language topics allowed users to not only filter interesting topics based on inferences of self-focus bias (Section 3.10), but also get a big picture view of how much topics were being discussed in different language editions. Finally, users began to comprehend the magnitude of information that was not available to them in the English Wikipedia.

## 7.4 Visualization Approach

While Omnipedia represents the first system to our knowledge that allows users to interact with a large number of Wikipedia language editions simultaneously, doing so required addressing many challenges. We discuss our solutions to the myriad “back-end” challenges involved with building Omnipedia in Chapter 3. Here, we briefly touch on those that occur in the “front end.” The most significant such challenge is at the very core of Omnipedia: the visualization of ego-centric multigraphs with the egos (i.e. the multilingual articles) and any other vertex (i.e. a topic) being connected by as many as 25 edges. In the early stages of the Omnipedia project, we experimented with many different visualization approaches for this particular data structure. The approach used in the current version of Omnipedia casts the visualization of this complex network into the visualization of discrete, overlapping sets of categories with each category representing the topics mentioned in a given language edition.

The visualization of discrete categories is of course a well-understood problem and has many successful applications. However, Omnipedia occasionally deviates from the best practices for this type of visualization. Primarily, it is generally not advisable to attempt to depict more than 5-9 discrete categories at once, as doing so will overload users’ ability to successfully interpret the categories [123]. Omnipedia, on the other hand, allows users to visualize up to 25

categories at once.

Omnipedia also takes a non-standard approach when it comes to color selection. Typically, it is advisable when visualizing discrete categories to ensure that categories displayed with similar hues are similar in some way [16]. While we initially considered grouping languages in the same language family, we decided that doing so would ignore similarities and differences in the corresponding language-defined cultures and abandoned the idea. For instance, French and German speakers share more than their language families would suggest. The same goes for Hungarian and Czech speakers, and so on. Similarly, it is generally recommended that the “lightest, darkest, and most saturated hues” in a color scheme be assigned to “categories that warrant emphasis” [16], which is again a recommendation we did not consider. Of course, in Omnipedia we seek to not emphasize or favor any particular language edition.

The decision to violate these perceptual and color theory guidelines, however, was intentional and, for two reasons, justified given the goals of Omnipedia. First, from a static visualization perspective, our only goal is that Omnipedia users gain a high-level understanding of the extent to which the data they are examining is from different language-defined communities. Because we chose to make size, color, and position redundant in certain respects, users are able to “at a glance” get a general idea of the relative numbers of single-language entities, globally-mentioned entities, and those that are not in either of these two groups. Similarly, users are also able to easily see that the content comes from a variety of language editions rather than just English. For example, consider Figure 7.1-a. It is immediately clear that there is a significant number of single-language topics in the “Conspiracy theory” multilingual article, and that these single-language concepts come from a number of different language editions. Given the extent to which the English-as-Superset hypothesis is assumed even in the

research community, making this type of point to end users is a key goal of Omnipedia.

Second, Omnipedia allows for many different types of interactions with the visualized data, and in our interaction choices, we carefully follow best practices such as those suggested by Card et al. [20], Shneiderman [184], and Yi et al. [216]. These interactions are in many cases designed to address the issues introduced by our decision to allow users to visualize 25 different discrete sets simultaneously. For instance, users are easily able to zoom in on specific mentioned topics, they can filter these topics by language edition, and by clicking on a topic's circle, they can receive details on demand about the context of how the topic is mentioned. The ability to restrict the number of language editions is a particularly important point. When a user selects nine or fewer language editions, Omnipedia immediately conforms to the best practices for visualization of discrete categories as the number of colors has been reduced to a manageable quantity. Indeed, in our study we often found that, after exploring a multilingual article at a high level with 25 language editions, users often reduced the number of visualized languages greatly in order to investigate more detailed hypotheses about the multilingual article.

With regard to our color choices, we made the judgement call that having consistent colors for each language edition was more important than following best practices for each individual visualization. While this results in non-optimal color selection in certain cases, we believe the benefits of consistency outweigh any small amount of additional perceptual effort required by users for some visualizations.

Of course, Omnipedia's visualization choices do involve a number of tradeoffs. For instance, it is difficult in Omnipedia to distinguish between topics that are mentioned in, say four language editions, from those that are mentioned in, say, six language editions. While position helps here, Omnipedia is optimized for identifying large differences in the mentioning of topics.

That said, we have recently implemented a “grouping” functionality that allows users to cluster topics based on any variable, although the only variable we support at the moment is the number of language editions in which topics are mentioned.

It is also difficult in Omnipedia to ask some of the questions we did in Section 3.5. For instance, in the current version of Omnipedia it is not possible to show all topics mentioned in a given set of language editions that are *not* mentioned in English. We are developing a feature, however, that will allow users to do this by dimming the English portion of topic circles (or that of any other language edition).

Before concluding, it is important to note that these visualization challenges are not merely idiosyncratic problems relevant only to Omnipedia. Any future applications that allow users to interact with culturally-faceted user-generated content will likely encounter similar issues, especially if they take a pure visualization approach. As we have seen in this thesis, user-generated content is contextualized by many different types of cultures, and by many different instances of those types. As such, finding a way to successfully communicate the complex, overlapping nature of the cultural context embedded in UGC will often involve developing effective visualizations of a large number of overlapping discrete (or even fuzzy) sets. With Omnipedia, we have taken steps towards demonstrating how this can be done. However, it is likely that we will be able to improve on our approach in future work.

## 7.5 Future Work and Conclusion

We have taken very initial steps towards using Omnipedia to visualize other culturally contextualized repositories of user-generated content. Sites such as Twitter and Flickr suffer from the same language barriers as Wikipedia and have also been shown to display important

differences across languages [33, 95]. Future work might treat a Twitter hashtag as an “article” and mine tweets posted in many languages that contain the hashtag for discussed topics. Similarly, a group of related photos (e.g. of the same event) could be used as the “article” and the photos’ tags could be considered topics.

Another area of future work – and one that is more pressing – is doing the engineering labor necessary to release Omnipedia to the wider public. The many advancements in WikAPIdia 0.3 discussed in Section 3.12 have made great strides in this direction, but additional work is needed specifically with regard to speeding up wikification. Another critical concern is obtaining a donation of machine translation API calls. Without such a donation, supporting the click-to-get-context functionality of Omnipedia would be prohibitively expensive.

## 8 Atlasify

*Note: Much of this work originally appeared in the Proceedings of SIGIR 2012 [78]. However, the detailed description of how cultural context is leveraged in Atlasify is new to this thesis.*

While Omnipedia is one application enabled by the cultural context in UGC, its entire purpose is quite closely tied to this cultural context. That is, the main contribution of Omnipedia is providing users with the ability to visualize the cultural context Wikipedia-based UGC. In this chapter, we show how the diversity in UGC can enrich a technology whose primary contributions lie in other areas. Specifically, we introduce Atlasify, a system we built that, through innovations in thematic cartography and natural language processing, enables a novel approach in an area known as exploratory search. The chapter begins with a description of the exploratory search approach and its contributions and ends with a demonstration of how this approach can be enhanced by leveraging the cultural context in user-generated content.

Exploratory search is usually defined as an open-ended information seeking activity in which a user aims to better understand a complex concept [205, 206]. While exploratory search has historically accounted for roughly a quarter of Web search query volume [172], it remains challenging using today's search engines due to their focus on closed information requests and navigational queries [206].

Atlasify supports an entirely new exploratory search approach that leverages thematic cartography's well-known ability to communicate complex geographic distributions [12, 20, 32, 187] to help users understand the complex concepts encountered in exploratory search. While the benefits of cartography are usually limited to geographic inquiries, our approach is made domain-neutral by harnessing general relational knowledge mined from Wikipedia. This means

that users can employ thematic cartography to explore concepts not only from a geographic perspective, but also from a chemistry perspective, a politics perspective, a music perspective, or a perspective from any other topic area (even user-defined topic areas).

Given a query concept, Atlasify automatically generates an interactive thematic cartography layer (e.g. a choropleth or heat map) on top of a spatial reference system from any domain, such as a periodic table, a U.S. senate seating chart, or a world map. The layer illustrates the degree to which the query concept is related to each spatial concept in the reference system (e.g. chemical elements, senators, countries). By clicking on a spatial concept, users see natural language explanations of exactly how that concept or region is related to the query concept. Users can enter any query that corresponds to a Wikipedia article (i.e. a page title, anchor text, or redirect; see Section 3.2.2).

To make this process more concrete, consider the Atlasify use case in Figures 8-a through 8-d. In Figure 8-a, a user (e.g. an intelligence analyst) who wants to learn about nuclear weapons has queried Atlasify for “Nuclear weapon” and selected “Periodic Table” as the desired spatial reference system. As is typical with choropleth maps, the dark green areas in Figure 8-a are very related to nuclear weapons, and the lighter green areas are less related. Exploring further, the user may wish to understand why, for example, cobalt is related to nuclear weapons. By clicking on cobalt in the visualization, the user is presented with natural language explanations of the relationships between nuclear weapons and cobalt. Seeking a geographic perspective on nuclear weapons, the user then changes to the “World Map” reference system (Figure 8-b). The user does the same for a temporal perspective in Figure 8-c (the “Timeline” reference system) and a United States politics perspective in Figure 8-d (the “U.S. Senate Seating Chart” reference system). Note that Atlasify correctly highlights the United States, Russia, and Iran in the world map, the

various important eras in the history of nuclear weapons on the timeline, and so on. Atlasify currently supports 13 reference systems in total and adding new reference systems is straightforward.

The effectiveness of thematic cartography is well established in geographic domains (e.g. [12, 20, 32, 187]). The goal of our exploratory search approach is to extend the strengths of thematic cartography to the wide variety of domains and query concepts encountered in exploratory search. This goal can be broken down into three key challenges, the solutions to which are additional contributions of this chapter and have implications outside of exploratory search.

The first challenge involves generalizing the visualization strategy used in Figure 8-b to non-geographic reference systems (e.g. periodic tables, anatomical charts, timelines, and many other figures and diagrams). Our solution is *explicit spatialization* (ES), which enables cartographic and geographic information retrieval (GIR) methods to be applied in any figure or diagram. As discussed in Section 8.2, ES accomplishes this by “spatializing” concepts into pre-defined reference systems and generalizing the canonical model of geographic information to incorporate domain-neutral spatial information. In doing so, ES can extend the ongoing advances in online mapping and GIR to many domains outside of geography.

The second challenge involves automatically estimating the degree of relatedness between any of the millions of possible query concepts (e.g. nuclear weapons) and every spatial concept in each reference system (e.g. chemical elements, countries). These estimates determine the value of the visual variables manipulated in thematic cartography, such as color and text size (e.g. the shades of green and font sizes in Figures 8-a through 8-d). We show how the Wikipedia-based semantic relatedness (SR) measures we discussed in Chapter 6 can solve this problem. We

also introduce a new SR measure, *AtlasifySR+E*, which uses a learned model to combine six separate SR measures, thus capturing all of the types of relationships understood by each individual SR measure. Experiments on several SR benchmarks show that *AtlasifySR+E* achieves state-of-the-art performance while also remaining language-neutral and using only open, easily accessible data, overcoming two limitations of the current state-of-the-art SR measure.

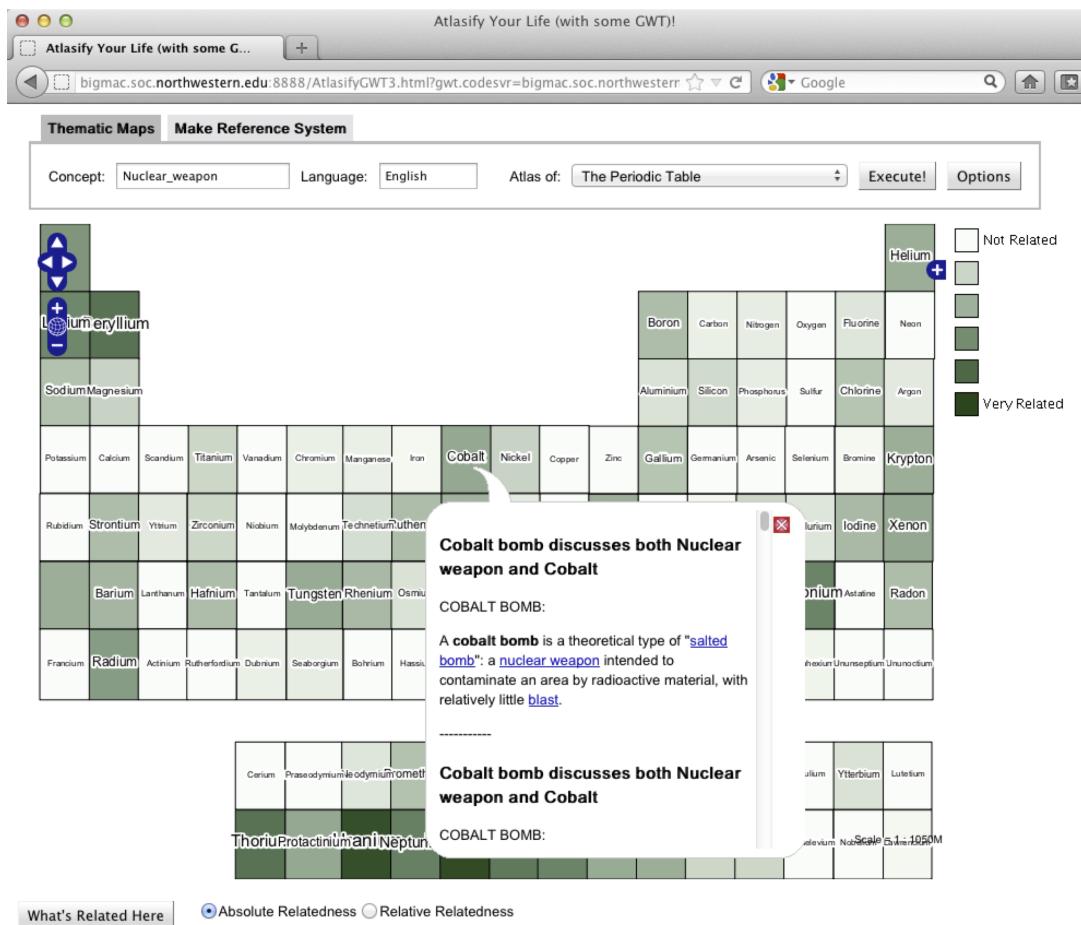


Figure 8-a: Atlasify's visualization of the query “Nuclear weapon” on the “Periodic Table” spatial reference system. If users click on cobalt, they receive a list of explanations of how nuclear weapons and cobalt are related.



*Figure 8-b: Atlasify visualizing the query “Nuclear weapon” on the “World Map” reference system. The user is able to see that, for instance, sub-Saharan Africa is not very related to nuclear weapons, while the United States, Russia, and North Korea are quite related. The “World Map” reference system is the largest of Atlasify’s spatial reference systems. For each query concept, the AtlasifySR+E semantic relatedness between all entities and the query concept must be calculated.*

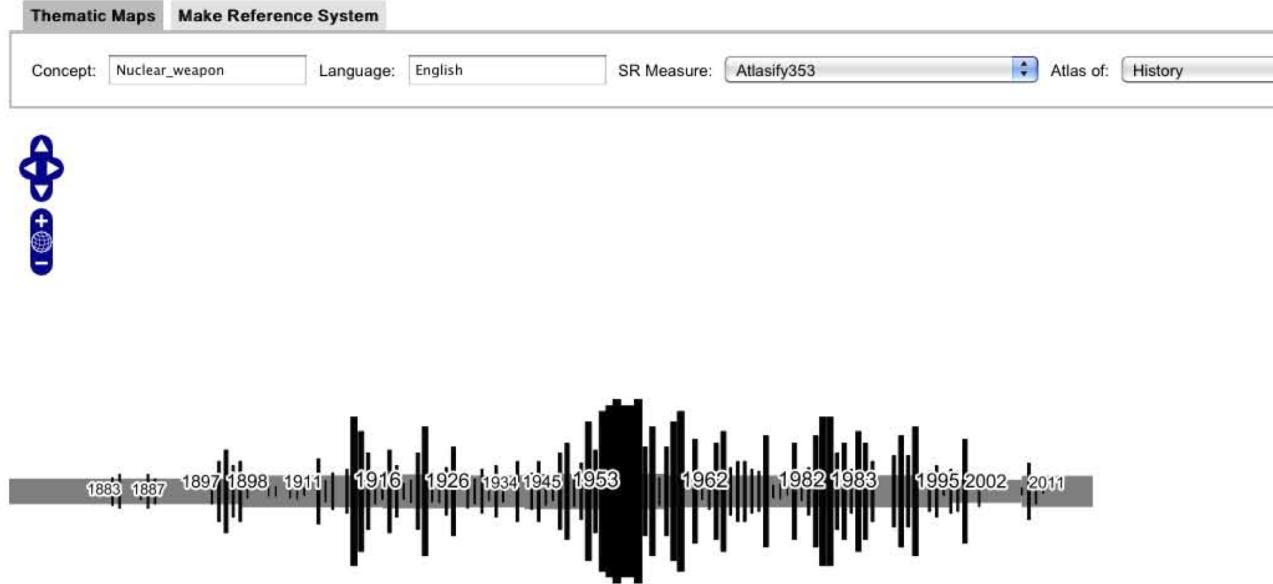


Figure 8-c: “Nuclear weapon” visualized on the “Timeline” reference system.

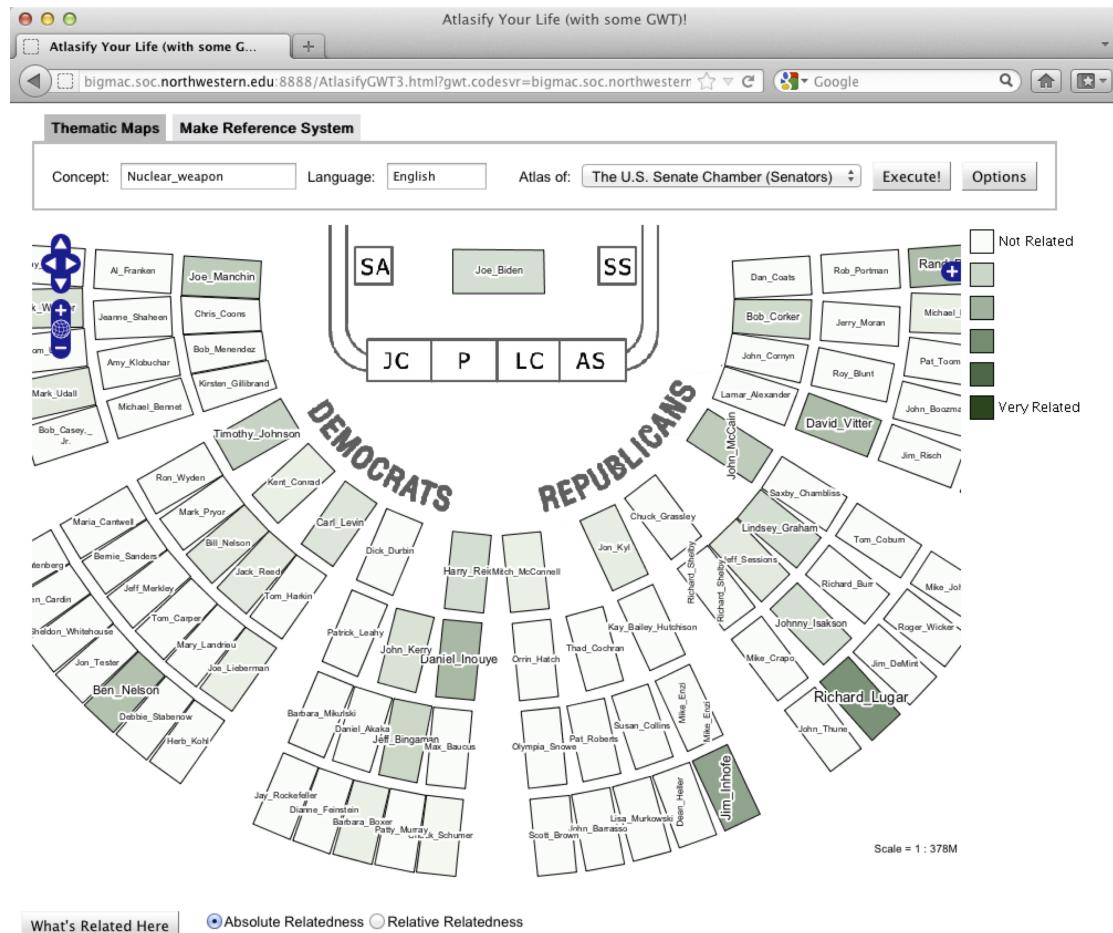


Figure 8-d: “Nuclear weapon” visualized on the “U.S. Senate Seating Chart” reference system.

The final challenge concerns generating natural language explanations of the relationships between the query concept and any spatial concept. These explanations realize the key paradigm of modern interactive cartography that users be able to click on a part of a map to obtain additional details [183, 184]. To address this challenge, we introduce the notion of *explanatory semantic relatedness measures (SR+E)*, which not only return a numeric estimate of the semantic relatedness between two concepts, but also *explain the identified relationships to end users*. We show how Wikipedia-based SR measures can be made explanatory by using machine learning to mine informative snippets of Wikipedia text. Furthermore, we describe how our SR+E measure, *AtlasifySR+E*, also uses machine learning to combine the explanations of its six constituent measures. Again, the approach of integrating the perspectives of each SR measure results in improved performance: our experiments demonstrate that *AtlasifySR+E*'s explanations outperform those of any single measure alone and other baselines.

In summary, our Atlasify work represents both a novel method for leveraging thematic cartography for domain-neutral exploratory search and the innovations in SR and information spatialization required to make that possible. In the following sections, we first describe related work and then discuss our solutions to each of the above challenges in more detail. Finally, we close by showing how the cultural context in user-generated content creates entirely new use cases for Atlasify.

## 8.1 Related Work

In this section, we cover research related to this chapter at a high level, with additional related work specific to each section of the chapter discussed in context. Our research falls into the area of exploratory search. White et al. write that exploratory search systems aid users with

information seeking problems that are “open-ended, persistent and multi-faceted” [206]. This stands in contrast to traditional Web search, which is primarily concerned with navigational queries and closed information requests. Despite the prevalence of exploratory queries, exploratory search is a relatively new research area with many open questions [206].

The field of cartography has identified several reasons why humans find thematic mapping useful for understanding complex geographic patterns. The known benefits of thematic maps are the communication of specific information [124, 187], the communication of regional/general information [114, 169], straightforward comparisons between maps showing different distributions [187], and straightforward comparisons between a mapped distribution and one’s mental model of depicted entities and regions [140, 187]. We enable these benefits in a wide variety of domains outside geography. For instance, in Figure 8-a it is easy to see that uranium specifically is quite related to nuclear weapons, but so is the entire “region” of actinides (the bottom row). An Atlasify user may recall from chemistry class that actinides have to do with the atomic age, so the fact that this region is highlighted reinforces the user’s mental model. Finally, comparing Figure 8-a with a periodic table visualization of, say, chemical weapons, it is easy to identify differences in the chemistry of the two concepts.

Our work within geographic reference systems is related to research in language models associated with geographic places. For example, Google Correlate [60] provides an interface to models based on georeferenced query logs. Others have leveraged geographic language models to study the geographic distribution of zeitgeist terms [87], to explore the use of relatedness-like metrics in a geographic context [85, 158], to make recommendations in social search [89], and for various other applications (e.g. [37, 100]). Some of this research has been echoed in the temporal domain (e.g. [158]). We extend this work by generalizing the notion of geographic

language models to arbitrary spatial reference systems, rather than just geographic and temporal ones. This research is also the first to our knowledge to (1) use geographic language models for exploratory search, (2) apply robust SR measures to geographic language models, and (3) use explanatory SR measures in this context (or any other).

## 8.2 Explicit Spatialization

Explicit spatialization (ES) is a novel form of information spatialization that, diverging from the existing spatialization literature, uses pre-defined reference systems (e.g. maps, figures, and diagrams) instead of data-driven reference systems. While ES is essential to our exploratory search approach, it also has implications beyond this work. Namely, it provides a new means by which advances in online mapping and geographic information retrieval (GIR) can be extended to domains outside of geography.

### 8.2.1 Definition of Explicit Spatialization

Explicit spatialization (ES) “spatializes” or “projects” any object  $o$  into a pre-defined reference system such as a periodic table, map, or seating chart. More formally, ES defines a process that represents an object  $o$  in terms of the spatial concepts  $E^{77}$  in a reference system  $rs$  according to the output of an ES function  $f_{ES}(o, E)$ . We clarify the key elements of this process below.

Let us consider Atlasify’s implementation of explicit spatialization. In Atlasify, each object  $o$  is a query concept (e.g. nuclear weapons) and the system’s ES function is our SR+E measure  $AtlasifySR+E$ . The spatial concepts considered include countries (and cities, landmarks, etc.) in

---

<sup>77</sup> In our original publication, we used the term “spatial entity” instead of “spatial concept.” Spatial concept is more appropriate in this thesis given the discussion in Section 3.10. However, in order to maintain a consistent formal definition of explicit spatialization, we keep the variable names from our original publication.

the “World Map” reference system, chemical elements in the “Periodic Table” reference system, and so on. Atlasify therefore spatializes each query concept into each reference system by running *AtlasifySR+E* on each query concept/spatial concept pair.

In explicit spatialization, each spatial concept  $e \in E$  in a reference system  $rs$  is comprised of a tuple  $\langle x, d \rangle$ , where  $x$  is a location (spatial footprint) in  $rs$ , and  $d$  is one or more data resources describing the entity. These data resources are mined by the ES function to spatialize the object  $o$ . In Atlasify,  $d$  consists of a single Wikipedia article describing each spatial entity.

The output of an ES function is a spatial distribution (“layer”) whose data model is a generalization of the canonical model of geographic information [58] (see Figure 8.2-a). The canonical geographic model formalizes an atomic unit of geographic information as a tuple  $\langle x, z \rangle$ , where  $x$  is a location in space-time of an entity on or near the surface of the Earth (e.g. its latitude / longitude coordinate or its polygonal representation) and  $z$  is a set of attributes

Figure 8.2-a illustrates two data models:

**a)** Canonical data model of geographic information:

	$x$	$z$				
		footprint	name	population	avg. mean July temp.	...
a)			Montana	989415	15.9°C	
			Nebraska	1826341	23.5°	

An arrow points from this table to a box labeled "Standard Cartographic and GIR Methods".

**b)** Explicit spatialization data model:

	$x$	$d$	$z$			
$rs = \text{Periodic Table}$	footprint	data resource	SR to query concept	explanations	...	
b)			Hydrogen	0.79	Hydrogen is related to query because...	
			Sodium	0.52	Sodium is related to query because...	

An arrow points from this table to a box labeled "Standard Cartographic and GIR Methods".

Figure 8.2-a: An example of the canonical data model of geographic information (a, top) and the explicit spatialization data model (b, bottom).

corresponding to that entity (e.g. temperatures, population counts). ES generalizes the geographic information model by replacing “the Earth” with an arbitrary reference system  $rs$  (such as the periodic table, an anatomical chart, etc.). The new model is equivalent to the geographic information model for a single fixed  $rs$  (except, of course, for the domain of the concepts). It is via this reduction that ES can use traditional cartographic and GIR methods with little to no modification.

The flexibility of the ES data model makes it adaptable to nearly any reference system in any domain. As one example of ES’s generality, consider a Web browser reference system that,

The screenshot shows the New York Times homepage with a search bar at the top containing the query "Arab People". Below the search bar, there are several news articles. On the left side, there is a sidebar with various news categories like World, U.S., Politics, etc. The main content area features several articles:

- Taking Lead, Iraqis Hope Commandos From U.S. Stay** (By TIM ARANGO) - A story about American troops staying longer in Iraq.
- President's Health Adds to Uncertainty in Yemen** (By LAURA KASINOF) - A story about President Ali Abdullah Saleh returning to power.
- Hezbollah Rejects Charges Over '05 Killing of Hariri** (By NADINE BARAKI) - A story about Hezbollah rejecting charges related to the killing of Rafik Hariri.
- Strauss-Kahn Case Adds to Doubts on Prosecutor** (By ALICE RYAN, JENNIFER ECKER and WILLIAM K. MARSHALL) - A story about the Strauss-Kahn case.
- Oil Spills Into Yellowstone River** (By ANAHAD O'CONNOR) - A story about an oil spill in the Yellowstone River.
- South Carolina's Young Governor Has High Hopes** (By KIM SEVERSON) - A story about Nikki Haley, the new South Carolina governor.
- MORE FROM THE REGION**
- Syrian President Fires**

On the right side, there is a "Sunday Review" section with columns for "GAY RIGHTS", "News Analysis", "Editorial", "Business Day", and a sidebar for "DealBook". There is also a sidebar for "see what's inside" with icons for different sections like Home Delivery and Personalized Weather.

Figure 8.2-b: Atlasify’s explicit spatialization of the query concept “Arab People” on the “New York Times Homepage” reference system. Atlasify correctly understands that the left column consists of Arab people-related stories. It also detects small increases in relatedness near “crude oil,” etc.

as a user browses the web, shows heat maps visualizing relatedness to a persistent concept of interest. We have implemented a static proof-of-concept of this idea in Atlasify’s “New York Times Homepage” reference system (Figure 8.2-b).

It is important to note that Atlasify’s implementation of explicit spatialization is far from the only possible approach. Other ES functions could include topic detection techniques or an algorithm that calculates concept-level sentiment. Similarly, a projected object  $o$  could be a blog post, an academic paper, or even an entire document collection, and data resources  $\mathbf{d}$  considered for each spatial concept could include tweets, images, or photo tags.

### **8.2.2 Relationship to Traditional Spatialization**

Traditional spatialization produces data-driven, abstract reference systems generally by applying dimension reduction to document collections for the purpose of visualizing those collections (see [76, 186] for an overview). ES, on the other hand, leverages existing reference systems (e.g. the periodic table, the human body, the surface of the Earth) for general IR applications. As a result, ES avoids the pitfalls that can make traditional spatialization undesirable for search [76], such as the error introduced by dimension reduction [76, 96, 186] and the imposition of a single, static visualization for an entire document collection [76]. While a few commercial document visualization systems (e.g. [168, 171]) have begun to explore extensible reference systems as in ES, they still rely on traditional spatialization as their primary paradigm.

### **8.2.3 Spatiotagging**

Preparing a new reference system for an ES application like Atlasify is a straightforward process that we call *spatiotagging*. Spatiotagging is a generalization of geotagging to arbitrary

spatial reference systems. To construct a reference system using spatiotagging, one simply identifies the spatial footprint (**x**) of the spatial entities in the reference system (e.g. chemical elements, Senate chamber seats, countries), and matches those entities to data resources (**d**) (e.g. corresponding Wikipedia articles). Spatial footprints can be identified by manually tracing the shapes of entities over a figure, diagram or image, obtaining pre-existing spatial representations (e.g. KML files or shapefiles), leveraging computer vision (e.g. OCR), or utilizing other techniques.

#### **8.2.4 User-defined Reference Systems**

While spatiotagging a new reference system is straightforward, it requires some effort. Further, users may not be able to find an existing reference system appropriate for their needs. In this section, we show how it is possible to extend ES to support ad-hoc, user-defined reference systems through the leveraging of semantic relatedness measures (Chapter 6).

Explicit spatialization enables the automatic construction of user-defined reference systems through three components: (1) predefined templates, which describe the general layout of the reference systems, (2) a category of concepts to act as spatial concepts, and (3) SR algorithms. Figures 8.2-c and 8.2-d provide a small use case of user-defined reference systems generated in Atlasify using the “spectrum” and “simplex” predefined templates respectively. In both figures, the spatial concepts are members of the Wikipedia category “Grammy Award winners,” and the SR measure is *AtlasifySR+E*.

Predefined templates and their “anchor concepts” make user-defined reference systems explicit. The “spectrum” template supports two anchor concepts and the “simplex” supports three. In Atlasify’s implementation, users can set these anchor concepts to any concept covered

by a Wikipedia article (e.g. Rock music, Hip hop music). Note that a reference system defined by a given set of anchor concepts remains fixed, independent of which category of concepts serves as spatial concepts or which concept is the query concept (i.e. it is not data-driven). As noted above, this is the key distinction between explicit and traditional spatialization.

In user-defined reference systems, the exact position of each spatial concept is defined by the SR between the corresponding concept and each of the anchor concepts. In other words, spatiotagging is done automatically in these reference systems using SR. If a spatial concept is very close to an anchor, this indicates that the corresponding concept is significantly more related to the nearby anchor than to the others. Only concepts that are non-trivially related to all anchors are included as spatial entities. In Atlasify's implementation, the category of concepts to act as spatial entities can be any Wikipedia category (Section 3.2.2).

As shown in Figure 8.2-d, user-defined reference systems are intended to be used as the basis for thematic cartography visualizations of query concepts just like standard ES reference systems. However, user-defined reference systems may also have value as exploratory search tools in and of themselves, without the thematic layer (e.g. Figure 8.2-c), but this more closely resembles traditional spatialization.

### **8.2.5 Spatial Information Retrieval**

Our exploratory search approach focuses on the cartographic benefits of explicit spatialization, but ES also has implications for geographic information retrieval (GIR). Namely, ES generalizes GIR to spatial information retrieval (SIR). In SIR, many GIR research areas – from understanding vague regions to toponym (place name) resolution to geographic relevance ranking to local search – can become relevant in non-geographic domains. For instance, Jones et

al.'s work on modeling vague geographic regions [100] like the English Midlands could be applied to numerous other reference systems, e.g. to model the “belly” or “tummy” vague regions in an anatomical reference system.

To demonstrate the possibilities of SIR, we have implemented in Atlasify one of the most basic GIR features: the simple bounding box spatial query. Users can issue these spatial queries by clicking Atlasify's “What's Related Here” button. Users are then presented with a list of concepts ranked by relatedness to the spatial region defined by the current view frame, which can then be filtered by Wikipedia category. This allows users to, for example, find out the concepts most related to the actinide elements or to the longest-serving members of the Democratic caucus (in the middle left of the seating chart).



Figure 8.2-c: A spectrum user-defined reference system of Grammy Award winners plotted from rock music to hip hop music. When used for thematic cartography visualizations, relatedness to the query concept is displayed in a similar fashion to Figure 8-c .

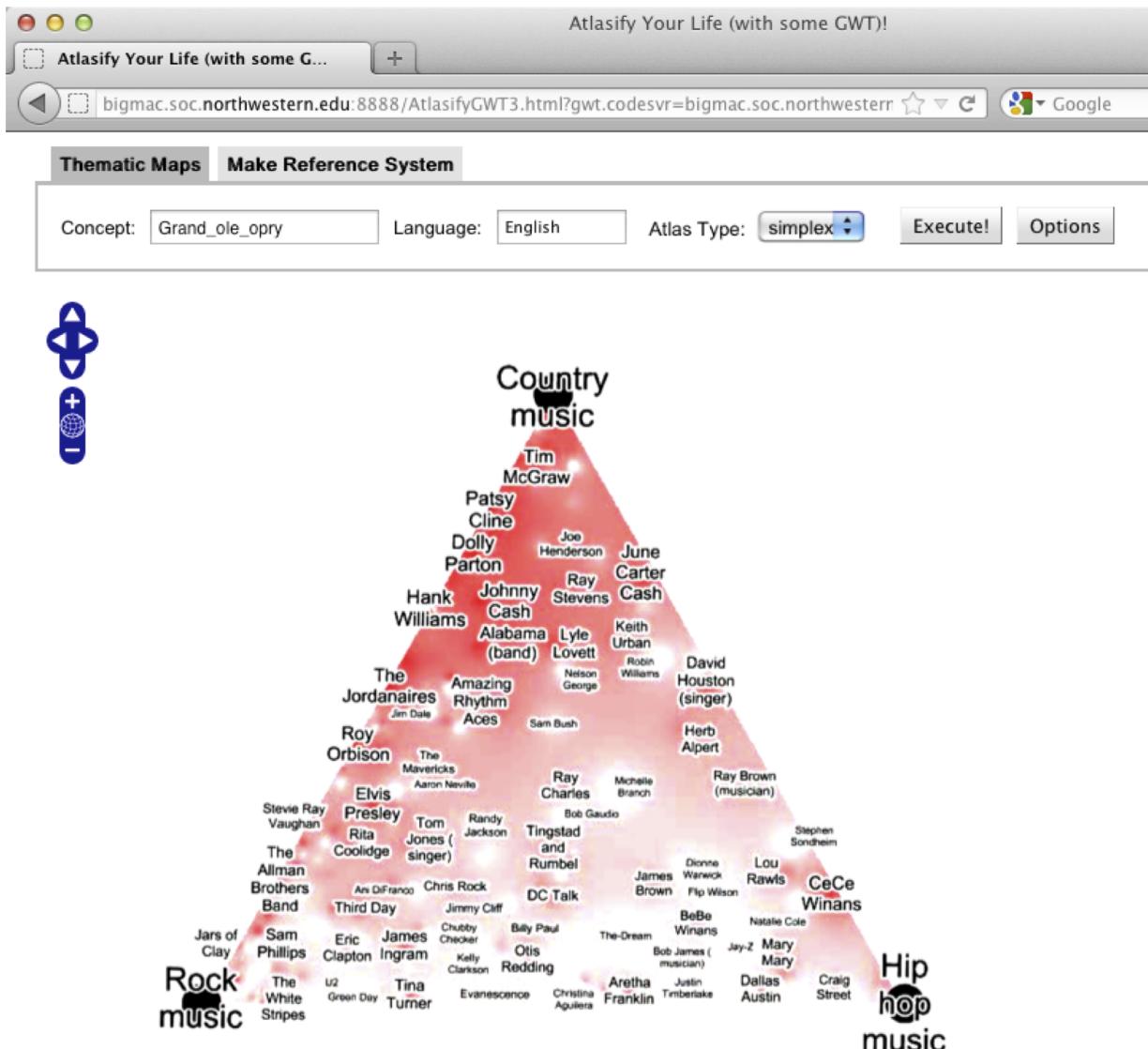


Figure 8.2-d: The query “Grand Ole Opry” is visualized on a simplex reference system defined by music genres, with the spatial concepts being members of the Wikipedia category “Grammy Award Winners.” It is clear that “Grand Ole Opry” is more related to country music than, say, rock music.

### 8.3 Explanatory Semantic Relatedness Measures (SR+E)

In this section, we introduce explanatory semantic relatedness measures (SR+E). Like traditional semantic relatedness (SR) measures, SR+E measures return a value that summarizes the number and strength of relationships between a given pair of concepts, e.g. (nuclear weapons, cobalt) [82]. However, along with each value, SR+E measures also provide a ranked list of *natural language explanations* of the various relationships underlying the value, in descending order of informativeness.

As noted above, SR+E measures play an integral role in our exploratory search approach. Each spatial distribution that our approach visualizes with thematic cartography is made up of SR estimates between the query concept and all spatial concepts in a reference system. While these visualizations show users the degree to which a query concept is related to a given spatial concept, the natural language explanations produced by SR+E measures describe *why* they are related. In doing so, the explanations provide users with “details-on-demand” [184] for a clicked spatial entity, following the principles of interactive cartography [183].

We begin our detailed discussion of SR+E measures below by introducing methods for generating explanations for *WikiRelate* [156, 192], *MilneWitten* [136, 137], and *Explicit Semantic Analysis* [47, 50] – and then do the same for the new measures we discussed in Chapter 6. We describe how each resulting SR+E measure mines Wikipedia’s text, links, or category graph to create explanations that reflect the relationships captured by the corresponding SR measure. We next cover *AtlasifySR+E*, the SR+E measure used in our exploratory search approach and implemented in the Atlasify system. *AtlasifySR+E* combines the benefits of the individual SR+E measures discussed above using a learned model. We hypothesized that this

ensemble approach could produce better SR estimates and explanations than any single measure alone. While we discuss the design of *AtlasifySR+E* here, our evidence that supports this hypothesis and descriptions of the related machine learning experiments are in Section 8.4.

Finally, we note that while our focus in this chapter is on utilizing SR+E for exploratory search, we expect the explanation mechanisms and improved SR measures will have broader applicability as well. As is discussed in more detail in Chapter 6, SR estimates are frequently utilized in NLP, AI, and IR [18, 50, 163, 219], and have been applied in tasks such as information extraction [18], clustering [11], and search [163].

### 8.3.1 Adding Explanations to SR Measures

As noted above, each SR+E measure must return a list of natural language relationship explanations ranked by informativeness. While there are other possible approaches, here we define the informativeness of each explanation to be based on two factors: the strength of the described relationship and the quality of the textual description. As such, each explanation must consist of natural language text, a relationship strength value, and a text quality value.

In all of our SR+E implementations, the text of relationship explanations is mined from Wikipedia. Several of the SR measures we considered implicitly calculate relationship strength when computing SR values. Where this is not true, we have developed strength metrics that are consistent with the SR measure’s overall algorithm. As is described in Section 8.4, we utilize machine learning techniques to map features of the textual explanations to an estimate of text quality, and combine this with relationship strength using heuristics to arrive at a final ranking. The heuristics differ for each SR+E measure, but they generally weigh relationship strength more heavily than text quality.

We now turn our attention to the approach we took to adding explanatory capabilities to each of the individual SR measures considered here.

### 8.3.2 WikiRelate Explanations

*WikiRelate* leverages a variant of the WCG path length between the articles  $a$  and  $b$  to estimate  $\text{SR}(a,b)$ . The insight behind this design is that each path represents a relationship between  $a$  and  $b$ , and the shorter the path, the stronger the relationship. We constructed *WikiRelate* explanations to elucidate these relationship paths to the user in natural language. For example, Figure 8.3-a displays a *WikiRelate* explanation for the strongest relationship between Chemistry and Mathematics (the shortest WCG path between the two articles). In the case of *WikiRelate*, text quality is not considered in the informativeness function as the natural language is automatically determined in the same way for all explanations.

- Chemistry and Mathematics both belong to Category:Academic Disciplines**
- Chemistry**
- 1) Chemistry is a member of Category:Natural sciences, which is a member of Category:Academic disciplines
- Mathematics**
- Mathematics is a member of Category:Formal sciences, which is a member of Category:Academic disciplines
- Cheese links to France**
- 2) CHEESE: WORLD PRODUCTION AND CONSUMPTION:  
The biggest exporter of cheese, by monetary value, is France; the second, Germany (although it is first by quantity)
- Beer in Ireland discusses both Ireland and Beer**
- 3) Though Ireland is better known for stout, 63% of the beer sold in the country is lager.
- Life and Death link to Organism**
- LIFE:
- In biology, the science of living organisms, life is the condition that distinguishes active organisms from inorganic matter. Living organisms undergo metabolism, maintain homeostasis, possess a capacity to grow, respond to stimuli, reproduce and, through natural selection, adapt to their environment in successive generations.
- DEATH:
- Death is the termination of the biological functions that sustain a living organism.
- Chocolate chip links to United States and Chocolate**
- CHOCOLATE CHIP:
- Chocolate chips are small chunks of chocolate.
- CHOCOLATE CHIP: AVAILABILITY
- Today, chocolate chips are very popular as a baking ingredient in the United States and the chocolate chip cookie is regarded as a quintessential American dessert.

Figure 8.3-a: Sample explanations from AtlasifySR+E's constituent SR+E measures. The format of these explanation mimic that of the Atlasify interface. (1) The top WikiRelate explanation for the concept pair (Chemistry, Mathematics). (2) The top WAGDirect explanation for the concept pair (Cheese, France). (3) The top ESA explanation for the concept pair (Ireland, Beer). (4) The top OutlinkOverlap explanation for the concept pair (Life, Death). (5) The top MilneWitten explanation for the concept pair (United States, Chocolate).

### 8.3.3 MilneWitten, OutlinkOverlap, and WAGDirect Explanations

As discussed in Section 6.1, *MilneWitten* operates by comparing the set of articles that link to the articles  $a$  and  $b$ . The intuition is that if  $a$  and  $b$  share many inlinks, they should be assigned a high SR score. The relationships considered here are indirect: a shared inlink means that an article  $c$  links to both  $a$  and  $b$ . Explanations based on *MilneWitten* must therefore elucidate the nature of these  $a \leftarrow c \rightarrow b$  relationships. Figure 8.3-a displays the most informative *MilneWitten* explanation for the concept pair (United States, Chocolate) ( $c$  = “Chocolate chip”).

However, in order to establish that the explanation in Figure 8.3-a was the top explanation – recall that explanations are ranked by informativeness, which is a function of strength and text quality – our *MilneWitten+E* implementation needed a way to measure the strength of each  $a \leftarrow c \rightarrow b$  relationship. In other words, we required some method of determining that  $c$  = “Chocolate chip” represents a stronger  $a \leftarrow c \rightarrow b$  relationship than, say,  $c$  = “List of Viva Piñata Episodes,” which also links to both “Chocolate” and “United States.” To solve this problem, we use bootstrapping to calculate  $\text{MilneWitten}(a,c)$  and  $\text{MilneWitten}(b,c)$ . The strength of each  $a \leftarrow c \rightarrow b$  relationship is then computed by taking  $\text{MilneWitten}(a,c) * \text{MilneWitten}(b,c)$ . This algorithm results in the relationship involving “Chocolate chip” being deemed the strongest relationship, with that involving “List of Viva Piñata Episodes” much further down the list.

We have also implemented a modified version of *MilneWitten*, *WeightedMW*, that more heavily weights the links that occur in the gloss of the article. The experiments in Section 8.4 show that this weighted measure estimates SR values somewhat better than our implementation of *MilneWitten*. Explanations are generated in the same fashion as in standard *MilneWitten*.

*MilneWitten* and *WeightedMW* cannot detect two important types of relationships present in

the WAG. Recall from Chapter 6 that we designed *WAGDirect* to capture one of these types of relationships, those that occur when  $a$  links directly to  $b$  ( $a \rightarrow b$ ) or vice versa ( $b \rightarrow a$ ).

Explanations of *WAGDirect* relationships thus consist of text snippets from article  $a$  that discuss  $b$ , and/or vice versa (Figure 8.3-a), without any intermediary article  $c$ .

Recall also that *OutlinkOverlap* captures the other type of WAG relationships not considered by *MilneWitten* and *WeightedMW*: the overlap of the set of outlinks of  $a$  and  $b$ . *OutlinkOverlap* explanations thus describe how  $a$  and  $b$  discuss these mutually outlinked articles. In other words, they include text snippets from  $a$  and  $b$  that explicate the  $a \rightarrow c \leftarrow b$  relationships considered by this SR measure (see Figure 8.3-a). *OutlinkOverlap* relationship strengths are calculated in a similar manner as *MilneWitten* strengths.

### 8.3.4 Explicit Semantic Analysis Explanations

As is discussed in Section 6.1, to produce SR estimates, *ESA* considers the co-occurrence of  $a$  and  $b$  in a large number of Wikipedia articles  $C$ . Specifically, *ESA* represents  $a$  and  $b$  as vectors of bag-of-words similarity to each article  $c$  in  $C$ . It then compares these vectors using cosine similarity. The relationships considered by *ESA* are thus co-occurring mentions of  $a$  and  $b$  in each Wikipedia article in the concept space. Stronger relationships are defined by articles in  $C$  that more frequently mention both  $a$  and  $b$  (with consideration for document frequency as well), and strength can be estimated by comparing the combined values in each vector dimension while calculating the cosine similarity. Explanations derived from *ESA* thus describe the co-occurrence of mentions of  $a$  and  $b$  in each article  $c$  in  $C$  in a human-readable fashion (Figure 8.3-a).

### 8.3.5 AtlasifySR+E

The SR+E measures discussed above capture distinct relationship types. *WikiRelate* tends to

operate on classical relations [18] such as *isA* (hyponymy/hypernymy) and *hasA* (meronymy/holonymy) [192]. The WAG-based SR measures are more capable of discovering non-classical relations [18], such as *isTheBiggestExporterOf* (Figure 8.3-a). Finally, *ESA* discovers the “distributional” relationships [18] inherent to text co-occurrence.

*AtlasifySR+E*, the algorithm employed in our exploratory search approach, combines all six previously discussed SR measures. The goal in doing so was to develop an SR+E measure that understands all three types of relationships. We hypothesized that such an ensemble measure would produce both (1) better SR estimates and (2) better relationship explanations.

*AtlasifySR+E*’s SR estimate for a pair of terms is the output of a learned model whose features include the estimates of each constituent SR measure as well as features like the word sense entropy of the pair. *AtlasifySR+E* generates explanations for these estimates using a different learned model to select the best explanation among those output by each constituent measure.

*AtlasifySR+E* then iterates, choosing the next best explanation, resulting in a long ranked list of explanations.

The experiments section that follows describes in detail each of the learned models and their associated machine learning experiments. We also show below that both of our hypotheses related to the combining of SR measures for improved performance were supported.

## 8.4 Evaluation Experiments

Evaluation of exploratory search systems is a notoriously difficult problem [205, 206]. In this chapter, our evaluation strategy is to investigate the performance of the individual components of our exploratory search approach. Specifically, we focus the evaluation on our method of projecting query concepts into spatial distributions using *AtlasifySR+E*’s relatedness

estimates and explanations. This has the added value of confirming these components as independent contributions. Once these spatial distributions have been created, thematic cartography's well-evaluated techniques (see [187] for an overview) can be employed.

Below, we first describe experiments that demonstrate the state-of-the-art accuracy of our SR estimates. Next, we discuss how we collected over 2,500 human judgments of explanation quality and used these judgments to train a ranker whose performance significantly exceeds baseline approaches.

#### **8.4.1 SR Value Estimates**

Accurate SR value estimates are integral to our exploratory search approach. The colors, text sizes, and other visual variables in Figures 8-a - 8-d, 8.2-c and 8.2-d are defined by *AtlasifySR+E*'s estimates of the SR between each spatial concept and the query concept. Our method for achieving high-quality SR is to combine the estimates of the six SR measures mentioned above using machine learning, and use the resultant trained model to generate *AtlasifySR+E*'s estimates. In this section, we describe this machine learning approach and evaluate the accuracy of *AtlasifySR+E*'s SR estimates against benchmark SR data sets.

We first ran an experiment to validate the performance of our implementations of SR measures from previous work. Following standard practice, we evaluated each implementation by comparing its SR estimates with datasets of human gold standard estimates using Spearman's  $r_s$  and Pearson's  $r$  (Table 8.4-a). As is discussed in more detail in Chapter 6, these datasets consist of term pairs and associated SR values, which are averaged across all human annotators of a dataset. We used two long-standing SR datasets, *WordSim353* [43] and *MC30* [133], as well as *TSA287* [163] and *Atlasify240*, the SR dataset we developed as part of the experiment

described in the following section and also heavily used in Chapter 6. The results in Table 8.4-a indicate that our implementations are satisfactory.

Our approach to combining the estimates of each constituent SR measure was to use a regression model to predict the human gold standard judgments in *WordSim353*, the most common SR dataset in the literature. We then used this trained model to predict the gold standard judgments in the four SR datasets discussed above. The regression model employed a variety of features, including the SR estimates produced by each constituent measure, along with numerous properties of the Wikipedia article corresponding to each term in a term pair (e.g. article length, link density). Our model also included as a feature the entropy of the word sense disambiguation task required to identify matching articles for each term. *AtlasifySR+E* uses a pairwise maximization approach for word sense disambiguation [136, 192], wherein word sense candidates are identified using anchor texts.

We found that a boosted implementation of Quinlan's M5 algorithm for smoothed trees of linear regression models achieved good performance using 10-fold cross validation (mean  $r_s = 0.75$  with gold standard values). Among the most predictive features in the model were the SR scores generated by the constituent algorithms and the word sense entropy of the term pair. The constituent SR measure with the most predictive power was *ESA*.

We then evaluated the performance of our new *AtlasifySR+E* measure using the same experimental setup as above. The full results can be seen in Table 8.4-a. *AtlasifySR+E* performs better than all Wikipedia-specific measures on every dataset but *MC30* for both correlation metrics, and the *MC30* differences are not significant. Further, we could not detect a statistically significant difference between *AtlasifySR+E*'s Pearson's correlations and the inter-annotator agreement in every case.

We also could not detect a significant difference between the accuracy of SR estimates generated by *AtlasifySR+E* and those generated by *TSA*, which is the current state-of-the-art SR algorithm. *AtlasifySR+E* relies only on Wikipedia data while *TSA* additionally uses exogenous information in the form of a large set of *New York Times* abstracts stretching over decades. This data is language-specific, less accessible than Wikipedia, and less open. *AtlasifySR+E* may thus be preferable to *TSA* in, for example, for-profit settings and non-English contexts (see Section 8.5). We also note that *AtlasifySR+E*'s ensemble approach – improving performance by combining different perspectives on the relatedness between concepts – can incorporate additional perspectives on relatedness, such as *TSA*'s temporal approach and future innovations.

SR Algorithm		MC30		WordSim353		TSA287		Atlasify240	
		$r_s$	$r$	$r_s$	$r$	$r_s$	$r$	$r_s$	$r$
WikiRelate	AtlasifySR+E	.78	.82	.49	.48	.40	.47	.52	.53
	Published	-	.57	-	.53	-	-	-	-
MilneWitten	AtlasifySR+E	.64	.65	.56	.52	.49	.45	.68	.69
	Published	.70	-	.69	-	-	-	-	-
WeightedMW		.65	.65	.66	.57	.53	.46	.74	.72
WAGDirect		.71	.73	.64	.58	.49	.53	.60	.56
OutlinkOverlap		.64	.67	.52	.42	.48	.42	.61	.51
ESA	AtlasifySR+E	.74	.77 <sup>†</sup>	.72	.70 <sup>†</sup>	.58	.62 <sup>†</sup>	.71	.72 <sup>†</sup>
	Published	.72	-	.75	-	-	-	-	-
TSA (current SoA)	Published	-	-	.80	-	.63	-	-	-
AtlasifySR+E		.75	.81	.78 <sup>‡</sup>	.76 <sup>‡</sup>	.64	.68	.78	.77
Inter-annotator Agreement		n/a	.90	n/a	.55-.73	n/a	-	n/a	.77

Table 8.4-a: The performance of the SR measures considered in this chapter; in context with that of their published versions. Where inter-annotator agreement (InterAA) is available, bold indicates results with which we could not detect a significant difference with InterAA using the method in [47] and  $p < 0.05$ . Where it is not available, bold indicates the top result and those with which we could not detect a significant difference with the top result. InterAA is not included for Spearman's  $r$  ( $r_s$ ) due to the prevalence of ties [219]. Note that AtlasifySR+E is the only measure that is bold in all columns, including those for which there is data for the current state-of-the-art, TSA.

<sup>‡</sup> The model was trained on this dataset.

<sup>†</sup> The log of the estimates has been used for improved performance

## 8.4.2 Explanation Ranking Experiments

Each of *AtlasifySR+E*'s constituent SR+E measures returns a list of explanations ranked by their informativeness (Section 8.3). *AtlasifySR+E* must then consolidate and rank the explanations from each measure into a single list to return to the user when they click on a spatial entity. We approached this explanation ranking task as follows: given a concept pair and the up to six top-ranked (most informative) explanations from the constituent measures, *AtlasifySR+E* is to select the best explanation. *AtlasifySR+E* then iterates, removing the explanation it judged to be most interesting at each iteration and placing it in order in the list of explanations to be returned to the user. In the case of the constituent SR measure whose explanation was placed in the returned list, the next most informative explanation is considered in the subsequent iteration. Solving this ranking problem involved gathering a dataset from human judges and then using this dataset to train, develop, and test a ranker. We describe this effort below.

### 8.4.2.1 Data Collection

Our training data was based on 268 manually selected concept pairs. Each concept mapped unambiguously to a Wikipedia article, and, following one approach in the literature, the concept pairs were hypothesized to uniformly cover the spectrum of semantic relatedness. While 28 of these concept pairs come from *WordSim353*, 240 are original pairs not seen before in the SR literature. These 240 pairs make up the *Atlasify240* dataset, which is focused on named entities. Named entities make up a large majority of concepts in spatial reference systems (e.g. John McCain, Israel, Helium). Existing datasets (e.g. [43, 133, 163]) include relatively few named entities, necessitating new concept pairs for our evaluation.

Each of the most informative (top-ranked) explanations from *AtlasifySR+E*'s constituent SR+E measures was generated for all of the 268 pairs and placed in a Web interface (when there was an explanation available). The interface allowed human annotators to rank the explanations for each pair of articles using drag-and-drop techniques. The presentation order of both the pairs and the explanations were randomized. Prior to ranking explanations for a pair, annotators were required to provide an SR estimate. Following the existing SR literature [152, 163], annotators were able to rank SR on a limited scale, in our case from 0 (not related) to 4 (very related). After ranking the available explanations, annotators were asked if they thought that their top-ranked explanation was a good explanation of a relationship between the two concepts.

Ten annotators finished all pairs. On average, annotators said 66% of their top-ranked explanations were good explanations of the relationship between the two concepts. As hypothesized, *WAGDirect* was by far the best algorithm, with 55% of its explanations being chosen as the best on average. However, *WAGDirect* was only able to produce an explanation in 26.8% of cases because only that many of the article pairs had at least one link between them. *WikiRelate* was worst performing algorithm, but was still selected 13.8% of time when it was available. The *MilneWitten* algorithms were the most prolific and were each able to generate an explanation for over 80% of the samples.

For 18 (6.7%) pairs, no algorithm was able to generate an explanation. This is to be expected for pairs with very low SR; where there is no relatedness, there is no relationship to explain. Indeed, the average mean SR judgment for these pairs was 0.52 (in a 0-4 range). In contrast, the average mean SR judgment for pairs for which all six algorithms generated explanations was 3.78.

#### 8.4.2.2 Machine Learning

Using the hand-annotated ranks from our data collection process, we developed a dataset that consisted of numerous features for each explanation, including: (1) the SR value estimate from the constituent SR+E measure, (2) the textual quality of the explanation (described in the following section), and (3) an indicator of which SR+E constituent measure produced the explanation. For each pair, we assigned the explanation with the lowest (i.e. best) mean rank a “1” and every other explanation a “2.” We trained a ranker to predict the best (“1”) explanation using SVMRank [98].

The results of this experiment can be found in Table 8.4-b. We report these results in terms of coverage, which is the percentage of pairs for which one or more explanations were available, and precision, which is the percentage of pairs for which *AtlasifySR+E* correctly identified the best explanation (when one or more were available).

Using 10-fold cross-validation, our best performing model had a precision of 56%, which is significantly better than random guessing ( $X^2 = 13.2, p < .01$ ) and only 2% lower than mean inter-annotator agreement (58%). In other words, the model predicts the best explanation almost as well as humans agree on the best explanation. The difference between the model and the inter-annotator agreement is in fact not significant ( $X^2 = 0.51, p = .48$ ). Moreover, this model results in

Model features	Precision	Coverage
All features	56%	93%
Measure indicators only	51%	93%
Random	39%	93%
<i>WAGDirect</i> (Highest Precision Single SR Measure)	55%	27%
<i>MilneWitten</i> (Highest Coverage Single SR Measure)	35%	80%

Table 8.4-b: Results of our explanation ranking experiment.

Model features	<i>r</i> with gold standard
All features	0.32
Contextual features only	0.29
Syntactic features only	0.24
Human Inter-rater Agreement	0.51

Table 8.4-c: The results of our text snippet quality experiment.

a slightly better precision (insignificantly so) than *WAGDirect*, the best SR+E algorithm for explanations, and has a much higher coverage; it can return an explanation when any of the constituent algorithms can find an explanation. In our experiment, this was 93.3% of the time, compared to *WAGDirect*'s 26.8%.

It is important to note that a model based only on which SR+E method was used (“Measure indicators only”) performs nearly as well as the full model, and the difference between them is not significant ( $X^2 = 1.27, p = .15$ ). That is, the relative performance of the constituent SR+E explanation generators accounts for most of the predictive power of our ranking model.

#### 8.4.3 Quality of Mined Text

The final machine learning experiment we will discuss assesses the quality of text mined from Wikipedia. This quality assessment, along with relationship strength estimates, is used to calculate the informativeness of the explanations for each of *AtlasifySR+E*'s constituent SR+E measures (Section 8.3). This informativeness is then used to rank explanations within each individual measure.

Hand-annotated data was supplied by four annotators, each of whom rated 500 snippets on a scale from 0 to 4 according to the quality of the natural language. Each snippet describes one “leg” of a relationship (e.g. some explanations in Figure 8.3-a have two snippets, while others only have one). Quality was assessed using several factors, including readability and clarity of

relationship described. Inter-annotator reliability was  $r = 0.51$  (calculated with Fisher's z-value transformation [217]).

For training, each snippet was assigned two types of features: syntactic (e.g. lack of a verb) and contextual (e.g. at the top of the page). After experimenting with a variety of regression models, we found a linear regression model to be the most accurate. Using 10-fold cross-validation, this model was able to achieve a mean correlation of  $r = 0.32$  (Table 3). While the combined model outperforms a model trained on only a single type of feature, models trained on either type of feature alone were not found to have significantly worse predictive power.

## 8.5 Atlasify and Cultural Context in User-Generated Content

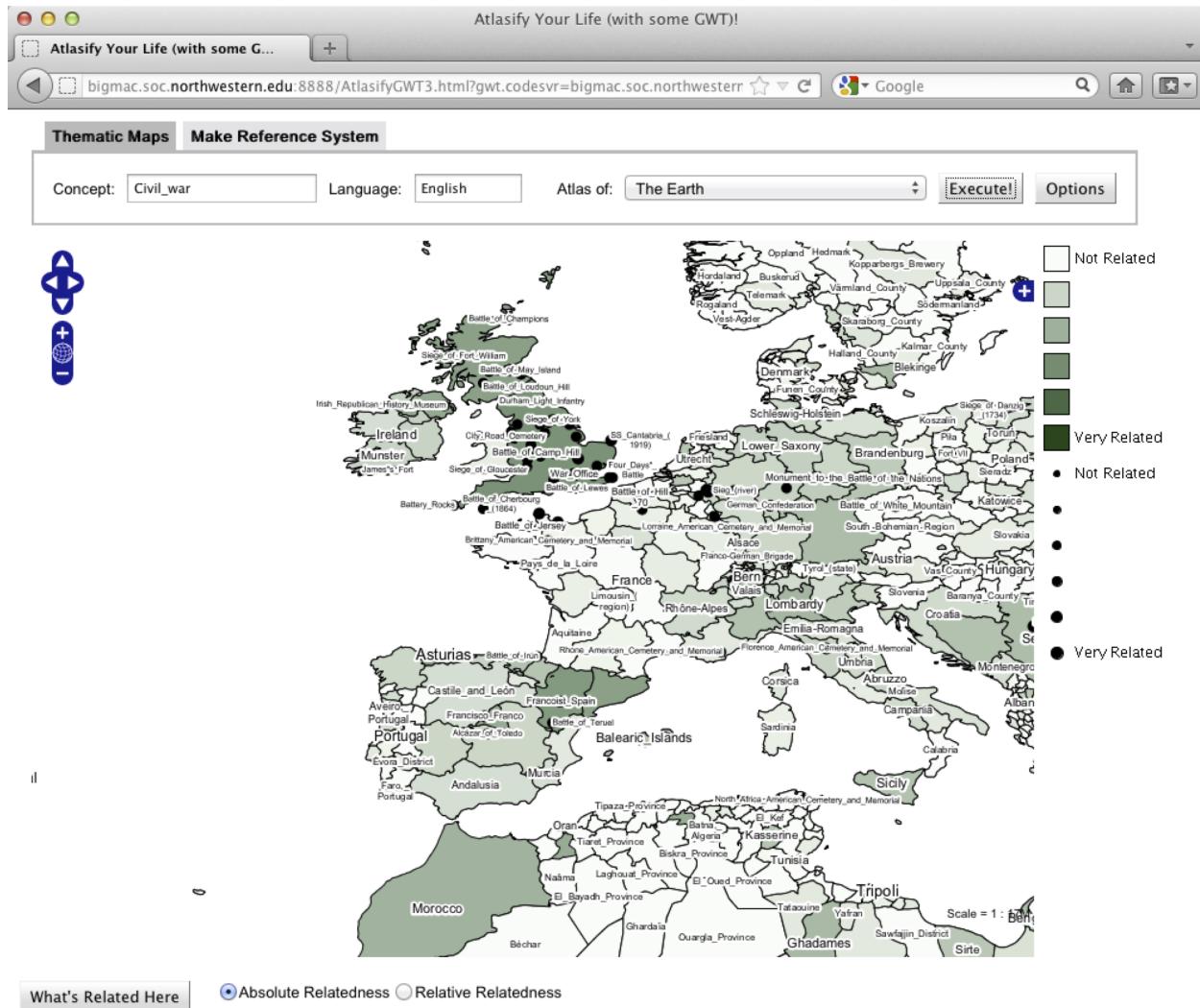
Thus far in this chapter, we have described in detail how Atlasify enables an entirely new approach to exploratory search and how it makes contributions in thematic cartography and natural language processing in order to support this approach. Throughout this discussion, we have been exclusively focused on Atlasify’s performance on experiments and use cases when it is leveraging the English Wikipedia as world knowledge. However, the attentive reader will have noticed that in many of the screenshots above, next to the query input box there is another input box marked “Language.” Using this input box, Atlasify allows users to, by typing a couple of letters, switch Atlasify’s world knowledge to any of the 25 language editions we have considered throughout this thesis. In the context of the nuclear weapons example, this means that users are able to engage with interactive cartographic visualizations of the geography, history, chemistry, politics, etc. of nuclear weapons as understood by the Japanese Wikipedia just as easily as they are able to with the English Wikipedia. Moreover, the same is true for the other 23 language editions.

Students in any introductory cartography class are taught that maps have reflected the cultural contexts of their cartographers effectively ever since the first map was produced [187]. From maps in which geographic features important to minority groups are ignored (e.g. Native American burial mounds) to propaganda maps that are deliberately manipulated to communicate a certain point of view [139], cartography provides an excellent use case for the importance of accessing information from multiple cultural perspectives. There have been many enlightening qualitative comparisons of maps from different cultures, focusing on both professionally-made maps [139] and those made by amateurs (e.g. the mental maps research discussed in Section

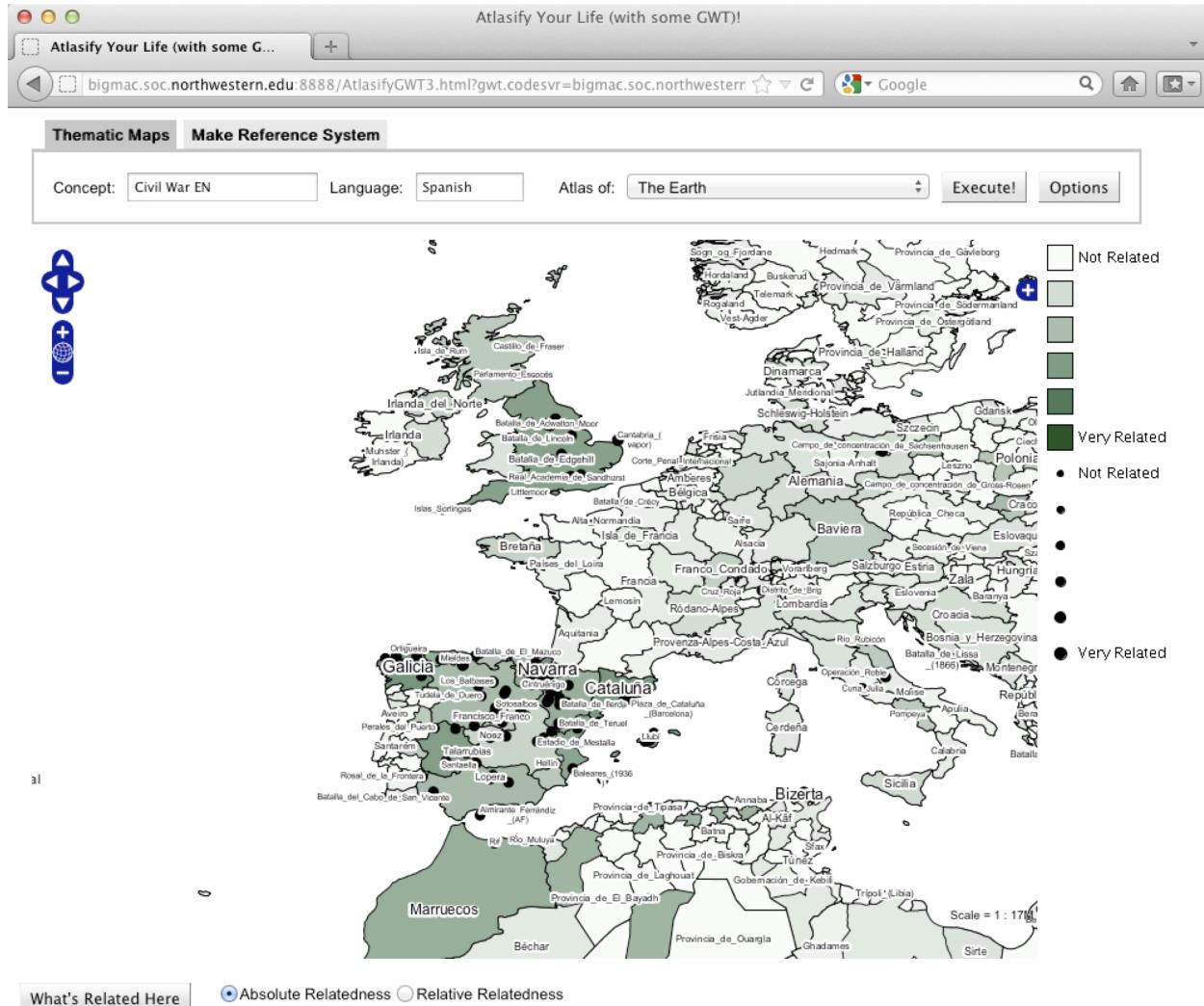
2.2.3). However, to our knowledge, Atlasify represents the first system that allows users to automatically generate these comparisons, let alone do so on a concept-by-concept basis for over 8.67 million concepts and do so in domains *other than geography*. While these comparisons will be inherently less nuanced than qualitative investigations that are the product of long-term, detailed study, they certainly make it much easier to take the first steps in this direction.

Consider, for instance, Figures 8.5-a and 8.5-b, which show the results of an Atlasify query for the concept of civil wars using the English (Figure 8.5-a) and Spanish (Figure 8.5-b) language editions. The map generated from the world knowledge in the English Wikipedia shows an obvious bias towards the English civil wars, with places most important to these wars (and the United Kingdom as a whole) indicated as being very related to the query concept. In the map generated using knowledge from the Spanish Wikipedia, it is Spain that is shown to be most related to the query concept, with key locations in the Spanish Civil War highlighted.

Taking a step back here, what we have effectively done with Atlasify is turn the challenge to existing technologies presented by cultural context in UGC (Chapter 6) into an advantage. That is, by taking care to support multiple language editions in almost every aspect of the Atlasify system (and its contributions), we have embraced – rather than ignored – the global diversity hypothesis as it applies to SR measures. In doing so, we have made the cultural context in UGC a feature, not a bug.



*Figure 8.5-a: The query concept civil war plotted by Atlasify on the “World Map” reference system using world knowledge from the English Wikipedia. Note that the most-related places to civil wars are important locations in the English civil wars, with the entirety of the United Kingdom being assessed as quite related as a result.*



*Figure 8.5-b: The query concept civil war plotted by Atlasify on the “World Map” reference system using world knowledge from the Spanish Wikipedia. The center of emphasis has shifted from England in the preceding figure to Spain, with places related to the Spanish Civil War indicated as being very related.*

Finally, before concluding this chapter, it is important to note that we are also exploring the application of explicit spatialization for the study of cultural context in UGC outside of the immediate Atlasify setting. Recall that explicit spatialization involves “projecting” an object  $o$  into a reference system defined by spatial concepts  $E$  using an explicit spatialization function  $f_{ES}(o, E)$ . While in Atlasify  $o$  is the query concept and the spatialization function is *AtlasifySR+E*, as we noted above, there is no reason that any of these parameters cannot be varied. We have recently begun to investigate extending our self-focus bias work in Section 3.10 to non-geographic reference systems by doing just this. Specifically, we are setting  $o$  = the WAG of a given language edition and  $f_{ES}$  = any of the WAG-based prominence metrics discussed in Section 3.10. Initial results have shown explicit spatialization to be a powerful way of communicating self-focus bias (i.e. cultural context) outside of the domain of geography. For instance, consider Figure 8.5-c, which is analogous to figures like Figure 3.10-m in that it shows the relative normalized PageRank sums of two language editions, in this case Japanese and English. However, instead of using the “World Map” reference system, Figure 8.5-c uses the “Periodic Table” reference system. It is through this visualization paradigm that it becomes quite clear that the Japanese Wikipedia considers certain elements to be much more important to world knowledge than the English Wikipedia does. The three elements for which this is most strongly the case – Plutonium, Cesium, and Strontium – all have to do with major events in Japan related to nuclear warfare and nuclear power (e.g. Fukushima, World War II).

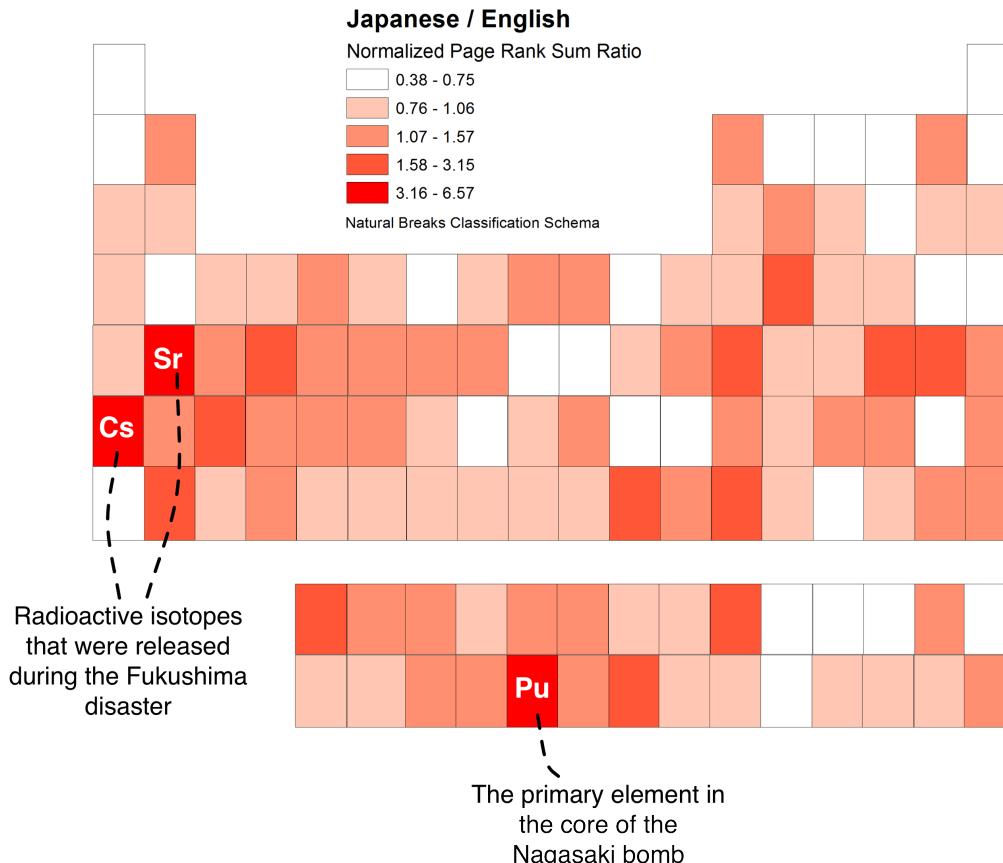


Figure 8.5-c: The Japanese PageRank score sum over the English PageRank score sum for the “Periodic Table” reference system. Darker red indicates a greater prominence in the Japanese Wikipedia than the English Wikipedia.

## 9 Conclusion and Future Work

When the term “user-generated content” first appeared in the popular lexicon, the notion that content voluntarily contributed by Internet users could be the lifeblood of important technologies was both foreign and fascinating. As we conclude this thesis in early 2013, this is so common as to make the term “user-generated content” somewhat blasé and old-fashioned. From Wikipedia articles to Instagram photos to Yelp reviews to Pinterest pins to Twitter microblogs, a growing proportion of the information with which we interact – both directly and through UGC-based technologies – has been produced by our peers. Indeed, user-generated content forms an enormously important component of the “big data” revolution that is likely to have had a large impact on any reader of this thesis more than few years into the future.

As we move further and further into this “big data” era, this thesis reminds us that despite their seemingly antiseptic nature, many “big data” repositories of information remain *decidedly human*. Consider the case of Wikipedia. Here we have a repository of encyclopedic world knowledge so large, popular, and well-regarded that it is a part of the daily experience of millions of people and serves as the “brains” of hundreds of computer science research projects and computing technologies. However, we saw in Chapter 3 that this knowledge is encoded in a fashion that intensely reflects the cultural memberships of its editors. Nowhere is this more evident than in Section 3.10, in which we showed that almost every language edition of Wikipedia is constructed in manner that puts the corresponding language-defined community’s home cultural regions at the very center of all of encyclopedic world knowledge.

This thesis argues that the cultural context in user-generated content represents both a challenge and an opportunity for the computer science community. The challenge is that if we

continue to ignore UGC’s cultural contextualization, we risk injecting cultural bias into the increasing number of “big data” technologies that are only able to exist thanks to user-generated content. This is something we saw quite clearly in Chapter 6, where we showed that the same algorithm operating on world knowledge from different language editions of Wikipedia produced substantially different results.

The opportunity lies in the enormous possibility for a new class of technologies that is enabled by the cultural context reflected in user-generated content, a class of technologies that is highlighted by our Omnipedia and Atlasify applications discussed in Chapters 7 and 8. Without the cultural context in UGC, Omnipedia would not be able to show to users how over 8.67 million concepts are understood across 25 different language-defined cultures. Similarly, without this cultural context, Atlasify would not be able elucidate the geography, history, chemistry, politics, and so on of these 8.67 million concepts as it is defined by this same group of language-defined communities.

To summarize, the three major contributions of this thesis are as follows:

- We mined and measured the cultural diversity in several databases of user-generated content. In doing so, we showed that UGC strongly reflects the many cultural memberships of its contributors, often to an extent far larger than that which has been assumed in the literature.
- We showed that the traces of contributors’ cultural memberships embedded in user-generated content have important implications for the many existing technologies that leverage UGC as a source of world knowledge. Specifically, we demonstrated that by using a single culture’s user-generated content as its “brains,” a technology can adopt the viewpoint of that particular culture at the expense of the perspectives of many other cultures.
- Finally, we highlighted the substantial upside to the cultural context in UGC for designers

of UGC-based technologies. Specifically, we demonstrated through two applications we built – Omnipedia and Atlasify – that the traces of contributors' cultural memberships in UGC can enable an entirely new class of technologies that make visible the similarities and differences in cultural perspectives around the world.

Before closing this thesis, it is important to discuss what we believe to be the two most important directions for future work in this area. First and foremost, the extent to which UGC reflects the cultural memberships of a larger variety of cultures must be established. In this thesis, we have focused on language- and geography-defined cultures. However, as noted by Clark [24], there are a plethora of other cultural communities to consider. These include age-defined cultures, ethnicity-defined cultures, income-defined cultures, profession-defined cultures, religion-defined cultures, and so on. For our part, we are beginning to investigate these issues with a large-scale extension of the self-focus bias work in Section 3.10. Specifically, through the combination of geographically-referenced census data with geographically-referenced UGC from repositories including Flickr, Twitter, Foursquare, and Wikipedia, we are exploring the extent to which the perspectives of the cultural groups thought to be most active in these UGC communities dominate the corresponding repositories. The census data allows us to explore this question along the lines of a variety of cultural communities, from those defined by income to those defined by profession to those defined by age.

The second vital direction for future work is the investigation of the impact of new technologies on the degree of cultural context in user-generated content. In Chapter 4, we saw evidence that the properties of technologies related to a given UGC repository can substantially shape the cultural character of the repository. This study was focused on geographically-referenced user-generated content, so we referred to these technologies as the “spatial content

production model” (SCPM) of a given repository. SCPMs that required more local contributions – i.e. “you have to be there” SCPMs – resulted in a greater diversity of content across geographically-defined cultures. The reverse was true of SCPMs that facilitated “total time-space compression” [73] by allowing anyone to contribute content about anywhere in the world, regardless of their geographic cultural memberships (i.e. “flat Earth” SCPMs).

As we noted in Chapter 4, this same principle applies in a non-geographic context. For instance, the language technologies of today (e.g. machine translation) facilitate a content production model in Wikipedia that is more “local” across language-defined cultures. That is, the absence of high-quality translation or similar capabilities has created an environment in which language-defined cultures can express their own perspective on world knowledge. If the technological context were to shift such that the linguistic equivalent of a “flat Earth” content production model is possible, the diverse language-defined cultural perspectives in Wikipedia would be put at risk.

New technologies that alter the content product model of a given UGC community are usually developed with a laudable objective in mind. For instance, Wikidata’s goal to increase information access for speakers of languages without high-quality Wikipedia language editions is very important. However, as we noted in Chapter 3, Wikidata will also likely dilute the cultural signal in the language editions of Wikipedia, despite its best efforts to avoid this outcome. The challenge for us as a computer science community is to continue to make major innovations in UGC-related technologies while maintaining, and indeed increasing, the ability for cultural communities to express their diverse perspectives about the world.

## 10 References

- [1] Adafre, S.F. and De Rijke, M. 2006. Finding Similar Sentences Across Multiple Languages in Wikipedia. (2006), 62–69.
- [2] Adar, E., Skinner, M. and Weld, D.S. 2009. Information Arbitrage Across Multi-lingual Wikipedia. *WSDM '09: Second ACM International Conference on Web Search and Data Mining* (Barcelona, Spain, 2009), 94–103.
- [3] Ahuja, A. and Downey, D. 2010. Improved extraction assessment through better language models. *NAACL-HLT 2010* (2010).
- [4] Alexa Top 500 Global Sites: <http://www.alexa.com/topsites>. Accessed: 2011-12-24.
- [5] Alpha Phi - Wikipedia, the free encyclopedia: [http://en.wikipedia.org/wiki/Alpha\\_Phi](http://en.wikipedia.org/wiki/Alpha_Phi). Accessed: 2013-01-16.
- [6] Aragon, P., Andreas, K., Laniado, D. and Volkovich, Y. 2012. Biographical Social Networks on Wikipedia. *WikiSym '12: 8th International Symposium on Wikis and Open Collaboration* (Linz, Austria, 2012).
- [7] Au Yeung, C., Duh, K. and Nagata, M. 2011. Providing Cross-Lingual Editing Assistance to Wikipedia Editors. *CICL '11: Computational Linguistics and Intelligent Text Processing* (Berlin, Heidelberg, 2011), 377–389.
- [8] Bailey, G., Hogan, B., Graham, M. and Mohammed, A.M. 2012. Interactive Mapping of Wikipedia's Geographies: Visualizing Variation in Participation and Representation. *WikiSym '12: 8th International Symposium on Wikis and Open Collaboration* (Linz, Austria, 2012).
- [9] Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M. and Gergle, D. 2012. Omnipedia: Bridging the Wikipedia Language Gap. *CHI '12: 30th International Conference on Human Factors in Computing Systems* (2012).
- [10] Baxter, R.N. 2009. New technologies and terminological pressure in lesser-used languages: The Breton Wikipedia, from terminology consumer to potential terminology provider. *LPLP*. 33, 1 (2009), 60–80.
- [11] Bergstrom, T. and Karahalios, K. 2009. Conversation clusters: grouping conversation topics through human-computer dialog. *CHI '09: 27th International Conference on Human Factors in Computing Systems* (Boston, MA, 2009), 2349–2352.
- [12] Bertin, J. and Berg, W.J. 1989. *Semiology of Graphics*. University of Wisconsin Press.
- [13] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R. and Hellman, S. 2009. DBpedia - A Crystallization Point for the Web of Data. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*. 7 (2009), 154–165.
- [14] Bollen, J., Mao, H. and Zeng, X. 2011. Twitter mood predicts the stock market. *Journal of*

- Computational Science.* 2, 1 (Mar. 2011), 1–8.
- [15] Borg, E. 2003. Discourse community. *ELT Journal.* 4 (2003), 398–400.
  - [16] Brewer, C. 1994. Color use guidelines for mapping and visualization. *Visualization in modern cartography.* A.M. MacEachren and Taylor, D.R.F., eds.
  - [17] Brin, S. and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. *WWW '98: Seventh International World Wide Web Conference* (Apr. 1998), 107–117.
  - [18] Budanitsky, A. and Hirst, G. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics.* 32, 1 (2006), 13–47.
  - [19] Callahan, E.S. and Herring, S.C. 2011. Cultural bias in Wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology.* 62, 10 (2011).
  - [20] Card, S.K., Mackinlay, J.D. and Shneiderman, B. 1999. *Readings in Information Visualization: Using Vision to Think.* Morgan Kaufmann.
  - [21] Chaey, C. 2012. 4 Ways The New Wikidata Will Improve Wikipedia. *Fast Company.*
  - [22] Cheng, Z., Caverlee, J. and Lee, K. 2010. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. *CIKM '10: 19th ACM International Conference on Information and Knowledge Management* (Toronto, Canada, 2010).
  - [23] Christaller, W. 1993. *Die zentralen Orte in Suddeutschland.* Gustav Fischer.
  - [24] Clark, H.H. 1996. *Using Language.* Cambridge University Press.
  - [25] Cohen, N. 2011. Define Gender Gap? Look Up Wikipedia's Contributor List. *The New York Times.* Accessed: 2012-05-26.
  - [26] Cohen, P.R. and Feigenbaum, E.A. 1982. The handbook of artificial intelligence. (1982).
  - [27] Crandall, D.J., Backstrom, L., Huttenlocher, D. and Kleinberg, J. 2009. Mapping the World's Photos. *WWW '09: 2009 International World Wide Web Conference* (Madrid, Spain, 2009), 761–770.
  - [28] CSISS Classics - Walter Christaller: Hierarchical Patterns of Urbanization: 2001. <http://www.csiss.org/classics/content/67>. Accessed: 2013-01-23.
  - [29] Dandala, B., Mihalcea, R. and Bunescu, R. 2012. Towards Building a Multilingual Semantic Network: Identifying Interlingual Links in Wikipedia. (Montreal, Quebec, Canada, 2012), 30–37.
  - [30] Dhar, V. and Chang, E.A. 2009. Does Chatter Matter? The Impact of User-Generated Content on Music Sales. *Journal of Interactive Marketing.* 23, 4 (Nov. 2009), 300–307.
  - [31] Van Dijk, Z. 2009. Wikipedia and lesser-resourced languages. *Language Problems & Language Planning.* 33, 3 (2009), 234–250.

- [32] Dodge, M., Kitchin, R. and Perkins, C. 2011. Introduction to The Map Reader. *The Map Reader: Theories of Mapping Practice and Cartographic Representation*. Wiley.
- [33] Dong, W. and Fu, W.-T. 2010. Cultural difference in image tagging. *CHI '10: 28th International Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA, 2010), 981.
- [34] Dourish, P. 2006. Re-space-ing place: place and space ten years on. *CSCW '06: 2006 ACM Conference on Computer Supported Cooperative Work* (Banff, Alberta, Canada, 2006), 299–308.
- [35] Duolingo: <http://duolingo.com/>. Accessed: 2011-09-13.
- [36] Dwight D. Eisenhower - Wikipedia, the free encyclopedia: <http://en.wikipedia.org/wiki/Eisenhower>. Accessed: 2013-01-26.
- [37] Eisenstein, J., O'Connor, B., Smith, N.A. and Xing, Eric P. 2010. A Latent Variable Model for Geographic Lexical Variation. *EMNLP '10: 2010 Conference on Empirical Methods in Natural Language Processing* (Boston, Massachusetts, United States, 2010), 1277–1287.
- [38] Erdmann, M., Nakayama, K., Hara, T. and Nishio, S. 2008. A bilingual dictionary extracted from the Wikipedia link structure. *Database Systems for Advanced Applications* (2008), 686–689.
- [39] Etzioni, O., Banko, M., Soderland, S. and Weld, D.S. 2008. Open information extraction from the web. *Communications of the ACM*. 51, 12 (Dec. 2008), 68–74.
- [40] Fellmann, J.D., Getis, A. and Getis, J. 2007. *Human Geography: Landscapes of Human Activity*. McGraw-Hill.
- [41] Filatova 2009. Directions for Exploiting Asymmetries in Multilingual Wikipedia. *CLIAWS3 '09: Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies* (Denver, Colorado, 2009).
- [42] Filatova, E. 2009. Multilingual Wikipedia, Summarization, and Information Trustworthiness. *CLIR 2009 : SIGIR 2009 Workshop on Information Access in a Multilingual World* (Stroudsburg, PA, USA, 2009).
- [43] Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. and Ruppin, E. 2002. Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems*. 20, 1 (2002), 116–131.
- [44] Fonstad, M., Pugatch, W. and Vogt, B. 2003. Kansas Is Flatter Than a Pancake. *Annals of Improbable Research*. (2003).
- [45] Franklin, C. and Hane, P. 1992. An introduction to GIS: linking maps to databases. *Database*. 15, 2 (Apr. 1992), 17–22.
- [46] Fussell, S.R. and Krauss, R.M. 1992. Coordination of Knowledge in Communication: Effects of Speakers' Assumptions About What Others Know. *Journal of Personality and*

- Social Psychology.* 62, 3 (1992), 378–391.
- [47] Gabrilovich, E. and Markovitch, S. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *IJCAI '07: Twentieth Joint Conference for Artificial Intelligence* (Hyderabad, India, 2007).
  - [48] Gabrilovich, E. and Markovitch, S. 2005. Feature Generation for Text Categorization Using World Knowledge. *IJCAI '05: 19th International Joint Conference on Artificial Intelligence* (Edinburgh, Scotland, 2005).
  - [49] Gabrilovich, E. and Markovitch, S. 2006. Overcoming the brittleness bottleneck using wikipedia: enhancing text categorization with encyclopedic knowledge. *AAAI '06: The Twenty-First National Conference on Artificial Intelligence* (2006), 1301–1306.
  - [50] Gabrilovich, E. and Markovitch, S. 2009. Wikipedia-based Semantic Interpretation for Natural Language Processing. *Journal of Artificial Intelligence Research.* 34, (2009), 443–498.
  - [51] Gärdenfors, P. 2000. *Conceptual Spaces: The Geometry of Thought.* The MIT Press.
  - [52] Gentzkow, M. and Shapiro, J.M. 2010. What Drives Media Slant? Evidence from U.S. Daily Newspapers. *Econometrica.* 78, 1 (2010), 35–71.
  - [53] German Community:  
[http://www.utas.edu.au/library/companion\\_to\\_tasmanian\\_history/G/German%20community.htm](http://www.utas.edu.au/library/companion_to_tasmanian_history/G/German%20community.htm). Accessed: 2013-01-13.
  - [54] Glott, R., Schmidt, P. and Ghosh, R. 2010. *Wikipedia Survey - Overview of Results.* Collaborative Creative Group, UNU-MERIT.
  - [55] Goddard and Wierzbicka eds. 1994. *Semantic and Lexical Universals.* John Benjamins Publishing Co.
  - [56] Goodchild, M. 2001. A Geographer Looks at Spatial Information Theory. *COSIT '01: 5th International Conference on Spatial Information Theory* (Morro Bay, CA, 2001).
  - [57] Goodchild, M.F. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal.* 69, 4 (Nov. 2007), 211–221.
  - [58] Goodchild, M.F., Yuan, M. and Cova, T.J. 2007. Towards a general theory of geographic representation in GIS. *International Journal of Geographical Information Science.* 21, 3 (2007), 239–260.
  - [59] Goodwin, D. 2012. Bing, Not Google, Favors Wikipedia More Often in Search Results [Study]. *Search Engine Watch.*
  - [60] Google Correlate: <http://correlate.googlelabs.com/>. Accessed: 2011-07-28.
  - [61] Gould, P. 1965. On mental maps. *Michigan Inter-university Community of Mathematical Geographers* (Ann Arbor, Michigan, 1965).

- [62] Gould, P. and White, R. 1986. *Mental Maps*. Psychology Press.
- [63] Graham, M. 2012. The Problem With Wikidata. *The Atlantic*.
- [64] Graham, M. 2012. Zero Geography: comparing the geographies of Arabic, Hebrew, and Persian Wikipedias. *Zero Geography*. Accessed: 2012-09-05.
- [65] Graham, M. 2012. Zero Geography: More Digital Divisions of Labour: a comparison of English and Arabic Wikipedias. *Zero Geography*. Accessed: 2012-09-05.
- [66] Greenstein, S. and Zhu, F. Is Wikipedia Biased? *American Economic Review*. 102, 3, 343–348.
- [67] Halavais, A. and Lackaff, D. 2008. An Analysis of Topical Coverage of Wikipedia. *Journal of Computer-Mediated Communication*. 13, (2008), 429–440.
- [68] Hale, S.A. 2012. Net Increase? Cross-Lingual Linking in the Blogosphere. *Journal of Computer-Mediated Communication*. 17, 1 (2012), 135–151.
- [69] Hara, N., Shachaf, P. and Hew, K.F. 2010. Cross-cultural analysis of the Wikipedia community. *Journal of the American Society of Information Science and Technology (JASIST)*. 61, 10 (2010), 2097-2108.
- [70] Hardy, D., Frew, J. and Goodchild, M.F. 2012. Volunteered geographic information production as a spatial process. *International Journal of Geographical Information Science*. (2012), 1–22.
- [71] Hargittai, E. and Litt, E. 2011. The tweet smell of celebrity success: Explaining variation in Twitter adoption among a diverse group of young adults. *New Media & Society*. (May. 2011).
- [72] Harkin, B. 2012. Is baseball turning into Latin America's game? - Baseball- NBC Sports. Accessed: 2013-01-22.
- [73] Harvey, D. 1991. *The Condition of Postmodernity: An Enquiry into the Origins of Cultural Change*. Wiley-Blackwell.
- [74] Hassan, S. and Mihalcea, R. 2009. Cross-lingual semantic relatedness using encyclopedic knowledge. *EMNLP '09: 2009 Conference on Empirical Methods in Natural Language Processing* (Singapore, 2009), 1192–1201.
- [75] Hays, J. and Efros, A.A. 2008. IM2GPS: estimating geographic information from a single image. *CVPR '08: IEEE Conference on Computer Vision and Pattern Recognition* (Jun. 2008), 1–8.
- [76] Hearst, M.A. 2009. *Search User Interfaces*. Cambridge University Press.
- [77] Hecht, B. 2007. *Utilizing Wikipedia as a Spatiotemporal Knowledge Repository*. University of California, Santa Barbara.
- [78] Hecht, B., Carton, S., Quaderi, M., Schöning, J., Raubal, M., Gergle, D. and Downey, D.

2012. Explanatory Semantic Relatedness and Explicit Spatialization for Exploratory Search. *SIGIR '12* (Portland, OR, 2012).
- [79] Hecht, B. and Gergle, D. 2011. A Beginner's Guide to Geographic Virtual Communities Research. *Handbook of Research on Methods and Techniques for Studying Virtual Communities: Paradigms and Phenomena*. IGI Global. 333–347.
- [80] Hecht, B. and Gergle, D. 2009. Measuring Self-Focus Bias in Community-Maintained Knowledge Repositories. *Communities and Technologies 2009: 4th International Conference on Communities and Technologies* (State College, PA, 2009), 11–19.
- [81] Hecht, B. and Gergle, D. 2010. On The “Localness” of User-Generated Content. *CSCW '10: 2010 ACM Conference on Computer Supported Cooperative Work* (Savannah, GA, 2010), 229–232.
- [82] Hecht, B. and Gergle, D. 2010. The Tower of Babel Meets Web 2.0: User-Generated Content and Its Applications in a Multilingual Context. *CHI '10: 28th International Conference on Human Factors in Computing Systems* (Atlanta, GA, 2010), 291–300.
- [83] Hecht, B., Hong, L., Suh, B. and Chi, E.H. 2011. Tweets from Justin Bieber's Heart: The Dynamics of the “Location” Field in User Profiles. *CHI '11: 29th ACM Conference on Human Factors in Computing Systems* (Vancouver, B.C., Canada, 2011), 237–246.
- [84] Hecht, B. and Moxley, E. 2009. Terabytes of Tobler: Evaluating the First Law in a Massive, Domain-Neutral Representation of World Knowledge. *COSIT '09: Ninth International Conference on Spatial Information Theory* (2009), 88–105.
- [85] Hecht, B. and Raubal, M. 2008. GeoSR: Geographically explore semantic relations in world knowledge. *AGILE '08: Eleventh AGILE International Conference on Geographic Information Science* (Girona, Spain, 2008), 95–114.
- [86] Hecht, B., Rohs, M., Schöning, J. and Krüger, A. 2007. WikEye: Using Magic Lenses to Explore Georeferenced Wikipedia Content. *PERMID '07: 3rd International Workshop on Pervasive Mobile Interaction Devices* (Toronto, Canada, 2007).
- [87] Hecht, B. and Schöning, J. 2008. Mapping the Zeitgeist. *GIScience '08: 5th International Conference on Geographic Information Science (Extended Abstracts)* (Park City, UT, 2008).
- [88] Hecht, B., Starosielski, N. and Dara-Abrams, D. 2007. Generating Educational Tourism Narratives from Wikipedia. *AAAI-INT '07: Association for the Advancement of Artificial Intelligence Fall Symposium on Intelligent Narrative Technologies* (Arlington, VA, 2007), 37–44.
- [89] Hecht, B., Teevan, J., Morris, M.R. and Liebling, D. 2012. SearchBuddies: Bringing Search Engines into the Conversation. *CHI '12: 30th International Conference on Human Factors in Computing Systems* (2012).
- [90] Help:Wiki markup - Wikipedia, the free encyclopedia:

- [http://en.wikipedia.org/wiki/Help:Wiki\\_markup](http://en.wikipedia.org/wiki/Help:Wiki_markup). Accessed: 2013-01-28.
- [91] Herring, S.C., Paolillo, J.C., Ramos-Velba, I., Kouper, I., Wright, E., Stoerger, S., Scheidt, L.A. and Clark, B. 2007. Language Networks on LiveJournal. *HICSS '07: 41st Hawaii International Conference on System Sciences* (Los Alamitos, CA, USA, 2007).
  - [92] Hoffart, J., Suchanek, F.M., Berberich, K. and Weikum, G. 2012. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. *Artificial Intelligence Journal*. (2012).
  - [93] Hofstede, G. 2002. Dimensions do not exist - a reply to Brendan McSweeney. *Human Relations*. 55, 11 (2002), 1355–1361.
  - [94] Hofstede, G., Hofstede, G.J. and Minkov, M. 2010. *Cultures and Organizations: Software for the Mind, Third Edition*. McGraw-Hill.
  - [95] Hong, L., Convertino, G. and Chi, E.H. 2011. Language Matters in Twitter: A Large Scale Study. *ICWSM '11: 5th International AAAI Conference on Weblogs and Social Media* (Barcelona, Spain, 2011).
  - [96] Hornbæk, K. and Frøkjær, E. 1999. Do Thematic Maps Improve Information Retrieval? *INTERACT '99* (1999), 1–8.
  - [97] Janowicz, K. 2012. Place and Location on the Web of Linked Data | STKO Lab. *Place and Location on the Web of Linked Data*. Accessed: 2013-01-24.
  - [98] Joachims, T. 2006. Training Linear SVMs in Linear Time. *KDD '06: 12th ACM Conference on Knowledge Discovery and Data Mining* (Philadelphia, PA, 2006).
  - [99] Joachims, T., Cristianini, N. and Shawe-Taylor, J. 2001. Composite kernels for hypertext categorisation. *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-* (2001), 250–257.
  - [100] Jones, C.B., Purves, R.S., Clough, P.D. and Joho, H. 2008. Modelling Vague Places with Knowledge from the Web. *International Journal of Geographical Information Science*. 22, 10 (2008), 1045–1065.
  - [101] Kennedy, L., Naaman, M., Ahern, S., Nair, R. and Rattenbury, T. 2007. How Flickr Helps us Make Sense of the World: Context and Content in Community-Contributed Media Collections. *ACM MM'08: Fifteenth ACM International Conference on Multimedia* (Augsburg, Germany, 2007).
  - [102] Kingsbury, D. 1968. *Manipulating the amount of information obtained from a person giving directions*. Harvard University.
  - [103] Kittur, A., Chi, E.H. and Suh, B. 2009. What's in Wikipedia? Mapping Topics and Conflict Using Socially Annotated Category Structures. *CHI '09: 27th International Conference on Human Factors in Computing Systems* (2009).
  - [104] Kittur, A. and Kraut, R.E. 2008. Harnessing the wisdom of crowds in wikipedia: quality

- through coordination. *Proceedings of the 2008 ACM conference on Computer supported cooperative work* (New York, NY, USA, 2008), 37–46.
- [105] Kleiman, M. 2013. *Web*.
  - [106] Knowledge – Inside Search – Google: <http://www.google.com/insidesearch/features/search/knowledge.html>. Accessed: 2012-09-11.
  - [107] Kramer, A.D.I. 2010. An unobtrusive behavioral model of gross national happiness. *CHI '10: 28th ACM Conference on Human Factors in Computing Systems* (2010), 287–290.
  - [108] Kramsch, C. 1998. *Language and Culture*. Oxford University Press.
  - [109] Krumm, J., Davies, N. and Narayanaswami, C. 2008. User-Generated Content. *Pervasive Computing, IEEE*. 7, 4 (Dec. 2008), 10–11.
  - [110] Ladd, F. 1967. A note on “the world across the street”. *Harvard Graduate School Education Association Bulletin*. 12, (1967), 47–49.
  - [111] Lam, S.K., Uduwage, A., Dong, Z., Sen, S., Musicant, D.R., Terveen, L. and Riedl, J. 2011. WP:Clubhouse? An Exploration of Wikipedia’s Gender Imbalance. *WikiSym '11: 7th International Symposium on Wikis and Open Collaboration* (Mountain View, CA, 2011), 1–10.
  - [112] Lanegran, D.A. and Natoli, S. 1984. *Guidelines for Geographic Education in the Elementary and Secondary Schools*. Association of American Geographers.
  - [113] Law, M.R., Mintzes, B. and Morgan, S.G. 2011. The Sources and Popularity of Online Drug Information: An Analysis of Top Search Engine Results and Web Page Views. *The Annals of Pharmacotherapy*. 45, 3 (Mar. 2011), 350–356.
  - [114] Lenat, D.B., Guha, R.V., Pittman, K., Pratt, D. and Shepherd, M. 1990. Cyc: toward programs with common sense. *Communications of the ACM*. 33, 8 (1990), 30–49.
  - [115] Lenat, D.B., Prakash, M. and Shepherd, M. 1985. CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *AI magazine*. 6, 4 (1985), 65.
  - [116] Lewis, M.P. ed. 2009. *Ethnologue: Languages of the World, 16th Edition*. SIL International.
  - [117] Liao, H.-T. 2009. Conflict and Consensus in the Chinese Version of Wikipedia. *IEEE Technology and Society Magazine*. Summer 2009 (2009), 49–56.
  - [118] Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P. and Tomkins, A. 2005. Geographic routing in social networks. *Proceedings of the National Academy of Sciences*. 102, 33 (2005), 11623–11628.
  - [119] Lieberman, M.D. and Lin, J. 2009. You Are Where You Edit: Locating Wikipedia Users

- Through Edit Histories. *ICWSM '09: 3rd Int'l AAAI Conference on Weblogs and Social Media* (2009).
- [120] Lih, A. 2009. *The Wikipedia Revolution: How a Bunch of Nobodies Created the World's Greatest Encyclopedia*. Hyperion.
- [121] Lindamood, J., Heatherly, R., Kantacioglu, M. and Thuraisingham, B. 2009. Inferring Private Information Using Social Network Data. *WWW '09: 2009 International World Wide Web Conference* (Madrid, 2009).
- [122] List of official languages - Wikipedia, the free encyclopedia: [http://en.wikipedia.org/wiki/List\\_of\\_official\\_languages#cite\\_note-6](http://en.wikipedia.org/wiki/List_of_official_languages#cite_note-6). Accessed: 2013-01-25.
- [123] Longley, P.A., Goodchild, M., Maguire, D.J. and Rhind, D.W. 2010. Cartography and Map Production. *Geographic Information Systems and Science*. Wiley.
- [124] MacEachren, A.M. 1982. The Role of Complexity and Symbolization Method in Thematic Maps. *Annals of the Association of American Geographers*. 72, 4 (1982), 495–513.
- [125] Magnini, B. and Cavaglià, G. 2000. Integrating Subject Field Codes into WordNet. *LREC '00: 2nd International Conference on Language Resources and Evaluation* (Athens, Greece, 2000), 1413–1418.
- [126] Mahmud, J., Nichols, J. and Drews, C. 2012. Where Is This Tweet From? Inferring Home Locations of Twitter Users. *ICWSM '12: Sixth International AAAI Conference on Weblogs and Social Media* (Dublin, Ireland, 2012).
- [127] Massa, P. and Scrinzi, F. 2012. Manypedia: Comparing Language Points of View of Wikipedia Communities. *WikiSym '12: 8th International Symposium on Wikis and Open Collaboration* (Linz, Austria, 2012).
- [128] Matei, S., Ball-Rokeach, S.J. and Qiu, J.L. 2001. Fear and Misperception of Los Angeles Urban Space A Spatial-Statistical Study of Communication-Shaped Mental Maps. *Communication Research*. 28, 4 (Aug. 2001), 429–463.
- [129] McCallum, A. and Nigam, K. 1998. A Comparison of Event Models for Naive Bayes Text Classification. *AAAI-98 Workshop on Learning for Text Categorization* (1998), 41–48.
- [130] De Melo, G. and Weikum, G. 2010. Untangling the Cross-Lingual Link Structure of Wikipedia. *ACL '10: 48th Annual Meeting of the Association for Computational Linguistics* (Uppsala, Sweden, 2010).
- [131] Mihalcea, R. and Csoma, A. 2007. Wikify!: linking documents to encyclopedic knowledge. (2007), 233.
- [132] Miller, B.N., Albert, I., Lam, S.K., Konstan, J.A. and Riedl, J. 2003. MovieLens unplugged: experiences with an occasionally connected recommender system. *Proceedings of the 8th international conference on Intelligent user interfaces* (New York, NY, USA,

- 2003), 263–266.
- [133] Miller, G.A. and Charles, W.G. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*. 6, 1 (1991), 1–28.
- [134] Miller, G.A. and others 1995. WordNet: a lexical database for English. *Communications of the ACM*. 38, 11 (1995), 39–41.
- [135] Milne, D. and Witten, I. 2013. An open-source toolkit for mining Wikipedia. *Artificial Intelligence*. 194, (2013), 222–239.
- [136] Milne, D. and Witten, I.H. 2008. An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. *WIKIAI '08: First AAAI Workshop on Wikipedia and Artificial Intelligence* (Chicago, IL, 2008).
- [137] Milne, D. and Witten, I.H. 2008. Learning to link with wikipedia. *CIKM '08: 17th ACM Conference on Information and Knowledge Management* (Napa Valley, California, USA, 2008), 509–518.
- [138] Minno, D., Wallach, H.M., Naradowsky, J., Smith, D.A. and McCallum, A. 2009. Polylingual Topic Models. *EMNLP '09: 2009 Conference on Empirical Methods in Natural Language Processing* (Singapore, 2009), 880–889.
- [139] Monmonier, M. and Blij, H.J. de 1996. *How to Lie with Maps*. University Of Chicago Press.
- [140] Montello, D.R. 2002. Cognitive Map-Design Research in the Twentieth Century: Theoretical and Empirical Approaches. *Cartography and Geographic Information Science*. 29, 3 (2002), 283–304.
- [141] Muller, M.J. 2007. Comparing tagging vocabularies among four enterprise tag-based services. (2007), 341–350.
- [142] Nash, E. 1996. The rise and fall of Spain's Machiavelli. *The Independent*.
- [143] Navigli, R. and Ponzetto, S.P. 2012. BabelRelate! A Joint Multilingual Approach to Computing Semantic Relatedness. *AAAI '12: Twenty-Sixth AAAI Conference on Artificial Intelligence* (Toronto, Canada, 2012).
- [144] Nemoto, K. and Gloor, P.A. 2010. Analyzing Cultural Differences in Collaborative Innovation Networks by Analyzing Editing Behavior in Different-Language Wikipedias. *COINs 2010: Collaborative Innovations Networks Conference* (Savannah, GA, 2010).
- [145] O'Madadhain, J., Fisher, D., White, S., Smyth, P. and Boey, Y. 2005. Analysis and Visualization of Network Data using JUNG. 10, 2 (2005), 1–25.
- [146] Oh, J.-H., Kawahara, D., Uchimoto, K., Kazama, J. and Torisawa, K. 2008. Enriching Multilingual Language Resources by Discovering Missing Cross-Language Links in Wikipedia. *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on* (Los Alamitos, CA, USA, 2008), 322–328.

- [147] Orleans, P. 1967. Differential cognition of urban residents: effects of social scale on mapping. *National Academy of Engineering Publication 1498*.
- [148] Ortega, F., Gonzalez-Barahona, J.M. and Robles, G. 2008. On the Inequality of Contributions to Wikipedia. *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual* (Jan. 2008), 304.
- [149] Orwell, G. 1943. *As I Please*. Tribune.
- [150] Over, P. and Yen, J. 2004. *An Introduction to DUC-2004: Intrinsic Evaluation of Generic News Text Summarization Systems*. National Institute of Standards and Technology.
- [151] Panciera, K., Priedhorsky, R., Erickson, T. and Terveen, L. 2010. Lurking? cyclopaths?: a quantitative lifecycle analysis of user behavior in a geowiki. *CHI '10: 28th International Conference on Human Factors in Computing Systems* (New York, NY, USA, 2010), 1917–1926.
- [152] Pedersen, T., Pakhomov, S.V.S., Patwardhan, S. and Chute, C.G. 2006. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*. 40, 3 (2006), 288–299.
- [153] Pennebaker, J.W., Booth, R.J. and Francis, M.E. 2007. *Linguistic Inquiry and Word Count: LIWC2007*. University of Texas at Austin.
- [154] Perez, S. 2012. Wikipedia's Next Big Thing: Wikidata, A Machine-Readable, User-Editable Database Funded By Google, Paul Allen And Others. *TechCrunch*.
- [155] Pfeil, U., Zaphiris, P. and Ang, C.S. 2006. Cultural Differences in Collaborative Authoring of Wikipedia. *Journal of Computer-Mediated Communication*. 12, 1 (Oct. 2006), 88–113.
- [156] Ponzetto, S.P. and Strube, M. 2007. Knowledge Derived From Wikipedia For Computing Semantic Relatedness. *Journal of Artificial Intelligence Research*. 30, (2007), 181–212.
- [157] Popescu, A. and Grefenstette, G. 2010. Mining User Home Location and Gender from Flickr Tags. *ICSWM '10: 4th International AAAI Conference on Weblogs and Social Media* (2010).
- [158] Popescu, A. and Grefenstette, G. 2010. Spatiotemporal mapping of Wikipedia concepts. *JCDL '10* (2010), 129–138.
- [159] Potthast, M., Stein, B. and Anderka, M. 2008. A Wikipedia-based multilingual retrieval model. *Proceedings of the IR research, 30th European conference on Advances in information retrieval* (Glasgow, UK, 2008), 522–530.
- [160] Priedhorsky, R., Chen, J., Lam, S. (Tony) K., Panciera, K., Terveen, L. and Riedl, J. 2007. Creating, destroying, and restoring value in wikipedia. *Group '07: 2007 International ACM Conference on Supporting Group Work* (Sanibel Island, Florida, USA, 2007), 259–268.

- [161] Priedhorsky, R. and Terveen, L. 2008. The Computational Geowiki: What, Why, and How. *CSCW '08: 2008 ACM Conference on Computer Supported Cooperative Work* (San Diego, CA, 2008).
- [162] Putnam, R.D. 2001. *Bowling Alone: The Collapse and Revival of American Community*. Touchstone Books by Simon & Schuster.
- [163] Radinsky, K., Agichtein, E., Gabrilovich, E. and Markovitch, S. 2011. A Word at a Time: Computing Word Relatedness using Temporal Semantic Analysis. *WWW '11: 20th International Conference on World Wide Web* (Hyderabad, India, 2011), 337–346.
- [164] Ranking sports' popularity: And the silver goes to... | The Economist:  
<http://www.economist.com/blogs/gametheory/2011/09/ranking-sports%E2%80%99-popularity>. Accessed: 2013-01-16.
- [165] Ratinov, L., Roth, D., Downey, D. and Anderson, M. 2011. Local and Global Algorithms for Disambiguation to Wikipedia. *ACL '11: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (2011).
- [166] Rattenbury, T., Good, N. and Naaman, M. 2007. Towards automatic extraction of event and place semantics from flickr tags. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (2007), 103–110.
- [167] Reagle, J. and Rhue, L. 2011. Gender Bias in Wikipedia and Britannica. *International Journal of Communication*. 5, (2011), 1138–1158.
- [168] Rex, B., Risch, J., Dowson, S. and Moon, B. 1999. Multiple Source Information Analysis, GIS and Starlight. *ACM GIS '99* (Kansas City, MO, 1999).
- [169] RIAA - Top Selling Artists - January 13, 2013:  
[http://www.riaa.com/goldandplatinum.php?content\\_selector=top-selling-artists](http://www.riaa.com/goldandplatinum.php?content_selector=top-selling-artists). Accessed: 2013-01-13.
- [170] Ribé, M.M. and Rodríguez, H. 2011. Cultural Configuration of Wikipedia: Measuring Autoreferentiality in Different Languages. *RANLP '11: Recent Advances in Natural Language Processing* (Hissar, Bulgaria, 2011), 316–322.
- [171] Risch, J.S., Rex, D.B., Dowson, S.T., Walters, T.B., May, R.A. and Moon, B.D. 1997. The STARLIGHT information visualization system. *INFOVIS '97* (London, U.K., 1997), 42–49.
- [172] Rose, D.E. and Levinson, D. 2004. Understanding user goals in web search. *WWW '04: 13th International Conference on World Wide Web* (New York, NY, USA, 2004), 13–19.
- [173] Rosenzweig, R. 2006. Can History Be Open Source? Wikipedia and the Future of the Past. *The Journal of American History*. 93, 1 (2006), 117–146.
- [174] Sakaki, T., Okazaki, M. and Matsuo, Y. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. *WWW '10: 19th International Conference on World*

- Wide Web* (Raleigh, North Carolina, USA, 2010), 851–860.
- [175] Sapir, E. and Mandelbaum, D.G. 1949. *Culture, Language and Personality*. University of California Press.
- [176] Scellato, S., Noulas, A., Lambiotte, R. and Mascolo, C. 2011. Socio-spatial Properties of Online Location-based Social Networks. *ICWSM '11: 5th International AAAI Conference on Weblogs and Social Media* (Barcelona, Spain, 2011).
- [177] Schmitz, P. 2006. Inducing ontology from flickr tags. *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland* (2006).
- [178] Schonfeld, E. 2010. Costolo: Twitter Now Has 190 Million Users Tweeting 65 Million Times A Day. *TechCrunch*.
- [179] Schöning, J., Cheverst, K., Löchtefeld, M., Krüger, A., Rohs, M. and Taher, F. 2009. PhotoMap: Using Spontaneously taken Images of Public Maps for Pedestrian Navigation Tasks on Mobile Devices. *Mobile HCI '09: 11th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Bonn, Germany, 2009).
- [180] Schöning, J., Hecht, B., Raubal, M., Krüger, A., Marsh, M. and Rohs, M. 2008. Improving Interaction with Virtual Globes through Spatial Thinking: Helping Users Ask “Why?” *IUI '08: 2008 International Conference on Intelligent User Interfaces* (Maspalomas, Gran Canaria, Spain, 2008), 129–138.
- [181] Schöning, J., Hecht, B., Rohs, M. and Starosielski, N. 2007. WikEar - Automatically Generated Location-Based Audio Stories Between Public City Map. *Ubicomp '07 EA: 9th International Conference on Ubiquitous Computing (Extended Abstracts)* (Innsbruck, Austria, 2007).
- [182] Semiocast 2010. *Half of messages on Twitter are not in English Japanese is the second most used language*. Semiocast.
- [183] Sheesley, B. 2009. Data Probing and Info Window Design on Web-based Maps - Axis Maps Blog. *Axis Maps Blog*. Accessed: 2012-06-04.
- [184] Shneiderman, B. 1996. The eyes have it: a task by data type taxonomy for information visualizations. *IEEE Symposium on Visual Languages '96* (Sep. 1996), 336 –343.
- [185] Singhal, A. 2012. Introducing the Knowledge Graph: things, not strings. *Google: Official Blog*.
- [186] Skupin, A. and Fabrikant, S.I. 2003. Spatialization Methods: A Cartographic Research Agenda for Non-geographic Information Visualization. *Cartography and Geographic Information Science*. 30, 2 (2003), 95–115.
- [187] Slocum, T.A., McMaster, R.B., Kessler, F.C. and Howard, H.H. 2009. *Thematic Cartography and Geovisualization*. Prentice Hall.
- [188] Sorg, P. and Cimiano, P. 2008. Enriching the Crosslingual Link Structure of Wikipedia -

- A Classification-based Approach. *WIKI-AI '08: AAAI 2008 Workshop on Wikipedia and Artificial Intelligence* (Chicago, IL, 2008).
- [189] Spolsky, B. 1997. Multilingualism in Israel. *Annual Review of Applied Linguistics*. 17, (1997), 138–150.
  - [190] Starbird, K., Palen, L., Hughes, A.L. and Vieweg, S. 2010. Chatter on the red: what hazards threat reveals about the social life of microblogged information. (Savannah, GA, 2010), 241–250.
  - [191] Stephens, M. 2012. Church or Beer? Americans on Twitter. *floating sheep*. Accessed: 2013-01-16.
  - [192] Strube, M. and Ponzetto, S.P. 2006. WikiRelate! Computing Semantic Relatedness Using Wikipedia. *AAAI '06: The Twenty-First National Conference on Artificial Intelligence* (Boston, MA, 2006), 1419–1424.
  - [193] Stvilia, B., Al-Faraj, A. and Jeong, Y.Y. 2009. Issues of Cross-Contextual Information Quality Evaluation — The Case of Arabic, English, and Korean Wikipedias. *Library & Information Science Research*. 31, 4 (2009), 232–239.
  - [194] Suchanek, F.M., Kasneci, G. and Weikum, G. 2007. Yago: a core of semantic knowledge. *WWW '07: 16th International Conference on World Wide Web* (New York, NY, USA, 2007), 697–706.
  - [195] Taraborelli, D. and Wikimedia Research Newsletter Contributors 2012. Given enough eyeballs, do articles become neutral? *Wikimedia Research Newsletter*. 2, 2 (Feb. 2012).
  - [196] Thij, M. ten, Volkovich, Y., Laniado, D. and Kaltenbrunner, A. 2012. Modeling and predicting page-view dynamics on Wikipedia. *arXiv:1212.5943*. (Dec. 2012).
  - [197] Top Ten origins of Tasmanians: <http://www.tasmaniatoften.com/lists/ancestries.php>. Accessed: 2013-01-13.
  - [198] Translating the world's information with Google Translator Toolkit: 2009. <http://googleblog.blogspot.com/2009/06/translating-worlds-information-with.html>. Accessed: 2011-09-16.
  - [199] VanderWal, T. 2004. You Down with Folksonomy? *vanderwal.net*. Accessed: 2012-12-26.
  - [200] Vieweg, S., Hughes, A.L., Starbird, K. and Palen, L. 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. *CHI '10: 28th ACM Conference on Human Factors in Computing Systems* (Atlanta, GA, 2010), 1079–1088.
  - [201] Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H. and Studer, R. 2006. Semantic Wikipedia. *WWW '06: 15th International Conference on World Wide Web* (New York, NY, USA, 2006), 585–594.

- [202] Wade, T. and Sommer, S. 2006. Multipoint Feature. *A to Z GIS: An Illustrated Dictionary of Geographic Information Systems*. ESRI Press.
- [203] Warncke-Wang, M., Uduwage, A., Dong, Z. and Riedl, J. 2012. In Search of the Ur-Wikipedia: Universality, Similarity, and Translation in the Wikipedia Inter-language Link Network. *WikiSym '12: 8th International Symposium on Wikis and Open Collaboration* (2012).
- [204] Weld, D.S., Wu, F., Adar, E., Amershi, S., Fogarty, J., Hoffman, R., Patel, K. and Skinner, M. 2008. Intelligence in Wikipedia. *AAAI '08* (2008).
- [205] White, R., Muresan, G. and Marchionini, G. 2006. Evaluating Exploratory Search Systems. *SIGIR '06 Workshop on Evaluating Exploratory Search* (2006).
- [206] White, R., Roth, R. and Marchionini, G. 2009. *Exploratory search: beyond the query-response paradigm*. Morgan & Claypool,.
- [207] Wierzbicka, A. 1997. *Understanding Cultures through Their Key Words: English, Russian, Polish, German, and Japanese*. Oxford University Press, USA.
- [208] WikiBhasha beta – A multi-lingual content creator for Wikipedia: <http://www.wikibhasha.org/>. Accessed: 2011-09-16.
- [209] Wikimania 2012 tackles diversity issues: 2012. [http://en.wikinews.org/wiki/Wikimania\\_2012\\_tackles\\_diversity\\_issues](http://en.wikinews.org/wiki/Wikimania_2012_tackles_diversity_issues). Accessed: 2013-02-10.
- [210] Wikipedia Report Card: <http://reportcard.wmflabs.org/>. Accessed: 2013-01-17.
- [211] Wikipedia-world: [http://de.wikipedia.org/wiki/Wikipedia:WikiProjekt\\_Georeferenzierung/Hauptseite/Wikipedia-World/en#Project\\_description](http://de.wikipedia.org/wiki/Wikipedia:WikiProjekt_Georeferenzierung/Hauptseite/Wikipedia-World/en#Project_description). Accessed: 2012-09-11.
- [212] Wikipedia:WikiProjekt Georeferenzierung/Hauptseite/Wikipedia-World/en – Wikipedia: [http://de.wikipedia.org/wiki/Wikipedia:WikiProjekt\\_Georeferenzierung/Hauptseite/Wikipedia-World/en#Usage\\_of\\_the\\_data](http://de.wikipedia.org/wiki/Wikipedia:WikiProjekt_Georeferenzierung/Hauptseite/Wikipedia-World/en#Usage_of_the_data). Accessed: 2013-01-24.
- [213] Wing, B.P. and Baldridge, J. 2011. Simple Supervised Document Geolocation with Geodesic Grids. *ACL '11: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (Portland, OR, 2011), 955–964.
- [214] Wunsch-Vincent, S. and Vickery, G. 2007. *Participative Web: User-Created Content*. Technical Report #JT03225396. Organisation for Economic Co-operation and Development (OECD) Directorate for Science, Technology, and Industry; Working Party on the Information Economy.
- [215] Yasseri, T., Sumi, R. and Kertész, J. 2012. Circadian Patterns of Wikipedia Editorial Activity: A Demographic Analysis. *PLoS One*. 7, 1 (Jan. 2012), 1–8.
- [216] Yi, J., Kang, Y., Stasko, J. and Jacko, J. 2008. Understanding and characterizing insights:

how do people gain insights using information visualization? *Proceedings of the 2008 conference on BEyond time and errors: novel evaLuation methods for Information Visualization* (2008), 1–6.

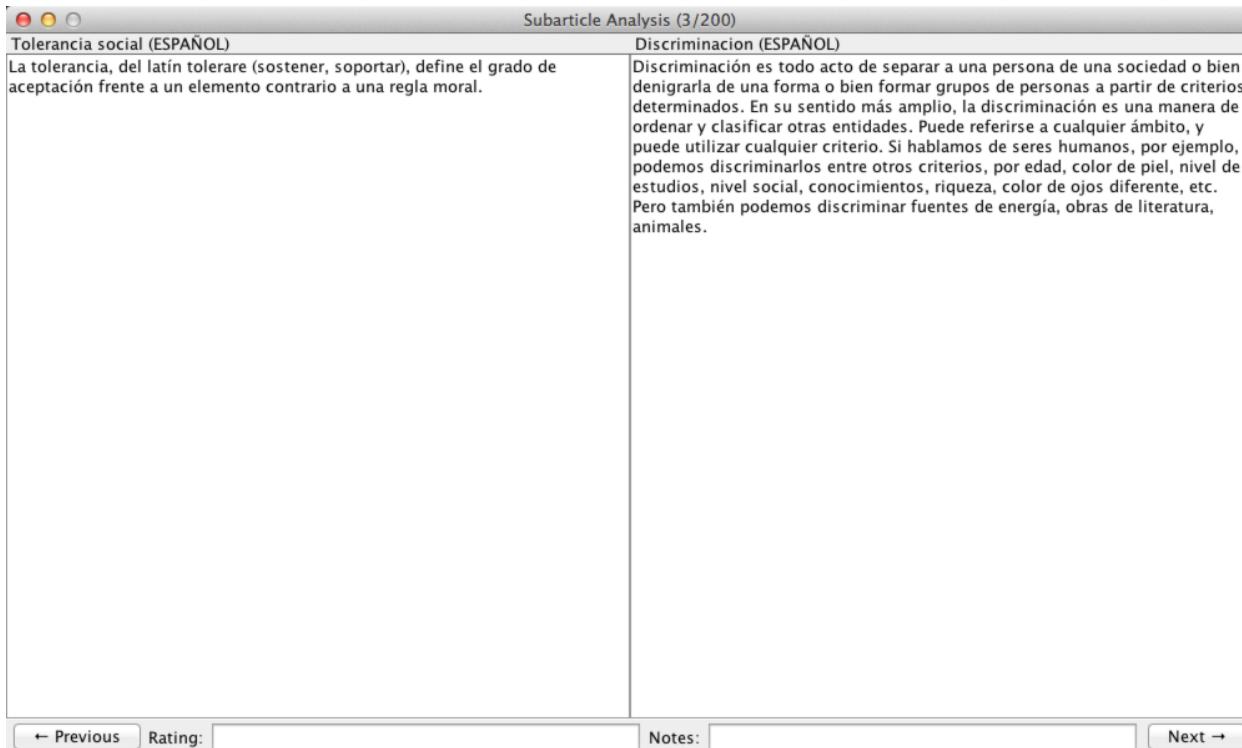
- [217] Zesch, T. and Gurevych, I. 2006. Automatically creating datasets for measures of semantic relatedness. *ACL-Workshop on Linguistic Distances* (Sydney, Australia, 2006), 16–24.
- [218] Zesch, T. and Gurevych, I. 2010. The More the Better? Assessing the Influence of Wikipedia’s Growth on Semantic Relatedness Measures. *LREC ’10: Seventh Conference on International Language Resources and Evaluation* (Valletta, Malta, 2010).
- [219] Zesch, T. and Gurevych, I. 2009. Wisdom of crowds versus wisdom of linguists – measuring the semantic relatedness of words. *Natural Language Engineering*. 16, 1 (2009), 25–59.
- [220] Zesch, T., Müller, C. and Gurevych, I. 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. *LREC ’08: 6th International conference on Language Resources and Evaluation* (Marrakech, Morocco, 2008).
- [221] Zlatić, V., Božićević, M., Štefančić, H. and Domazet, M. 2006. Wikipedias: Collaborative web-based encyclopedias as complex networks. *arXiv:physics*. (Jul. 2006).
- [222] 2012. Anchor text. *Wikipedia, the free encyclopedia*. Accessed: 2012-12-26.
- [223] 2010. Cuba launches its own Wikipedia. *BBC*. Accessed: 2012-09-05.
- [224] 2012. *English Proficiency Index*. Education First.
- [225] *Google Earth*. Google.
- [226] 2012. Help:Template. *Wikipedia, the free encyclopedia*. Accessed: 2012-12-31.
- [227] 2012. User-generated content. *Wikipedia, the free encyclopedia*. Accessed: 2012-05-26.
- [228] 2012. Wikipedia:Disambiguation. *Wikipedia, the free encyclopedia*. Accessed: 2012-12-26.
- [229] Wikipedia:Featured articles. *Wikipedia, the free encyclopedia*. Accessed: 2013-03-01.
- [230] 2012. Wikipedia:Systemic bias. *Wikipedia, the free encyclopedia*. Accessed: 2012-09-14.

# 11 Appendices

## A Randomly Selected Single-language and Global Concepts

Global Concepts (Randomly Selected)	Single-language Concepts (Randomly Selected)
25-Aug	Baia di Moore (it)
1057	Greg Marsden (en)
Roman mythology	MIND ASSASSIN (ja)
Nu metal	Beneficiarius (it)
Eurasia	头柱灯心草 (zh)
Zadar	IRE Legion (it)
MŠK Žilina	(כָּפַל) (he)
Medea	Шпак Федір (uk)
Umar	Sinseol-dong (en)
Homeostasis	Sickenhammer (de)
Allier	IC 4003 (uk)
2009	PhoneFactor (en)
Cicero	South Ice (en)
Bromine	Shkélzen (en)
455	Basil Hoffmann (it)
Solipsism	Te Rangiuamutu (en)
9th century BC	絶体延命 (ja)
Baffin Island	菊地豊 (ja)
Seismometer	Nib (pen) (en)
Syncretism	시사매거진 2580 (ko)
Ageing	Martin Caroe (en)
Frederick Soddy	Business case (fr)
Sensor	Lena Lassen (fr)
First language	Vöröses ösmoly (hu)
Antonio Puerta	Zezschwitz (de)
System of a Down	Oskar Lindberg (en)
1463	Tomàs Carnicer (ca)
Wolfgang Pauli	Bachy Ferenc (hu)
Assassination of John F.	Ellen Karcher (en)
Kennedy	Jim Colbert (en)
Parma	Bode Sowande (en)
Positron	西港鎮 (秦皇岛市) (zh)
12-May	新興県 (曖昧さ回避) (ja)
Fernando Verdasco	Dope Nose (it)
818	Don Abi (de)
Jon Bon Jovi	Ножига (ru)
Banjul	Supa Dam (en)
Caracalla	Schweißhände (de)
214	Bob Meinke (en)
Strait of Gibraltar	Rulers.org (en)
Mensa International	종이팩 (ko)
1733	Georg Kotowski (de)
19-May	ScerTF (en)
Tertullian	Дронкерс, Бен (ru)
C. S. Lewis	Guido Convents (fr)
1367	Paige Hemmis (pl)
	Felipe Torres (pt)
	Militia Dei (ru)
	Tommy Lewis (en)
	George Brann (en)
	Inédit (fr)
	Ryan Cutrona (en)
	Absol (nl)
	ACARM-ng (en)
	François-Louis (en)
	Muljinapura (ca)
	WNGH (en)
	マセラティ・228 (ja)
	Peter Sheridan (it)
	道摩法師 (ja)
	Colonial Dubs (en)
	Telemark Høyre (no)
	暢通運輸 (zh)
	Stick mantis (en)
	Paw Madsen (en)
	Наушера (ru)
	Кочар (ru)
	Evert Nilsson (sv)
	Kunståret 1364 (no)
	Cây Trưởng II (nl)
	天使の殺意 (ja)
	力順 (zh)
	Jan Ceuleers (nl)
	Lilo Friedrich (de)
	見て (ja)
	Jim Bianco (en)
	Rex de Rox (sv)
	難病対策 (ja)
	Andrew Stark (en)
	頭十位在位時間最短的教宗列表 (zh)
	○○都民 (ja)
	1982年の自転車競技 (ja)
	Jam (rivista) (it)
	中山大学附属中学 (zh)
	星屑ファンタジー (ja)
	Bokota people (en)
	William Pike (en)
	스카이 이자르 (ko)
	Калинович (uk)
	Walter Odede (pl)
	Ivano Bellodis (it)

## B GUI Application Used for Sub-article Coding



## C Sub-article Coding Guidelines

### INTRODUCTION

Thanks for agreeing to help us with our Wikipedia study! The goal of our study is to determine which articles are **sub-articles** of other articles.

According to the English Wikipedia, Wikipedia articles are “individual page[s] that display information on a topic.” However, for some topics there is too much information to put on a single article. In these cases, information is split off into “sub-articles” For example, the “Northwestern University” **main article** in the English Wikipedia has sub-articles that include “History of Northwestern University”, “Alumni of Northwestern University”, “Northwestern University buildings”, and so on.

While it is easy for a human to understand which articles are sub-articles of other articles, this is not the case for computer algorithms. Indeed, one of the five fundamental principles in Wikipedia is that “Wikipedia does not have firm rules”. Computers tend to need “firm rules” to very easily understand things.

Your goal is to use a provided application to manually indicate which articles are sub-articles of a given main article. From this data, we will teach a “machine learning” algorithm to recognize the patterns in your manual annotation. This will help our software understand which articles are sub-articles of a given main article.

## RATING SCHEMA

You will be asked to rate a series of *potential* sub-article/parent article relationships according to the following schema:

- **3:** The *only* reason the potential sub-article exists is to split the corresponding main article into more manageable subtopics. The potential sub-article really does *not deserve its own page*, and the corresponding main article is the best place to put the sub-article’s content.
- **2:** Same as above, but the potential sub-article’s topic is significant enough to warrant its own page.
- **1:** The potential sub-article contains information that would be useful to have on the main article, but contains its own, *unrelated (non-overlapping) content*.
- **0:** The potential sub-article is on a topic that is **trivially related** to the main article or has a large amount of **non-overlapping content**.

Below are some examples for each rating. Remember that **there is no correct answer** for any given potential main article/sub-article pair. We want your best guess for each potential main article/sub-article pair.

### Examples of “3”

- Main article: Caffeine
  - Health effects of caffeine
- Main article: Northwestern University
  - List of Northwestern University alumni
  - List of Northwestern University buildings
  - List of Northwestern University faculty
  - History of Northwestern University
- Britney Spears
  - Britney Spears (Discography)

### Examples of “2”

- Main article: Caffeine
  - Decaffeination
- Main article: War
  - Preventive war
  - Just war theory
  - Casus belli

- Civilian causalities
- Civil War
- List of ongoing military conflicts
- Main article: Thanksgiving
  - Black Friday
  - Thanksgiving (USA)
  - Thanksgiving (Canada)
  - Labor Thanksgiving Day

### **Examples of “1”**

- Main Article: Caffeine
  - History of coffee
  - Effects of psychoactive drugs on animals
  - Efectos del café en la salud
  - HIstory of chocolate
  - History of tea
  - Coffee
  - Tea
- Main Article: War
  - Alliance
  - Geneva Conventions
  - Military strategy
  - Military Keynesianism
  - Social conflict
- Main Article: Thanksgiving
  - Harvest festival

### **Examples of “0”**

- Caffeine
  - Koffein (film)
- War
  - International Physicians for the Prevention of Nuclear War
  - United Nations
- Thanksgiving
  - Colonial history of the United States
- Britney Spears
  - List of Most Expensive Music Videos

## D *RatioInRandom* Results Using All Links

<i>RatioInRandom</i> using All Links				
<i>Wikification Strategy</i>	<i>Mean</i>	% <i>RatioInRandom</i> = 1	<i>Mean</i> only-intersection	% <i>RatioInRandom</i> = 1 only-intersection
“Kitchen Sink” Upper-Bound <WikipediaTitle+Redirect+AnchorText, GoogleTranslateTitle+Redirect>	0.671	9.72%	0.740	15.5%
Moderate <WikipediaTitle+Redirect, GoogleTranslateNone>	0.588	4.17%	0.648	6.70%
“Just Links” Lower-Bound <WikipediaTitleNone, GoogleTranslateNone>	0.428	0.86%	0.493	1.72%

Table 11-a: These results are equivalent to those in Table 3.5-e, expect the considered BOLs included all links, not just parseable ones.

## E Overlap Coefficient Using All Links

Overlap Coefficient using All Links				
<i>Wikification Strategy</i>	<i>Mean OC</i>	% <i>OC</i> = 1	<i>Mean OC</i> only-intersection	% <i>OC</i> = 1 only-intersection
“Kitchen Sink” Upper-Bound <WikipediaTitle+Redirect+AnchorText, GoogleTranslateTitle+Redirect>	0.865	17.9%	0.905	29.1%
Moderate <WikipediaTitle+Redirect, GoogleTranslateNone>	0.781	8.20%	0.824	13.7%
“Just Links” Lower-Bound <WikipediaTitleNone, GoogleTranslateNone>	0.567	1.95%	0.640	4.20%

Table 11-b: These results are equivalent to those in Table 3.6-a, expect the considered BOLs included all links, not just parseable ones.

## F Intersection and xOR of English and German Concepts with the 100 Highest PageRank Scores

The intersection and exclusive-or of the sets of concepts in the German and English Wikipedias with the top 100 PageRank scores. The left-most column depicts the intersection and

the zero-based rank (0-99) of each concept in each language edition is to the right.

Intersection			Only English	Only German
Title	DE	EN	Title	Title
Association football	27	8	AllMusic	Actor
Australia	32	11	American Civil War	Ancient Greek
Austria	5	55	Animal	Ancient Rome
Belgium	45	53	Argentina	Baden-Württemberg
Brazil	60	28	Arthropod	Baroque
California	51	27	BBC	Basic Law for Germany
Canada	21	4	Census	Bavaria
Catholic Church	17	24	C. European Summer Time	Berlin
China	35	15	Central European Time	Bishop
Denmark	49	78	Chicago	CDU Party (Germany)
Departments of France	74	38	Chordate	Christianity
England	25	5	Communes of France	Classical antiquity
English language	9	12	Egypt	Cologne
Europe	39	25	Flowering plant	Czech Republic
European Union	33	57	Gene	Denkmalschutz
Finland	97	87	Genus	Doctor of Philosophy
France	2	1	Geog. Names Info. Sys.	Dresden
French language	34	43	Gmina	East Germany
German language	20	65	Indonesia	Frankfurt
Germany	1	3	Insect	Hamburg
Greece	75	66	Iran	Hungary
Greek language	41	52	Ireland	Ice hockey
India	38	7	Israel	Italian language
Italy	8	13	Lepidoptera	Jurisprudence
Japan	24	14	List of sovereign states	Leipzig
Latin	4	22	Los Angeles	Lower Saxony
London	13	16	Member of Parliament	Medicine
Mathematics	83	86	New York	Moscow
Mexico	76	42	New Zealand	Munich
Middle Ages	30	73	North America	Napoleon
National Reg. of Historic Places	86	49	Ontario	Nazism
Netherlands	26	26	Oxford University Press	North Rhine-Westphalia
New York City	16	18	Pakistan	Philosophy
Norway	64	46	Philippines	Politician
Paris	12	37	Plant	Pope
Poland	23	20	Population density	Protestant Reformation
Portugal	55	71	Powiat	Prussia
Rome	36	88	Protein	Renaissance
Russia	15	19	Romania	Rhineland-Palatinate
Soviet Union	29	32	Scotland	Romanization of Japanese
Spain	18	21	South Africa	SDP Party (GermanY)
Species	77	54	Spanish language	State (polity)
Sweden	31	34	The Guardian	States of Germany
Switzerland	6	44	The New York Times	Township (USA)
Turkey	48	48	US Census Bureau	University
U.S. state	43	76	Village	Vienna

<b>Intersection</b>	<b>Only English</b>	<b>Only German</b>
United Kingdom	10	2
United Nations	66	80
United States	0	0
United States dollar	63	84
Washington, D.C.	96	40
World War I	11	23
World War II	3	6

## G Culture Hearths by Language-defined Community

<b>Language-defined Culture</b>	<b>Countries in Culture Hearth</b>
Catalan speakers	Spain, France, Italy
Chinese speakers	China, Taiwan, Singapore
Czech speakers	Czech Republic, Slovakia
Danish speakers	Denmark
Dutch speakers	The Netherlands, Belgium, Suriname
English speakers	Antigua and Barbuda, Australia, The Bahamas, Bangladesh, Barbados, Belize, Botswana, Cameroon, Canada, Dominica, Eritrea, Fiji, The Gambia, Ghana, Grenada, Guyana, India, Republic of Ireland, Jamaica, Kenya, Kiribati, Lesotho, Liberia, Malawi, Malaysia, Malta, Marshall Islands, Mauritius, Micronesia, Namibia, Nauru, New Zealand, Nigeria, Pakistan, Palau, Papua New Guinea, Philippines, Rwanda, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Samoa, Seychelles, Sierra Leone, Singapore, Solomon Islands, Somaliland, South Africa, South Sudan, Sri Lanka, Sudan, Swaziland, Tanzania, Tonga, Trinidad and Tobago, Tuvalu, Uganda, United Kingdom, United States of America, Vanuatu, Zambia, Zimbabwe
Finnish speakers	Finland
French speakers	Belgium, Benin, Burkina Faso, Burundi, Cameroon, Canada, Central African Republic, Chad, Comoros, Ivory Coast, Democratic Republic of the Congo, Djibouti, Equatorial Guinea, France, French Guiana, French Polynesia, French Loyalty Islands, French Southern and Antarctic Lands, Guadeloupe, Martinique, Mayotte, New Caledonia, Réunion, Saint Barthélemy, Saint Martin, Saint Pierre and Miquelon, Wallis and Futuna, Gabon, Guinea, Haiti, Luxembourg, Madagascar, Mali, Mauritius, Monaco, Niger, Republic of the Congo, Rwanda, Senegal, Seychelles, Switzerland, Togo, Vanuatu
German speakers	Austria, Belgium, Germany, Liechtenstein, Luxembourg, Italy, Switzerland
Hebrew speakers	Israel
Hungarian speakers	Hungary

Indonesia speakers	Indonesia
Italian speakers	Italy, Switzerland, San Marino, Vatican City
Japanese speakers	Japan
Korean speakers	North Korea, South Korea
Norwegian speakers	Norway
Polish speakers	Poland
Portuguese speakers	Angola, Brazil, Cape Verde, East Timor, Equatorial Guinea, Guinea-Bissau
Romanian speakers	Romania, Moldova
Russian speakers	Russia, Abkhazia, Belarus, Kazakhstan, Kyrgyzstan, South Ossetia, Tajikistan, Transnistria
Slovak speakers	Slovakia, Czech Republic
Spanish speakers	Argentina, Bolivia, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Equatorial Guinea, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Spain, Uruguay, Venezuela
Swedish speakers	Sweden, Finland
Turkish speakers	Turkey, Cyprus
Ukrainian speakers	Ukraine

The countries that are defined to be in each language-defined culture's culture hearth are shown above. Note that "country" is defined loosely here. Some of the countries above are not widely recognized as independent. The loose definition is used here to match our sources.