Paul Laufer

Cultural Similarity, Understanding and Affinity on Wikipedia Cuisine Pages

Master Thesis

Graz University of Technology

Knowledge Management Institute
Head: Univ.-Prof. Dr. Stefanie Lindstaedt

Supervisor: Prof. Dr. Markus Strohmaier
Advisor: Dr. Claudia Wagner

Cologne, August 2014

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Graz, _____ _____

Date Signature

Eidesstattliche Erklärung[1]

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

Graz, am _____ _____

Datum Unterschrift

[1] Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008; Genehmigung des Senates am 1.12.2008

Abstract

Quantifying cultural relations is a difficult and tedious task which is often performed manually. This thesis presents a novel approach for automatically inferring cultural similarities, understanding and affinities between different cultures in online media. These three cultural relations provide valuable insight into the ties between cultural groups. The introduced method is explained and then evaluated on the basis of a single cultural dimension, food, and its representation on the online encyclopedia Wikipedia. 31 different European cuisines are analyzed and the findings of this analysis are presented. Finally, the strengths and weaknesses of the approach are discussed.

Acknowledgements

I would like to express my gratitude to both Markus Strohmaier and Claudia Wagner who supervised me during the writing of this thesis. Their feedback and criticism as well as the many fruitful discussions did not only result in valid contributions to this thesis, but also helped me develop personally and taught me to manage the scientific research process. I am particularly thankful for Claudia's continuous support and the valuable perspectives and ideas she shared with me. Additionally, I would like to thank GESIS - Leibnitz Institute for the Social Sciences[2] for financially supporting the research. I am also grateful for the excellent team at the Computational Social Science department that inspired me during our weekly meetings. At last, I would like to thank David García and Thiago Silva for providing the Eurovision and Foursquare datasets.

[2] http://www.gesis.org

Contents

List of Figures

1 Introduction

1.1 Motivation

In 1980, Hofstede laid the foundations to measure differences between cultures on a large scale in his publication "Culture's Consequences" [Hof80], in which he analyzed over 100,000 questionnaires answered by IBM employees in 40 different countries. He extended his research to more than 70 countries in 2010 [HHM10]. The many other survey-based studies which followed brought valuable insight into the distribution of cultural values. Such qualitative approaches are, however, limited as the cultural background of the researcher might introduce a bias (see e.g. [Ail08] and [Xin09]). Recent research in the field of computational social science shows that analyzing large-scale datasets such as the voting history of the Eurovision Song Contest [GT13], Foursquare check-ins [Sil+14] or the behavior on Twitter [GQJ13], may help to overcome these limitations, while potentially introducing other biases, as discussed in Section 6. In fact, the assumption that knowledge does not represent universal truths, but those of the culture which portrays the knowledge [Bro94], comes as an advantage when analyzing large-scale user-generated content on the Web. These different views on the same topic can be used as a proxy for measuring cultural relations.

This study addresses multiple problems that arise when measuring culture through surveys and other comparable methods. These approaches, which come from the social sciences, have in common that they are (i) costly, (ii) time-consuming and (iii) biased by the interviewing or data collection and interpretation process. The method presented in this thesis attempts to overcome these problems by automatically measuring cultural values using readily available data encoded in the different versions of online knowledge repositories, particularly Wikipedia.

Wikipedia is the largest and most successful encyclopedia that is globally and collaboratively generated. Although Wikipedia is a compilation of facts, it is likely that each language edition of Wikipedia is biased by the views of the people who speak it well enough to contribute to it [Bao+12; MS13]. Yasseri shows that the different language editions (except for the English Wikipedia) can be reasonably mapped to

nations predominantly speaking the respective language [YSK11]. Therefore, the different Wikipedae can be used as a proxy for cultural views [OR11; HG09].

Culture is expressed in many different ways spanning across many cultural dimensions, such as art, literature, music etc. [Xin09]. Although the methods presented in the following chapters are generally applicable, this thesis only focuses on one cultural dimension, which is particularly strongly related to culture: the perception of food and cuisines. This choice was taken as, according to Fischler [Fis88], cuisines are not only a mere combination of ingredients, but a representation of an incorporated world view. Therefore it is safe to assume that the perception of food and different cuisines differs from one culture to another. This definition also coincides with Hofstede's definition of culture as "a tendency to prefer certain states of affairs over others" [Hof80]. Hofstede also adds that culture can only be assessed in relative terms and not in absolute values.

1.2 Approach & Research Question

Based on the ideas of international culturology [Xin09], three different interdependent cultural dimensions, *cultural similarity, cultural understanding and cultural affinity* are analyzed. In order to infer these relations, the representation of culturally relevant resources, such as beliefs, ideology, art etc. on different language editions of Wikipedia are compared (an example of a cultural resource on Wikipedia is depicted in Figure 1.1). This approach used to explain culture has been discussed by Carley [Car91] and Lietz & Strohmaier [Lie+14], who describe culture as a distribution of referenced facts.

The first dimension, *cultural similarity*, shall provide a quantifiable measure which may be used to, for instance, monitor the convergence or divergence of different cultural groups over time. The second dimension, *cultural understanding*, is inferred from Xintian [Xin09], who noted that every cultural group can only understand other cultures through the distorted perspectives influenced by their own ethics, moral, beliefs etc. In this thesis, a comparative method which provides means to evaluate such understanding between cultures, is presented. Xintian further argued that the different cultural perspectives can easily lead to misunderstandings, influencing the political, social and economic affinity relations between different cultural groups or nations. The third dimension, *cultural affinity and biases*, provides an aggregated view on the affinities expressed by different language versions on Wikipedia.

In order to make the automated procedure presented in this thesis feasible, the following assumptions are made:

Paella

Da Wikipedia, l'enciclopedia libera.

La **paella** è un piatto tradizionale della cucina valenciana in Spagna, diffusosi in tutto il Mar Mediterraneo e nell'America Latina. Il piatto, a base di riso, zafferano e carne, è simile al risotto italiano, al pilaf turco o al biryani indo-pakistano e viene preparato nella tipica padella da cui prende il nome, la *paella* o *paellera*.

Indice [nascondi]

1 Etimologia del nome
2 Il piatto
3 Voci correlate
4 Altri progetti
5 Collegamenti esterni
6 Note

Etimologia del nome [modifica | modifica wikitesto]

Etimologicamente la parola valenciana[1] [2][3] *paella* deriva dal latino *patella*, dal quale sono derivati anche il francese *poêle*, lo spagnolo medievale *padilla* e l'italiano padella. In origine il termine

Paella

Origini	
IPA	/paˈeʎa/
Luogo d'origine	Spagna
Regione	Andalusia
Diffusione	Penisola iberica Mar Mediterraneo America Meridionale

Figure 1.1: **Example of a culturally relevant resource on Wikipedia:** The Spanish dish Paella as described on the Italian Wikipedia.

- Each language edition of Wikipedia can be mapped to one or many nations that predominantly speak this language. According to the statistics of the Wikipedia Media Foundations, the contributions of most language editions stem from a single country (with the notable exceptions of the Spanish and English Wikipedia). This assumption is also supported by the findings of Yasseri et al. who mapped Wikipedia language editions to nations by analyzing the timestamps of changes and comparing them to the respective time zones [YSK11].

- Cultural groups can be mapped to nations. This is a simplification assumed in almost all current inter-cultural research, which is mainly due to the fact that data is mostly only available on a national level.

- A culturally relevant resource can be directly mapped to nations. For instance, it is assumed that the waltz can be directly linked to the Austrian culture, whereas flamenco is associated with the Spanish culture.

The main research question which will be answered in this thesis is *Can cultural relations be extracted from the different perspectives inherent to the online knowledge repository Wikipedia?*

1.3 Contributions & Findings

The main contribution of this thesis is the algorithm which quantifies cultural relations automatically using Wikipedia (or potentially other online sources) by analyzing different representations of culturally relevant resources. Applying these methods on a particular sub-domain, in this case food, reveals some interesting insights into the different perspectives.

Firstly, the findings reveal that the attention a Wikipedia article receives (i.e. the number of views) correlates strongly with structural measures of the article, such as the number of outlinks or the number of words. This leads to the assumption that Wikipedia articles, which are heavily visited, also grow extensively, and that the amount of information used to define each concept can be used as a proxy for its importance. For example, if the article describing the Lithuanian cuisine is viewed more often on the German Wikipedia than on the Polish one, this can be a strong indicator that the article on the German Wikipedia is also longer and contains more outlinks than the Polish article. This causal relation also remains true when the direction is reversed (i.e. if the article is longer, it can be assumed that it is viewed more frequently).

Secondly, the findings show that cultural similarity on Wikipedia is very well approximated by considering multiple language editions and not only the pair the similarity of which is analyzed. This means that, assuming a considerable number of different perspectives (Wikipedia language editions), the similarity between two language groups can be approximated rather well, even if the two language groups being analyzed do not know anything of each other (e.g. do not describe each other's culturally relevant resources). Using the example of cuisines, one could say that the similarity between, for instance, the Bosnian and the Peruvian cuisine can be estimated rather well, regardless of whether the Bosnian Wikipedia describes the Peruvian cuisine (and vice versa) or not. However, the findings also show that additional cultural dimensions, such as music or literature, have to be considered before a reasonable estimation of one culture's perspectives can be made.

Concerning the understanding of other cultures, the findings demonstrate that popular cuisines, such as the Italian or French cuisine, are much better understood than less popular cuisines, such as the Bosnian or Bulgarian. Although these results are not surprising, the automatic inference of such information is valuable, as it allows for a repeated, comparable quantification and reveals interesting latent information (such as the surprisingly good understanding of the Turkish cuisine by many other European cultures).

Finally, the analysis of the bias inherent to Wikipedia cuisine articles confirms previous results which indicated that each Wikipedia edition focuses on their own cultur-

ally relevant resources as well as on those of geographically close regions more than on others [MK06; CH11; HG09]. However, further research is necessary in order to determine how well these detected affinities can be used as an approximation for cultural affinities.

1.4 Overview

The remaining part of this thesis is structured as follows: Chapter 2 discusses related research on identifying and evaluating cultural similarity, understanding and affinity. Chapters 3.1 and 3.2 describe the used dataset and applied methodologies that explain how Wikipedia was used to analyze the previously mentioned cultural values. The results for each topic are then presented in Chapter 4 including an evaluation of their reliability. Chapters 5 and 6 critically analyze the results and their shortcomings and Chapter 7 concludes this thesis.

2 Related Work

2.1 Cultural Diversity and Cultural Differences

In [Xin09], Xintian highlights the importance of international cultorology in order to support international communication. He notes that cultural pluralism is unavoidable and that a nation's interests, alliances and conflicts are increasingly influenced by culture. He also emphasizes the importance of media in maintaining and creating strong cultural relations, which can be reasonably extended to all platforms distributing information, including collaborative online systems such as Wikipedia.

The largest study on cultural differences has been performed by Hofstede et al. [Hof80] who performed interviews in settlements of IBM in more than 80 countries and derived a number of cultural dimensions, which are still used in many recent publications to quantify cultural differences. He later extended his work to more countries and refined his findings in [Hof02; HHM10].

Gao et. al [GHZ13] performed a survey on the employees of two big IT companies located in China and the U.S. After verifying the cultural relatedness of the participants with the respective nationalities, they investigated in the network factors that play an important role for selecting colleagues. They hypothesized that the two different cultural groups would value the structural positions of their potential business partners differently. As reciprocity and harmony in social networks are more important in the Chinese culture, whereas task-orientation and efficient access to resources are more common in North America, the authors assumed that Chinese workers would choose their potential colleagues within their networks whereas American employees would reach out to more distant partners. Their second hypothesis stated that in the Eastern culture, the hierarchical position of the potential associate is more important in contrast to the Western culture which values expertise more. Their first hypothesis was not proven, the second one, however, was supported by their findings, showing significant differences in the approaches that both cultural groups use to select their potential colleagues.

Another study which dealt with the cultural differences between Chinese and Americans was performed by Nguyen and Fussel [NF12]. They presented participants

from both cultures with an instant messaging system which allowed them to communicate with one another and performed a retrospective analysis. Apart from the expected difficulties in inter-cultural communication and perceived annoyance, their findings indicated a higher dissatisfaction among every participant working with a Chinese partner (regardless of the participant's cultural background). This supports the results of previous research by Wang et al. [WFS09], who discovered that people in high context cultures (such as Americans) can more easily adapt to match the style of low-context cultures (such as Chinese).

2.1.1 Cultural Diversity and Cultural Differences on Wikipedia

In [HG09], the authors hypothesized that the home region of each Wikipedia edition (i.e. the region where the language of the Wikipedia edition is either primary and/or has a significant number of speakers) is to a large extent the geographic focus of each Wikipedia edition. The authors suggested two measures (sum of in-degree and page rank) to measure the bias in 15 different Wikipedae, only focusing on Wikipedia articles which have a location (i.e. they are geotagged). In a similar manner, Overell and Rüger [OR11] analyzed whether different language editions of Wikipedia focus on the description of geographic regions where their language is spoken. They measured the self-focus bias of a language-specific Wikipedia edition using the ratio between the number of links made to locations where the language of this Wikipedia edition is primarily spoken and links to locations where the language of this Wikipedia edition is not spoken. Furthermore, they proposed a model which estimates the relevance of a location l for a person p based on the product of the subjective interestingness of location l to person p and the objective interestingness of the location. The subjective interestingness is based on the relationship between p and l, while the objective interestingness is based on properties of l. Both publications only considered articles with attached locations (i.e. longitude and latitude). The approach presented in this thesis differs in this regard, as a topical pre-selection was used by only considering articles describing cultural resources (e.g. the Italian cuisine) which can be related with specific geolocations (e.g. Italy). Using this approach, more culturally relevant resources (such as cuisines) can be incorporated, instead of only considering resources with a specific location, such as cities or monuments. It was assumed that this extension allows for a richer description of cultural values. This idea is supported by Maurer [MK06] who introduced the concept of local heroes, which indicates that famous people are considered more important (e.g. have longer and more extensive descriptions) on their home country's Wikipedia edition than on others. Callahan and Herring [CH11] later proved this idea when they compared descriptions of famous persons on the English and Polish Wikipedia. They found culturally biased differences for both the extent and the concepts with which

the persons were described. In another study, Pfeil [PZA06] analyzed the Wikipedia article about games in multiple languages, focusing on their edit histories. She then compared the edits with the cultural dimensions proposed by Hofstede and showed that one's editing behavior, expressed by deletions, additions or corrections, is heavily influenced by cultural values. She also found that the Internet in general and Wikipedia in specific are not a culturally neutral space. Again, this idea had already been brought up by Maurer [MK06] who stated that "... even if an article is written in compliance with the 'neutral point of view' the varying cultural, social, national and lingual backgrounds can have an enormous influence. Hence, content on Wikipedia can only be as professional and balanced as its authors and their demography are." Another study analyzing historically relevant persons on Wikipedia was performed by Aragon et al. [Ara+12]. They focused on the relationships between historical figures which appear in different language editions of Wikipedia. For each version, they extracted a directed network with nodes representing people and edges representing links between the articles describing them. The resulting networks were then analyzed using standard network analysis tools. The results indicated that, for instance, the clustering coefficient is very low for all language editions except for the Chinese one. By comparing the similarities of these networks between different languages, they found similar language- and geographical clusters as in this thesis (for instance, Scandinavian languages or Roman languages). Their findings suggest that biographical connections are recorded differently in different Wikipedae.

Similarly, in [ES13], the authors also applied a quantitative approach in order to perform a cross-cultural study on Wikipedia. They used three different ranking algorithms in different language editions of Wikipedia in order to determine the 30 most prominent articles about persons in each language edition. After they manually assigned these persons to fields of human activities and a corresponding origin culture, they compared the results of the ranking algorithms and calculated a locality for each of the persons. This method allowed them to define e.g. George W. Bush as belonging to 'Politics', 'English' and 'Local' for the English Wikipedia, while Jesus was categorized as 'Religion', 'World' and 'Non-Local'. Regardless of the ranking algorithm, their findings suggest that about 50% of the top 30 persons of each language edition belong to the culture of the edition, supporting the idea of a self-focus bias. For the other cross-cultural or 'global' heroes, they found that less links pointed to these articles, but that these links originated from more prominent articles. Examples for such globally important persons are for instance Napoleon, Michael Jackson or Adolf Hitler. Finally, the authors constructed a network of cultures by considering the culture of the prominent person in each Wikipedia edition and used these counts to create a directed network (e.g. if the English Wikipedia has 2 French persons among their top 30, this relation would result in a link with weight 2 from the English to the French culture). They found that their results were not sufficiently significant, as they had only used a small sample of language editions, most of which

originated from one geographical region. In this thesis, the idea of using resources (e.g. notable persons) which are relevant to or originate from a particular culture and their representations in different language editions is picked up. However, not only a simple count of existing articles was used, but the actual content of the articles was analyzed.

In [NG11], the authors chose a small sample of language editions (English, German, Japanese, Korean and Finnish) in order to analyze the cultural differences in behaviors. They particularly chose the Korean and Finnish Wikipedia, as these languages are both only spoken in one country, thus allowing a mostly unbiased analysis of their cultures. They noted that although both countries are among the heaviest users of Internet, the Finnish Wikipedia has significantly more articles per first-language speaker than the Korean one. By analyzing the user interactions with each other over an extended period of time, they found that the Japanese and Korean language editions show a much less stable collaboration network than their Western counterparts. In the second part of their analysis, they looked at the different ways conflicts are resolved in different language editions. Their findings indicate that egalitarian cultures (such as the Finnish culture) collaborate notably different from more hierarchical cultures (such as the Japanese culture).

In another paper, Wang et al. compared the different language editions of Wikipedia according to their concept overlap [War+12]. They tried to find the 'ur-Wikipedia', an agglomeration of knowledge which is part of all language editions and can therefore be considered important in all parts of the world, regardless of the local language. This idea is somewhat contradictive to the results of Hecht and Gergle [HG10] who found strong evidence against the existence of a global consensus of world knowledge. Wang et al., however, discovered that there is a set of almost 300 articles which are present in the majority of language editions. These articles cover mainly general topics, the biggest ones being time-related articles and descriptions of countries and cities. When they closely investigated the spread of certain articles, e.g. the 'True Jesus Church', which is present in 254 of 283 language editions, they found that most of those articles had been initiated by a very small group of users. This raises the question whether such articles in a language edition represent the true interest of the entire language group or just a subset. They also calculated the similarity between different language editions using a similar approach to the one applied in this thesis. However, instead of limiting themselves to a certain domain (e.g. food), they used the entire set of pages on each Wikipedia edition. Their results suggest that roughly 4% of similarity can be explained by geographical proximity and around 50% by the sizes of the respective Wikipedae. As only a very small and dedicated set of articles was used in the analysis presented in this thesis, the size only plays a minor role.

Liao and Petzold introduced a geographic and linguistic normalization model which allows for a better comparison between statistics of different Wikipedae [LP14]. Pre-

vious research showed that there is a strong correlation between the number of Wikipedia editors and the Internet population, as well as the total tertiary-educated population. This correlation was used to normalize different language editions, as an additional factor besides the mere sizes. The method introduced by Liao and Petzold aims to break down such an attempt in order to analyze geo-linguistic units (such as Egyptian Arabic, Saudi Arabia Arabic, etc.). They used data from the Language-Territory Information database to break down Wikipedia's edit and view statistics and showed that the contributions of different countries to each language edition can be reasonably approximated.

2.1.2 Cultural Diversity and Cultural Differences in Other Online Resources

Another interesting study [GQJ13] explored the link between users' activities on Twitter and the culture of their home countries. It is believed that the way in which people interact with one another, how they perceive and accept power and how they perceive time drastically differs among countries. The study presented in [GQJ13] showed that interactions on Twitter reveal interesting differences. For example, countries with a higher pace of life tend to be more predictable and people living in collectivistic countries tend to interact more with others. The authors tried to infer cultural descriptions from the individual activities of users, while the approach presented in this thesis explores the collective perception / description of a group of users on culturally relevant items.

Similarly, in [Rei+13], the authors analyzed 1.5 million polls from 211 countries on the popular scheduling platform Doodle. The authors used the theoretical background of Hofstede [HHM10] and Inglehart [IB00] to link the different behaviors in group decision processes and the perception of time to certain cultural dimensions. They assumed, for instance, that individualistic countries, which are often more monochromic in their perception of time, would create more polls and their time options would be more precise in order to organize their life which is thought to be more scheduled. They also hypothesized that the GDP per capita, which supports self-expression, would positively correlate with the number of polls. Using the empirical data from Doodle, their assumptions were mostly confirmed, with the exception of some outliers. As the study presented in this thesis, the authors tried to extract cultural values from quantitative online data. Different to the approach presented in the remainder of this thesis, they did not focus on cultural relations between countries, but on cultural dimensions used for comparing different nationalities. Their study demonstrates that the users of Doodle do show different behaviors which can reasonably be associated with different cultural values and beliefs, indicating that the group of internet users is in fact culturally diverse.

In [GT13] and [SV06] the authors used the voting history from the Eurovision Song Contest to explore cultural dynamics in Europe. In [SV06] the authors showed that geographical factors and religion strongly influence voting behavior. By using Turkish migration data in Europe they found evidence that ethnicity plays a role. Countries with a large proportion of Turkish inhabitants tend to favor Turkish songs. The results of these studies were also used as a ground truth to evaluate the findings presented in this thesis.

Another interesting paper discusses cultural diversity from a different perspective. Instead of working with textual manifestations of cultural beliefs and views, Yanai and Bingyu analyzed the content of geo-tagged photos extracted from Flickr. They applied several data processing methods originally created for the field of textual data mining, but instead of using word vectors they created vectors holding visual information derived from SIFT descriptors. By applying this approach, they were able to identify different interpretations of general concepts. For instance, they found that the concept 'noodles' is most often used to describe Spaghetti in photos from Europe, whereas it is primarily associated with Ramen, a Japanese noodle soup, in photos taken in Japan [YYQ09].

2.2 Wikipedia

Steiner provided a general overview of the changes made on Wikipedia in [Ste14]. He developed a tool which monitors the edit activity on all 287 language editions and found that, for instance, around 15% of all edits are made by bots and 26% of edits by anonymous users. He also provided reasonable justifications for the different language editions to grow and be modified in a different way (e.g. anonymously vs. logged-in) but did not further investigate into possible cultural explanations.

In [Ng12], the authors analyzed notable Americans' bibliographies published on Wikipedia and searched their descriptions for sociologically important characteristics which promoted their notability. They validated their method by using two characteristics, first names and birth places, and compared them to external data. Their findings show, for instance, that persons with rare first names are more likely to appear on Wikipedia. They suggest that such automated analyses could aid to validate and perform sociological studies investigating success, where self-reporting is often problematic.

2.2.1 Bias on Wikipedia

In [Ras08], Rask analyzed the connection between the development of a country and the activity of Wikipedia contributions. He concluded that countries with a lower Human Development Index (HDI), such as Russia or Poland, show less interest in editing and maintaining Wikipedia than more developed countries, such as Denmark or Germany. However, the findings may not be applicable nowadays, as they date back to 2007. Since then, the World Wide Web has drastically developed and grew in importance. In an even earlier publication, in 2005, Bellomi and Bonato analyzed the English Wikipedia using methods intended for structural network analysis and found that the English Wikipedia is generally biased towards Western cultures [BB05]. They, however, did not perform further evaluations on other language editions to validate their results. The study of Aragon et al. [Ara+12] revealed that many of the central nodes in their network of biographical articles in 15 different language editions represented Americans or Europeans. However, their initial data set was already biased, as they only considered people with an article on the English Wikipedia. The bias towards Europe and the United States is also supported by the findings of Wang et al. [War+12].

2.2.2 Editors of Wikipedia

In [JL12], Jurgens and Wu created a bipartite graph of users and articles of Wikipedia labelling both the user nodes with their type (anonymous, administrator, bot or registered user) and the edges with the type of interaction (major/minor add/delete or revert). Out of triplets of consecutive interactions they created what they called motifs as a combination of different users interacting with an article in different ways. They used these motifs to describe different articles and analyzed by how much different topics of articles support conflicting or cooperative behavior. Using this approach, they were able to provide insights into the motivations and behavior of users of Wikipedia. Finally, they applied their method on the historical data of the English Wikipedia and found that, although the growth of the number of interactions remained stable, different usage patterns developed differently throughout time.

Although not performed on Wikipedia, but on the collaborative bicycling platform Cyclopath, Panciera et al. [PMT14] analyzed the core group of editors, which produce the majority of contributions and take on much of the community maintenance work. Similar patterns of only a small proportion of users producing most of the content have been recorded and analyzed for Wikipedia. In contrast to many other studies dealing with this group of core contributors, which mainly address the quantity and quality of edits, this study focuses particularly on the skills, the knowledge and the experience of the central group. The same approach had previously been

used on Wikipedia itself, performed by Bryant et al. [BFB05]. Their main finding was that, other than expected, the core group did not consist of people with a strong tendency towards cycling and were therefore no experts. However, they all had in common that they were interested in open collaboration and free knowledge exchange and almost all of them were heavily involved on Wikipedia.

Ortega et al. analyzed the inequality of contributions (i.e. few editors contribute the majority of content) in order to find conclusive results to an open research question that had been frequently tackled [OGR08]. They applied methods employed in economical sciences to analyze the ten biggest Wikipedia language editions on a longitudinal basis and retrieved clear measures indicating a strong inequality. They concluded that 90% of users are responsible for less than 10% of the overall contributions in all language editions alike. However, differences arise when looking at the sizes of the different Wikipedae. Those editions with a large number of articles (with the exception of the Japanese Wikipedia) seem to have a more equal distribution of contributions. One possible explanation is that when more topics are covered, more potential authors feel like they can contribute valuable information. Additionally to those static measures, the authors also analyzed the development of inequality from the beginning of each language edition. They found that for the first 20 months, the inequality is relatively unstable, followed by a stable phase with an inequality of around 85%.

Yasseri et al. estimated the geographic origin of editors from their temporal behavior patterns by mapping the edit times to time zones [YSK11]. Using this approach, they were able to trace the origin of the editors of different language editions to a single time zone and assumed the related country. For language editions with editors from more than one time zone, they calculated the share of each country. For instance, their results indicate that although Quebec in Canada is French, the share of contributions to the French Wikipedia by North Americans only amounts to 5%.

2.2.3 Knowledge Distribution in Single and Multiple Wikipedia(s)

In her work from 2009, Filatova [Filo9b] analyzed the distribution of information used to describe a single concept across multiple language editions of Wikipedia. She used machine translation and compared the translated sentences with one another using the tf/idf measure in order to identify whether two versions contain the same information. The derived overlap was then used to create summaries of the Wikipedia articles, the quality of which was evaluated. She found that while the facts used to describe a concept in two different languages were not contradictory,

every language edition added further information. In another study, the same author analyzed the differences in articles describing 48 different persons in different language editions of Wikipedia and concluded that the descriptions varied in both the amount and choice of information [Filo9a]. In contrary to these publications discussing cultural differences on Wikipedia, the approach presented in this thesis does not take a qualitative approach to analyze the content on a subset of Wikipedia articles manually, but rather tries to propose methods to perform such inter-language comparisons automatically.

Hecht and Gergle showed in [HG10] that the diversity of information across Wikipedia language editions is much greater than initially estimated by literature and that only one tenth of a percent is comprised of common concepts. They stated that only a subset of the current research acknowledges and analyzes the diversity of knowledge. Furthermore, they introduced a measure to quantify the diversity inherent to different versions of Wikipedia. Finally, they concluded that one of the major research challenges for future information retrieval systems will be the automatic separation of culture-dependent and globally applicable information, such as birth dates. The methods described in the following sections tackle this problem to a certain extent, as the cultural influence on Wikipedia articles is analyzed and quantified.

The research presented in [Bao+12] combines different cultural views and aligns them side-by-side, as a vast amount of information is only available in some language editions. They created a system called Omnipedia which allows users to retrieve different views and compare them directly using machine translation. It identifies commonly and less commonly discussed concepts by looking at the outlinks. That way, one can visualize how one concept is described in different language editions of Wikipedia. Additionally, they improved the inter-language link graph to resolve ambiguities (if multiple concepts in one language are linked to one concept in another language). However, only around 1% of all multilingual articles are initially ambiguous. Similarly, in [MS13], Massa and Scrinzi presented their system Manypedia, which aims to help understand the Linguistic Point of View, as they defined it, inherent to the different language versions of Wikipedia. They claimed that the 'Neutral Point of View' policy promoted by Wikipedia is only valid within each language edition which led to the development of an online system that allows comparisons of different language versions, similar to the Omnipedia platform.

2.3 Food

In [Sil+14] the authors used food and drink related check-ins from Foursquare to assess the cultural distance between countries, cities and regions. The authors only focused on users who have a check-in history which is limited to one country since

they wanted to estimate the home-location of the user. Then, they represented each user as a vector of the categories of the locations he / she checked in. All users living in one region constituted the cultural make-up of the geographic area. The authors clustered regions by their cultural similarity inferred from their food and drink preferences and compared their results with the cultural map of the world based on the World Value Survey [IW10]. Their results reveal interesting similarities. Similar to their work, the approach presented in this thesis also deals with the estimation of cultural differences and the possibilities to assess them using non-reactive research methods. However, instead of considering the eating and drinking habits of users, the users' perception and views on different cuisines are taken into account. Also, the used dataset, Wikipedia, is larger and possibly less biased than the Foursquare check-ins. This is especially true when considering that the authors only used those check-ins which were published via Twitter, further restricting the already biased set of Foursquare users. As Wikipedia does not require a smartphone and has a greater coverage, one could argue that the users of Wikipedia are a better estimate of the entire population than Foursquare users who publish their check-ins on Twitter. Additionally, the authors of [Sil+14] assumed that users who only posted from one country also live there. This seems to be a realistic assumption, which leads to a high precision, but suffers from a low recall. In this thesis it was assumed that users who contribute to a specific language edition of Wikipedia are likely to live or have lived in one of the countries where this language is predominantly spoken. For example, someone who contributes to the Italian Wikipedia might be (i) an Italian living in Italy, (ii) an Italian living abroad or (iii) a non-Italian but someone who speaks the Italian language very well and probably has also spent a certain amount of time in Italy. It was assumed that all those groups of people are very familiar with the Italian culture.

Dixon et al. analyzed Tweets that revolved around food and performed a sentiment analysis on them [Dix+12]. They developed a system that monitors Twitter for English messages about food by using keyword-based methods. The messages were analyzed and stored in a database together with the identified food item, the location of the user and the sentiment of the message (positive or negative). This aggregated data was then compared to obesity levels and GDP per capita for each country separately. As they only considered English Tweets, their dataset is however biased and might not represent the true population of a (non-English speaking) country, but only those inhabitants who publicly tweet in English. However, besides some findings which were visible on a global scale, such as the popularity of meat and fast food, they also found references to culturally relevant food items and differences in the consumption that related to the wealth of the countries. Overall, their findings indicate global trends related to food, such as a flattening of the food culture and the high sentiments related to unhealthy food.

3 Materials and Methods

3.1 Description of Dataset

Altogether, 27 different language editions of Wikipedia and 31 different cuisines from across Europe were analyzed, as listed in Table 3.1. The articles describing the different cuisines were downloaded in all language editions to capture the different views on each cuisine. For instance, the article "Italienische Küche" (Italian cuisine) on the German Wikipedia can be seen as a description of the Italian cuisine from the perspective of the German culture. An extract of this article is shown in Figure 3.1. It has to be noted though that not all language editions of Wikipedia include articles about all cuisines, so that a cuisine is represented on only 13 versions of Wikipedia on average (with a variance of 3.3). Based upon these cuisine articles, four different datasets were used, as described below:

Figure 3.1: Extract of an article about the Italian cuisine on the German Wikipedia

Language	Code	Size	Related Cuisines	Avg. Editors	Avg. Views
Bulgarian	bg	158,130	Bulgarian cuisine	17.91	4669
Bosnian	bs	48,761	Bosnian cuisine	26.00	853
Catalan	ca	422,684	Catalan cuisine	28.58	2155
Czech	cs	289,551	Czech cuisine	31.54	11361
Danish	da	186,047	Danish cuisine	45.50	3122
German	de	1,692,696	Austrian and German cuisine	146.34	108501
English	en	4,462,417	British, English and Irish cuisine	222.77	468724
Spanish	es	1,084,184	Spanish cuisine	77.78	137567
Estonian	et	121,329	Estonian cuisine	13.25	1221
Hungarian	hu	256,215	Hungarian cuisine	41.10	8603
Croatian	hr	143,375	Croatian cuisine	11.67	1362
Finnish	fi	342,384	Finnish cuisine	22.42	9467
French	fr	1,481,635	French cuisine	74.64	66692
Italian	it	1,103,118	Italian cuisine	46.23	39720
Lithuanian	lt	163,546	Lithuanian cuisine	18.33	4202
Latvian	lv	52,871	Latvian cuisine	13.67	1269
Dutch	nl	1,763,752	Belgian and Dutch cuisine	40.88	13125
Norwegian	no	412,649	Norwegian cuisine	30.75	2544
Polish	pl	1,031,851	Polish cuisine	46.37	42972
Portuguese	pt	821,450	Portuguese cuisine	37.67	48972
Romanian	ro	241,239	Romanian cuisine	23.00	3575
Russian	ru	1,093,578	Russian cuisine	58.68	59685
Slovak	sk	190,907	Slovak cuisine	21.67	2243
Serbian	sr	243,268	Serbian cuisine	25.00	523
Swedish	sv	1,612,310	Swedish cuisine	32.76	16945
Turkish	tr	224,742	Turkish cuisine	41.71	17674
Ukrainian	uk	496,343	Ukrainian cuisine	28.66	9960

Table 3.1: Language editions of Wikipedia, their language codes, their sizes, the related cuisines that were used, their average number of unique editors of cuisine articles and the average monthly views of cuisine articles (as of May 2014).

Word count dataset

The first dataset contains the number of words used to describe each cuisine on the different language editions. It is used as an initial approximation for the extent, with which each cuisine is described.

First hop dataset

The second dataset simply consists of the outgoing links of all existing articles. Both the number of outlinks and the referenced concepts were stored, allowing for both an analysis of the importance which is more stable than the word count (as some articles are basically a list of links) and an analysis of the concepts which are used to describe each cuisine in different language editions.

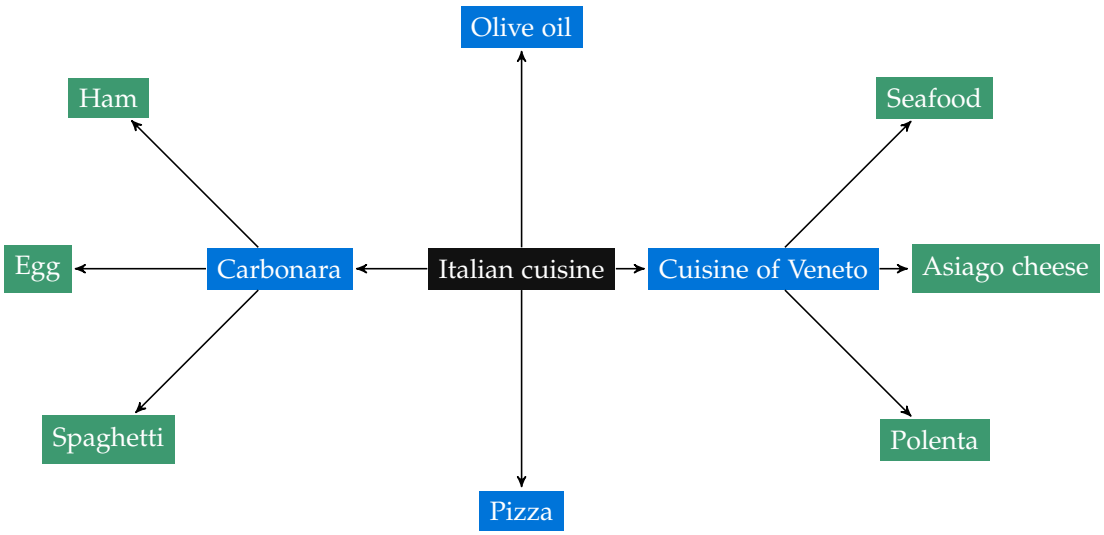


Figure 3.2: **First and Second hop dataset:** An example of how the first and second hop dataset were created. A simplified network of Wikipedia articles is shown. The arrows represent links from one article to another. Considering the seed article "Italian cuisine" (in black), the first hop dataset would consist of all articles to which the seed article links (in blue). The second hop dataset would additionally contain the articles to which the first hop articles link (in green).

Second hop dataset

As the quality of each cuisine article highly depends on the author(s) who wrote them, a third dataset was created by using a second hop link set originating from these cuisine articles. Figure 3.2 shows a simplified version of a sub graph of Wikipedia and the nodes that would end up in the first and second hop dataset. For obvious reasons, this third dataset is directly dependent on the second one, but allows for a greater range of concepts to be covered. As this approach introduced a considerable amount of noise (e.g. nonfood-related concepts, such as geographical entities), the Wikipedia category pages were used to clean the third dataset. Therefore, the category page for foods (e.g., `http://en.wikipedia.org/wiki/Category:Foods` for the English Wikipedia) was fetched in all language editions together with its subcategories up to a depth of 3. This number was chosen to limit the computational effort while maintaining an acceptable coverage and avoiding a semantic drift that was discovered when using more hierarchical levels. The threshold is derived from experimental results. Figure 3.3 shows such a category page. All articles in all resultant category pages were considered food-related items and were translated to the other languages using the inter-language-link-graph inherent to Wikipedia. This translation step was necessary, as not all language editions contain the same entries in their categories. The resultant article sets contained between 847 (for the Bosnian Wikipedia) and 33,574 (for the English Wikipedia) distinct food concepts for each language. Finally, the second hop dataset was truncated to only contain concepts

Figure 3.3: Extract of a Category Page about Foods on the Portuguese Wikipedia

which were considered "food-related" according to the category-based method. Although this approach may have removed some valid articles, as the category pages are far from maintaining an exhaustive list of foods, the remaining sub-graphs were still large enough to allow for a reasonable analysis.

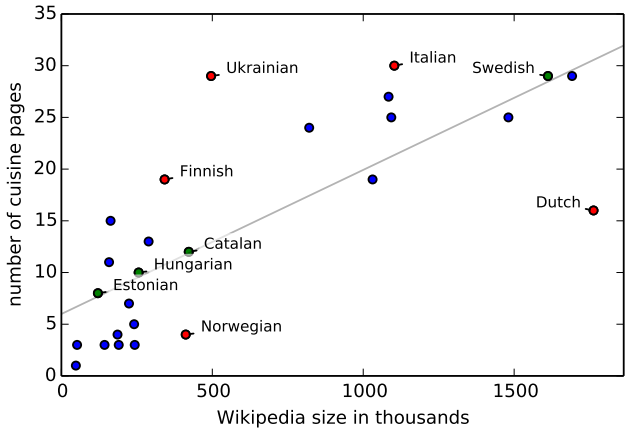
View counts dataset

Finally, a fourth dataset was fetched consisting of the view counts for each cuisine article (but not their first or second hop outlinks) in each language taken from an unofficial repository[1] between May 2013 and June 2014. Additionally to the length of the articles, which could theoretically be influenced heavily by single contributors, the view counts represent a more statistically stable dataset representing the attention of the different cuisines. However, the two measurements do correlate, as shown in the next section.

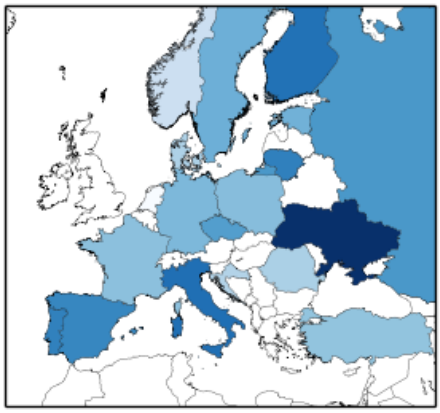
All data was fetched between April and June 2014.

Figure 3.4 shows the relation between the size of the different Wikipedae (i.e., the total number of pages they contain) and the number of (European) cuisine pages

[1] http://stats.grok.se/



(a)



(b)

Figure 3.4: **Do some language editions show a greater interest in cuisines than others?** The top plot (a) shows the relationship between the size of the respective Wikipedia and the number of cuisine articles it contains. The line is a linear approximation of the data. Data points in red are furthest from the norm, data points in green are closest to the norm. It has to be noted that the English Wikipedia was left out, as it is considerably bigger than all other language versions. The bottom plot (b) shows the distance of all countries to the first order approximation. Hence, the darker the country is plotted, the higher its interest in the European cuisines as expressed by the respective Wikipedia articles.

that they cover. One can see that no clear correlation exists, as some language versions tend to describe many of the cuisines, indicating a greater interest in the topic (such as the Italian, Ukrainian or Finnish Wikipedia), whereas others of comparable size only contain articles about very few cuisines (such as the Dutch or Norwegian Wikipedia). However, as only 31 cuisines were analyzed, such a comparison may not be generally valid, but should give a coarse overview of the used dataset.

3.2 Methods

This chapter describes the implemented approach to quantify cultural similarity, understanding and affinity and the steps taken to evaluate its results.

3.2.1 Approach

Cuisines were used as cultural resources as they are important in differentiating cultures or social groups [Fis88]. Furthermore, most cuisines can be directly related with a country or a society from which they originate. The approach itself is however open to other cultural resources, such as music or literature. The only requirements are that the chosen cultural resource can be associated with a country or a society and that Wikipedia articles about the resource exist in several language editions.

The approach uses the content of the article (specifically the outgoing links since they can be compared across different language editions) and the view statistics of the articles. *Cultural similarity* is measured by comparing the concepts used to describe two different cuisines (e.g. the cultural similarity between Italy and Spain is approximated by the concept overlap of their cuisine pages). *Cultural understanding* between two countries is measured by comparing their descriptions of the same resource which is culturally relevant to one of them (e.g. Germany's cultural understanding of France is approximated by the similarity with which both describe the French cuisine). *Cultural affinity and bias* is defined by the amount of attention one country pays to another country. Attention is measured by the level of detail of the created article as well as the number of views it receives. If the amount of attention exceeds what one would expect on average, the country is assigned a positive bias or affinity towards the other country. If the amount of attention is lower than expected, one could argue that a negative affinity or bias exists. Otherwise it is concluded that no bias exists. In the following, the three dimensions are described in more detail.

Cultural similarity

Cultural similarity is approximated by culinary similarity which is calculated with the overlap of concepts used to describe two different cuisines. Both a global perspective (i.e. the set of concepts used by all languages to describe each cuisine) and a local perspective (i.e. the set of concepts used by the two language editions of the cuisines' origin countries) are evaluated. It is assumed that the culinary similarity at least partially explains cultural similarity as it is one dimension of cultural identity

and can therefore be used as a proxy. The similarity is then calculated using Jaccard similarity (A and B refer to the sets of concepts used to describe two different cuisines):

$$\text{sim}(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{3.1}$$

Cultural Understanding

For calculating the cultural understanding of a target culture B by a source culture A, it is assumed that the respective article describing culture B on the Wikipedia of culture A is similar to the article that culture B uses to describe itself. On the contrary, if the resulting similarity is rather small, culture A seems to have a different definition of the cuisine of culture B than culture B itself. Therefore, culture A has a limited understanding of culture B. As cultures tend to describe their own cuisines in more detail than others (see later sections), the risk of a large Wikipedia describing cuisines more detailed than the actual target Wikipedia is minimal. In order to calculate the understanding, the concepts defined in the respective Wikipedia articles are used. Using the inter-language-links inherent to Wikipedia, same concepts in different languages can be identified. Finally, Jaccard similarity is calculated using the two sets of concepts, again using equation 3.1.

Cultural Affinity and Bias

The cultural affinity between a country A and B (or the bias of A for B) is quantified by measuring how much attention country A pays to the cultural resource of country B. The outcome of the collaborative edits (i.e. the content of the articles) are used as well as the view counts as a measure of attention. For instance, if the Finnish Wikipedia describes the Irish cuisine to a great extent (i.e. with many outgoing links) or is viewed heavily, then it can be assumed that the Irish cuisine is important to the Finnish people. In the following paragraphs, several formulas are described which take different approaches in calculating such affinity values.

The affinity / bias of one country towards another is measured using three formulas,

the first one being

$$bias_g(l,r) = \frac{\overbrace{f(l,r)}^{\text{attention twds. res. } r}}{\underbrace{\sum_{\bar{r}\in R} f(l,\bar{r})}_{\text{attention twds. all res.}}} - \overbrace{\frac{1}{\underbrace{|\{\bar{l} \in L \setminus l, \bar{r} \in R : f(\bar{l},\bar{r}) > 0\}|}_{\text{total number of resource descriptions}}} \sum_{\bar{l}\in L\setminus l} \frac{f(\bar{l},r)}{\sum_{\bar{r}\in R} f(\bar{l},\bar{r})}}^{\text{normalized rel. attention of others}} \tag{3.2}$$

L is the set of all languages, R is the set of all resources (e.g. cuisines). $bias_g(l,r)$ then calculates the bias of language l towards resource r. The function f returns the number of concepts used, the number of words in the articles or the number of page views between May 2013 and June 2014 respectively and is an indicator of the importance a language attaches to a resource. The first term of the equation normalizes for the size of the language edition, e.g. how important it considers the resource to be compared to all other resources. The second term then normalizes for the general importance of the resource under consideration, by subtracting the average importance each other language attaches to said resource in average. This yields a value between -1 and $+1$ which is positive if the language gives more attention to the respective resource than one would expect. Using this approach, the absence of resource articles in some languages is not considered as a factor expressing little or no interest. The following formula incorporates this idea and uses a slightly different normalization factor, where all language editions are considered and not only those who define the respective resource:

$$bias_m(l,r) = \frac{\overbrace{f(l,r)}^{\text{attention twds. res. } r}}{\underbrace{\sum_{\bar{r}\in R} f(l,\bar{r})}_{\text{attention twds. all res.}}} - \underbrace{\frac{1}{|L|-1} \sum_{\bar{l}\in L\setminus l} \frac{f(\bar{l},r)}{\sum_{\bar{r}\in R} f(\bar{l},\bar{r})}}_{\text{normalized rel. attention of others}} \tag{3.3}$$

However, it has to be noted that for all analyzed data, the two formulas mentioned above correlate with $\rho > 0.95$, which is why for most of the calculations, only the second equation is used.

Finally, for a simpler approach, a third formula expressing bias was used which simply considers the attention a language attaches to a certain resource as compared to all other resources. This method was also included as it expresses the attraction of a language towards a resource, independently of the general popularity of the resource and hence allows for a different insight into a culture's preferences. A similar approach has also been used in [Bao+12], although they only used outlinks. Examples for all three bias calculations are shown when describing the findings in the following sections.

$$bias_s(l,r) = \frac{\overbrace{f(l,r)}^{\text{attention twds. res. } r}}{\underbrace{\sum_{\bar{r}\in R} f(l,\bar{r})}_{\text{attention twds. all res.}}} \tag{3.4}$$

The three different measures, i.e. the (i) word count, the (ii) outlink count and the (iii) view count were chosen to allow for the analysis of the problem from three slightly different perspectives. The word and outlink count can be interpreted as the importance attached by the editors of the Wikipedia to the respective resource. The latter measure was introduced due to some articles containing little descriptive text but a long list of links (for instance, the article on the Italian Wikipedia about the Italian cuisine[2]). The last measure, view counts, represents the attention of people viewing Wikipedia, which is more statistically stable as significantly more people view Wikipedia than edit it. However, the view counts do correlate with the outlink counts (spearman correlation $\rho_g = 0.90$, $\rho_m = 0.91$ and $\rho_s = 0.66$ for the three formulas with all $p \ll 0.001$), which is why the structural measure is considered a valid proxy for attention. This is an interesting finding, as previous research [OGR08] has shown that 90% of all contributions on Wikipedia come from a small proportion, namely 10% of all authors. It is therefore surprising that the amount of content on Wikipedia is so highly correlated with the interest in the pages. There are two possible explanations for this phenomenon: Either the creation of longer articles triggers interest manifested in higher view counts or more interesting articles (i.e. that are viewed more frequently) are being edited more often. The latter explanation seems more reasonable, as Brändle already noticed that an increase in the quality of an article is heavily influenced by its interest and relevance [Bräo5].

Additionally to the first hop outlinks, the second hop outlinks were analyzed separately. For this extended link-based dataset, a further constraint was added by sorting the number of concepts according to their generality, as defined by the tf/idf score. Concepts that appeared in the sub-graphs of multiple cuisines were considered less representative whereas concepts which only appeared in the sub-graph of one cuisine received a higher score. Therefore concepts like ”Pasta” or ”Pizza” were considered less representative for the Italian cuisine as they were used for other cuisines also, whereas ”Caciocavallo” (a special kind of Italian cheese) or ”Allis shad” (an Italian dry red wine) received considerably high tf/idf scores. The bias was re-calculated for the subset of concepts with tf/idf scores that exceeded a certain, arbitrarily chosen threshold. The resultant ratio can be interpreted as an indicator for the different language editions to not necessarily contain more concepts related to the respective cuisine but to describe it using a more fine-grained vocabulary.

[2] http://it.wikipedia.org/wiki/Cucina_italiana

For any of the above equations, the calculation of the bias gains validity the more cuisines a language defines. Therefore, a minimum threshold of four cuisines was chosen. This decision led to the Bosnian, Hungarian, Slovakian and Serbian Wikipedia to be excluded from the analysis.

In a more detailed analysis, the so-called self-focus bias was calculated for each language edition. Self-focus bias was defined by Brent Hecht and Darren Gergle as "information within a knowledge repository which is important and correct for the majority of its contributors, but differs from the information available in other, comparable repositories" [HG09]. In the context of cuisine articles, the different knowledge repositories are the different language editions of Wikipedia. Self-focus bias can then be defined as the tendency of a single edition of Wikipedia (e.g. the Italian Wikipedia) to focus on a particular resource (e.g. the Italian cuisine) that is related to the language which is predominantly spoken in the geographical area to which the resource belongs (e.g. Italy). In order to analyze this bias, the three different formulas were used. The self-focus bias was then calculated by the following equation:

$$\textit{self-focus-bias}(l) = \underbrace{\frac{1}{|R_{\text{own}}|} \sum_{r \in R_{\text{own}}} \textit{bias}(l,r)}_{\text{avg. bias twds. own res.}} - \underbrace{\frac{1}{|R_{\text{other}}|} \sum_{r \in R_{\text{other}}} \textit{bias}(l,r)}_{\text{avg. bias twds. other res.}} \tag{3.5}$$

This calculates the difference between average attention towards a culture's (l) own resources R_{own} (there can be multiple, e.g. both Austrian and German cuisine for the German speaking culture) and the average attention towards other resources R_{other}. $\textit{bias}(l,r)$ is the bias as calculated by one of the respective formulas.

Similarly, a regional bias was calculated, which is defined as the attention that a country pays towards neighboring countries as compared to countries that are geographically more distant. For instance, the regional bias of the Austrian Wikipedia would be represented by the difference between its affinity towards Italy, Germany, Hungary, etc. as its neighboring countries and for instance France, Finland, Spain, etc. as distant countries. The information about the country adjacency was retrieved from https://github.com/P1sec/country_adjacency.

$$\textit{regional-bias}(l) = \underbrace{\frac{1}{|R_{\text{adjacent}}|} \sum_{r \in R_{\text{adjacent}}} \textit{bias}(l,r)}_{\text{avg. bias twds. neighbors}} - \underbrace{\frac{1}{|R_{\text{other}}|} \sum_{r \in R_{\text{other}}} \textit{bias}(l,r)}_{\textit{avg.biastwds.others}} \tag{3.6}$$

The equation expresses the difference between the average attention towards resources of countries adjacent (R_{adjacent}) to a culture's home country l and the average attention towards other resources (R_{other}).

3.2.2 Evaluation Setup

Cultural similarity

The cultural similarity was compared to an external index[3] which is based on the European Social Survey (ESS)[4]. The ESS is a survey that measures a rich set of social, economic, political and cultural indicators on a regular basis. Jochen Roose used the results of the survey to compute the cultural similarity between countries [Roo10]. He used the symmetric Rsquared similarity measure since it quantifies the difference between two groups. In this method, the explained variance is divided by the total variance. If the computed result equals to one, the distributions of values in both groups are identical. If it is zero, the distributions in both groups are different and the two cultural groups are not considered similar. The index therefore contains a list of country pairs and their respective cultural similarity. This data was used for correlation with the cultural similarity as retrieved from Wikipedia.

As the survey measures cultural similarity on a much greater set of cultural dimensions, a second approach was taken to determine whether the applied method is reasonable and leads to plausible results. Therefore, a task was launched on the crowd-sourcing platform Crowdflower[5], where human workers were asked to decide which one of two cuisine pairs was more similar (e.g. they were asked whether the Ukrainian cuisine is more similar to the Russian cuisine than the Catalan cuisine to the Latvian cuisine). Out of the 450 possible combinations of cuisines, the most and least similar 15 combinations were taken. Then, each of the most similar pairs was compared to each of the most dissimilar pairs, resulting in 225 distinct comparisons. Each comparison had to be performed by at least 10 different workers and their choices were aggregated.

Cultural understanding

In order to evaluate one culture's understanding of another, two different external datasets were used for the evaluation. Firstly, the cultural similarity index by Jochen Roose was re-used, as it was assumed that cultural understanding is potentially explained by cultural similarity. If two countries share rather similar views and perceptions of the world, one could conclude that they understand each other well. Another possible explanation which encourages cultural understanding is the frequent exchange between two cultures. To model this exchange, data from the Global

[3] http://userpage.fu-berlin.de/~jroose/indexkultahn/main_indexwerte.htm
[4] http://www.europeansocialsurvey.org/
[5] http://www.crowdflower.com/

Bilateral Migration Database published by The World Bank[6] was used. These migration flows may also explain cultural understanding as local residents will be exposed to the foreign cultural values of the immigrants. For both external sources, Spearman rank correlations with the findings presented in this thesis were calculated.

Affinity and Bias

For the evaluation of cultural affinity expressed by the Wikipedia cuisine pages, two external sources were used as comparisons. The first one are the Eurovision Song Contest votings from 1975 to 2013. Unlike in sport competitions, no objective evaluation criterion exists for the appreciation of music. It reflects the national taste, native rhythm and primordial meaning [Yai95]. The Eurovision Song Contest allows to observe how different countries vote for the songs of other countries. Suspicions about tactical and political votings are as old as the song contest and e.g., looking at certain values which Greece and Cyprus usually assigned to each other (between 1993 and 2003 they voted for each other using the maximal number of points) supports these accusations. An alternative explanation for a systematic bias might be cultural and linguistic similarities that might manifest in a strong common music taste. Either way the results of this contest highlight the stable cultural relations between countries [Yai95]. In [SV06] the authors showed that countries tend to prefer or dislike songs of geographically nearby countries even when correcting for other factors such as culture and language. Therefore, it is likely that geographic preferences reflect political votings. Further they found that also religion and ethnicity might help to explain systematic voting biases. In [GT13] the authors attempt to control for artistic quality of songs and reveal voting biases between countries which expose the affinity between these countries.

The second external dataset consists of Foursquare check-ins in different countries and is used as proxy for modelling which cuisines are most prominent in which countries. To achieve this, simple counts of check-ins into restaurant types such as 'Italian restaurant' or 'German restaurant' were used to express the affinity of the originating culture (the region where the check-in occurred) to the target cuisine (the restaurant type). However, only six restaurant types were identified that could be directly mapped to countries and only very low counts were found (ranging from 1 to 1476). The data is additionally biased as only those Foursquare check-ins were used that were posted publicly on Twitter [Sil+14]. The comparison with the second dataset can therefore only be seen as a vague reference instead of a ground-truth evaluation. Unfortunately, there are not many datasets on cultural affinity available at such scale.

[6] http://data.worldbank.org/data-catalog/global-bilateral-migration-database

4 Results

4.1 Cultural Similarity

As mentioned in the introduction, cultural similarity was approximated by culinary similarity. To evaluate the similarity of two cuisines, the overlap of concepts that describe them was used. For instance, if both the French and Italian cuisine were described using the concepts ”Wine” and ”Cheese”, then those descriptions would be an indicator for a higher similarity between the two cuisines. The perceived similarity between two cuisines (e.g. the Italian and French cuisine) might be different depending on the perspective of the cultural background of the judge. Therefore, both the perceived local similarity (i.e. the overlap with which the Italian and French Wikipedia describe their cuisines) and the perceived global similarity (the similarity of the two cuisines as described by all Wikipedia editions) was analyzed.

How similar are the cuisines on Wikipedia?

The similarity was calculated using both a global, aggregated perspective as defined by all Wikipedia editions and a local perspective, where for each country pair, only the descriptions of their cuisines were considered. The results are shown in figure 4.1 for the global perspective and in figure 4.2 for the local perspective. The two perspectives correlate with $\rho = 0.58, p = 1.18e^{-46}$ (spearman rank correlation), which is an indicator for the entire Wikipedia community to reflect the similarity of cuisines fairly well compared to how it is perceived by the participating countries. Apart from the visible geographical bias, which will be treated in the next section, some findings are noteworthy: From the local perspective, the Ukrainian cuisine appears to be similar to many other cuisines, which was not expected. This might also be influenced by the Ukrainian Wikipedia describing many different cuisines, therefore giving the opportunity to be compared with other cultures. If this is the case, then the global perspective should give a better estimation of cultural similarity than the local one, as all combinations of cultures can be evaluated. This would also be supported by the heatmap derived from the local similarity which contains values of or close to zero for many fields.

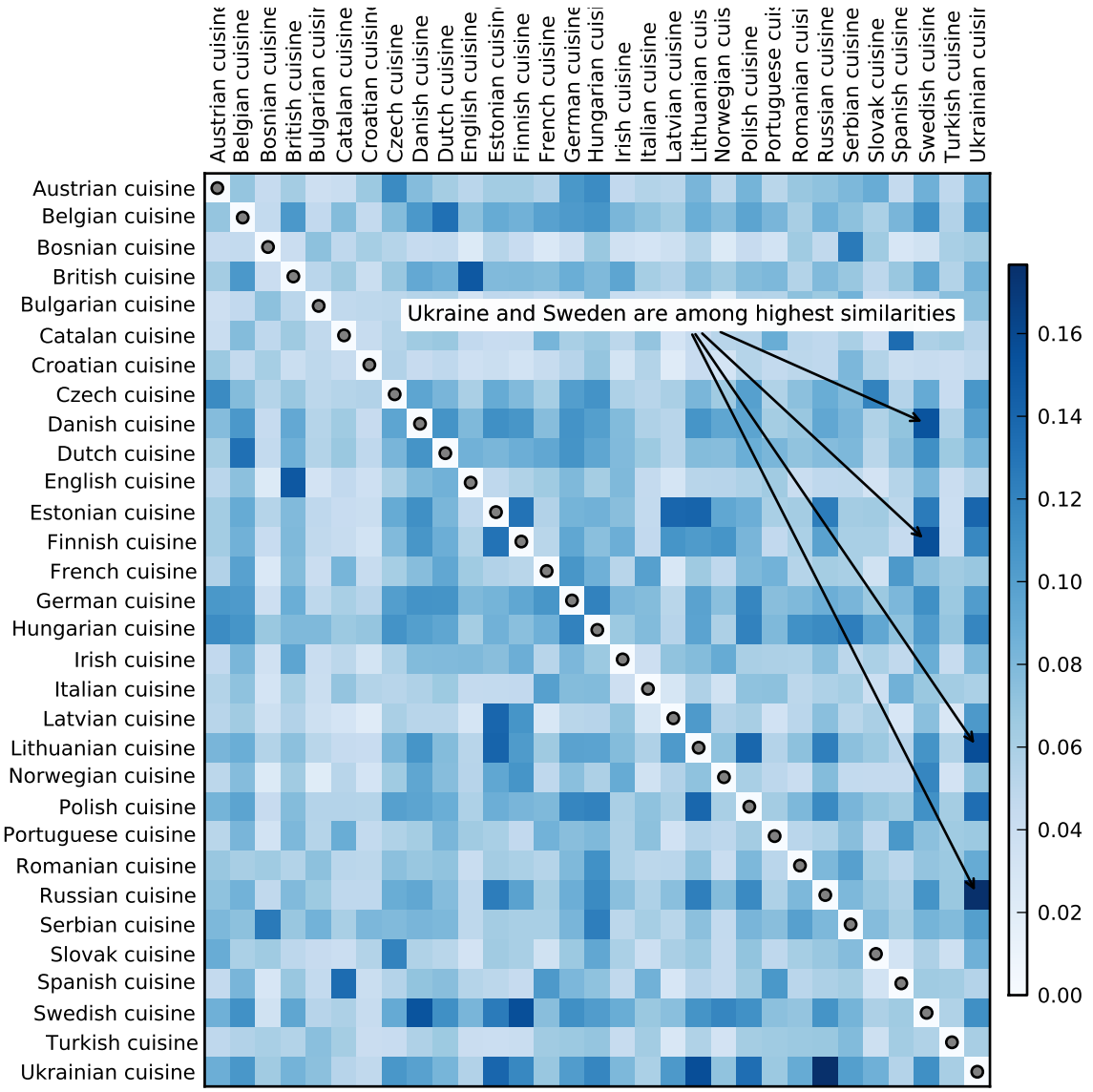


Figure 4.1: **How similar are different cuisines described on Wikipedia across all language editions?** The heatmap shows the culinary similarity from a global perspective. A higher similarity for geographically close cuisines is visible and will be analyzed later in this chapter.

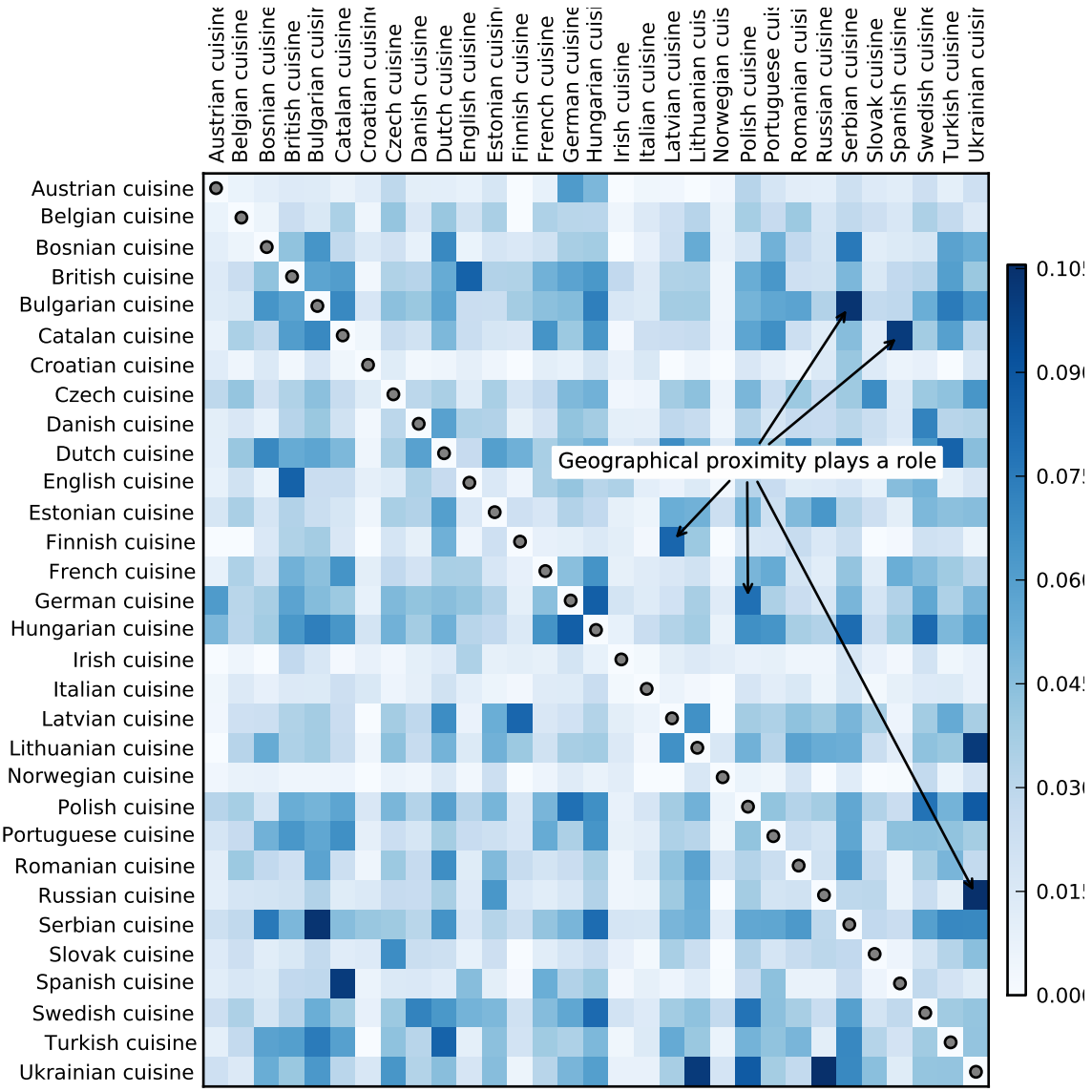


Figure 4.2: **How similar are different cuisines described on Wikipedia on their "native" editions?** The heatmap shows the culinary similarity from a local perspective. As not all language editions define each cuisine, the matrix contains many values of or close to zero, and generally lower similarity values are found. However, a certain geographical bias is also visible in this perspective.

Which are the most similar cuisines?

Using the local and global similarity ratings, a list of cuisines and their most similar counterparts is shown in Table 4.1. Clearly, a regional bias is visible and will be analyzed later. One can also see that the overlap is small for both methods, but greater and therefore more stable using the global perspective. This comes as no surprise, as more concepts are globally defined to describe a cuisine than by only a single article.

Cuisine	Local		Global	
Austrian	German	(0.06)	Czech	(0.12)
Belgian	Czech	(0.04)	Dutch	(0.13)
Bosnian	Serbian	(0.08)	Serbian	(0.13)
British	English	(0.08)	English	(0.15)
Bulgarian	Serbian	(0.10)	Serbian	(0.09)
Catalan	Spanish	(0.10)	Spanish	(0.14)
Croatian	Serbian	(0.04)	Serbian	(0.08)
Czech	Slovak	(0.07)	Slovak	(0.12)
Danish	Swedish	(0.07)	Swedish	(0.15)
Dutch	Turkish	(0.08)	Belgian	(0.13)
English	British	(0.08)	British	(0.15)
Estonian	Russian	(0.06)	Lithuanian	(0.14)
Finnish	Latvian	(0.08)	Swedish	(0.16)
French	Catalan	(0.06)	German	(0.11)
German	Hungarian	(0.09)	Hungarian	(0.12)
Hungarian	German	(0.09)	Serbian	(0.12)
Irish	English	(0.03)	British	(0.10)
Italian	Hungarian	(0.03)	French	(0.10)
Latvian	Finnish	(0.08)	Estonian	(0.14)
Lithuanian	Ukrainian	(0.10)	Ukrainian	(0.16)
Norwegian	Swedish	(0.03)	Swedish	(0.12)
Polish	Ukrainian	(0.09)	Lithuanian	(0.14)
Portuguese	Catalan	(0.07)	Spanish	(0.11)
Romanian	Dutch	(0.07)	Hungarian	(0.11)
Russian	Ukrainian	(0.11)	Ukrainian	(0.18)
Serbian	Bulgarian	(0.10)	Bosnian	(0.13)
Slovak	Czech	(0.07)	Czech	(0.12)
Spanish	Catalan	(0.10)	Catalan	(0.14)
Swedish	Hungarian	(0.08)	Finnish	(0.16)
Turkish	Dutch	(0.08)	Serbian	(0.08)
Ukrainian	Russian	(0.11)	Russian	(0.18)

Table 4.1: **Which are the most similar cuisines?** Most similar counterpart for each cuisine in both perspectives and their similarity measure calculated using Jaccard similarity.

Are cuisines more similar to other, geographically close cuisines than to distant ones?

The previous results already indicated the influence of a geographical factor to the similarity values. In order to evaluate the extent of this influence, a network of neighboring countries was retrieved from https://github.com/P1sec/country_adjacency. Using this data, the average similarity of each cuisine to both neighboring and distant cuisines was calculated using both the global and the local perspective. For instance, for the Portuguese cuisine, only the Spanish cuisine was considered a neighboring cuisine whereas all other European cuisines were not. The German cuisine, however, has many neighboring cuisines such as the Polish, the Austrian, the French, etc. Using this approach, a possible geographical influence of the similarity ratings was evaluated. It can be assumed that geographical distances play an important role for cuisines to influence each other (although other factors such as trade routes, religion etc. are also possible influences). This assumption is supported by the findings, as presented in in Table 4.2. Apart from a few exceptions when using the local perspective (the Bosnian, Dutch, Finnish and Hungarian cuisine), all cuisines show to be more similar to their neighboring cuisines than to others. For the global perspective this measure is relatively stable and indicates that each cuisine seems to be roughly 1.5 times as similar to its neighbors than to foreign cuisines (with a standard deviation of 0.2).

Cuisine	Local perspective	Global perspective
Austrian cuisine	+2.65	+1.51
Belgian cuisine	+1.68	+1.41
Bosnian cuisine	+0.51	+1.26
British cuisine	+1.12	+1.22
Bulgarian cuisine	+1.63	+1.42
Catalan cuisine		
Croatian cuisine	+1.82	+1.36
Czech cuisine	+1.56	+1.51
Danish cuisine	+2.43	+1.62
Dutch cuisine	+0.98	+1.58
English cuisine		
Estonian cuisine	+1.54	+1.70
Finnish cuisine	+0.79	+1.73
French cuisine	+1.47	+1.64
German cuisine	+1.51	+1.36
Hungarian cuisine	+0.82	+1.14
Irish cuisine	+3.09	+1.50
Italian cuisine	+1.05	+1.23
Latvian cuisine	+1.74	+2.00
Lithuanian cuisine	+1.64	+1.51
Norwegian cuisine	+1.14	+1.77
Polish cuisine	+1.49	+1.48
Portuguese cuisine	+1.36	+1.72
Romanian cuisine	+1.34	+1.42
Russian cuisine	+2.25	+1.50
Serbian cuisine		
Slovak cuisine	+2.01	+1.69
Spanish cuisine	+3.18	+1.86
Swedish cuisine	+1.02	+1.67
Turkish cuisine	+2.08	+1.35
Ukrainian cuisine	+1.74	+1.36
Average	+1.63	+1.52
Std.Dev	0.64	0.20

Table 4.2: **Which cuisines are most influenced by their geographical neighbors?** Ratio between the similarity of neighboring cuisines and distant cuisines. A value of e.g. 2 indicates that neighboring cuisines are twice as similar as distant ones. As visible, with the exception of the Bosnian, Dutch, Finish and Hungarian cuisine from the local perspective, all cuisines do seem to be more similar to geographically close counterparts than to others. The global perspective is clearly more stable, most likely due to the fact that more cuisine pairs can be compared, as all of them are defined on an aggregated level but only a few for each local perspective.

4.1.1 Evaluation

In order to evaluate how well the culinary similarity as described on Wikipedia approximates the cultural similarity between two countries, the dataset from the European Social Survey (ESS) was used and spearman rank correlation was calculated. For the local perspective, no significant correlation could be found, and for the global perspective, a correlation with $\rho = 0.25, p = 0.0002$ was detected. Figure 4.3 shows the correlation dependent on the k most similar countries according to ESS. From the results, two findings are noteworthy: Firstly, it can be concluded that the culinary similarity on Wikipedia can only contribute to a certain extent to calculating cultural similarity between countries. However, if further cultural dimensions are added to complement the narrow view of culinary aspects, the method will most likely perform better. Secondly, it is clear that the global description (by all Wikipedia editions) is a better indicator for similarity and hence should be used instead of the local descriptions. As mentioned earlier, this also allows for sparse datasets (where not every language edition describes each cuisine) to be analyzed.

In order to evaluate whether the method itself is a plausible description of culinary similarity, the results of a crowd-sourcing task were compared to the 15 most similar and 15 most dissimilar pairs. The human workers had to choose for each of the 225 combinations, which of the cuisine pairs was more similar. Their choices were aggregated and compared to the rankings of pairs according to their similarity. Out of the 225 combinations, only one was perceived differently by the human workers (they considered the Croatian cuisine more similar to the Latvian cuisine than the Estonian to the Ukrainian), resulting in 99.56% correctly ranked pairs. These findings show that the applied method itself is capable of describing culinary similarity.

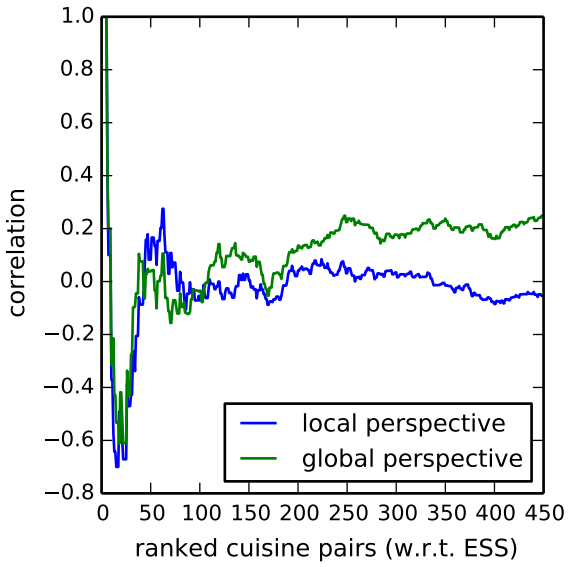


Figure 4.3: **How well does the culinary similarity approximate cultural similarity?** The Spearman rank correlation of the local and global perspective with the k most similar countries according to ESS on the x-axis. Both perspectives seem to rank the few most similar countries correctly, but when increasing k further, a strong negative correlation is visible, before the correlations converge to a low, but stable value.

4.2 Cultural Understanding

In this section, the analysis of the understanding of one country's culture by another culture is presented. The concept of cultural understanding is related to cultural similarity and bases upon the idea, that if culture A understands culture B very well, then they will both describe the resources relevant to culture B in a similar manner. To map this idea to cuisines, one could argue that the Germans understand the Italians well if they can accurately describe the Italian cuisine (i.e. similar to how the Italians themselves would describe it). For the analysis, both the first hop dataset (consisting of the outgoing links of all cuisine articles) and the second hop dataset (consisting of all outgoing links from all pages linked from the cuisine articles) were analyzed separately. For instance, if the article describing the Spanish cuisine links to the article "Paella", then the concept "Paella" would be in the first hop dataset describing the cuisine. If the article "Paella" further links to "Seafood", then the concept "Seafood" would be added to the second hop dataset. So the second hop dataset is an extension of the first hop dataset which contains more concepts.

Which concepts define the different cuisines on a global, aggregated level?

As a cuisine is defined in a different way on each language edition of Wikipedia, the question arises whether a globally valid description of a cuisine is possible. Using the first hop dataset, concepts which were used to describe each cuisine (outgoing links in the cuisine articles) were extracted and sorted according to the number of Wikipedae on which they were used to describe the cuisine. If the concept "Pizza" is for instance used on 5 different Wikipedae to describe the Italian cuisine, it would be ranked with a value of 5 for the Italian cuisine. Table 4.3 shows the 5 most frequently used concepts for each cuisine and Figure 4.4 visualizes the culinary concepts on a map. Although the data contains noise (mainly geographical entities and links to other cuisines), a rather good descriptor of each cuisine is given as an aggregated, common perspective. The noise is introduced by other terms being heavily used on the cuisine article pages, such as "...originates from [geographic entity]" or "...is influenced by [other cuisine]". One can see that more popular cuisines receive higher values which means that the number of Wikipedae which use a certain concept to describe a cuisine does not necessarily reflect the homogeneity of the description but also the popularity of the cuisine. Later sections will consider this factor and normalize for the number of Wikipedae that describe the cuisine (e.g. naturally the Italian and French cuisine receive high values, but not necessarily because they are so commonly defined but because many language versions have articles about them - i.e. they are popular).

Interestingly, most of the cuisines with the smallest relative overlap of concepts (Swedish, Romanian, Slovak and Estonian cuisine) are from countries which Wikipedia editions are also rather small (all less than 250.000 articles, except for the Swedish Wikipedia). However, the cuisines with the biggest overlap (Bosnian, Russian Portuguese, Turkish and German cuisine) are not related to the largest Wikipedae (English, Dutch, German and Swedish).

Also noteworthy is the fact that even when using the top 10 concepts, the Italian cuisine is the only one were no noise is introduced (only food-related concepts appear) and that no concept is used on all articles describing a single cuisine.

Figure 4.4: **Which are the most common terms on Wikipedia used to describe European cuisines?** The two culinary concepts used to describe European cuisines across most language editions.

Austrian cuisine (16)	**Belgian cuisine (13)**	**Bosnian (6)**	**British cuisine (11)**
en/Austria-Hungary (12)	en/Lambic (9)	en/Carrot (5)	en/Sandwich (7)
en/Austria (11)	en/French cuisine (9)	en/Tomato (5)	en/Scotland (7)
en/Sachertorte (10)	en/Belgium (9)	en/Potato (5)	en/Bread (7)
en/Vienna (9)	en/French fries (9)	en/Garlic (5)	en/Yorkshire pudding (7)
en/Wiener Schnitzel (9)	en/Beer (8)	en/Herzegovina (5)	en/Tomato (7)
Bulgarian cuisine (10)	**Catalan cuisine (11)**	**Croatian cuisine (9)**	**Czech cuisine (13)**
en/Bulgaria (9)	en/Catalonia (10)	en/Croatia (7)	en/Czech Republic (9)
en/Yogurt (8)	en/Pa amb tomàquet (7)	en/Karlovačko (6)	en/Beer (9)
en/Tarator (6)	en/Fuet (7)	en/Hungarian cuisine (6)	en/Austrian cuisine (9)
en/Balkans (5)	en/Aioli (7)	en/Italian cuisine (6)	en/Becherovka (8)
en/Banitsa (5)	en/Botifarra (7)	en/Maraschino (6)	en/Knödel (8)
Danish cuisine (13)	**Dutch cuisine (13)**	**English cuisine (8)**	**Estonian cuisine (10)**
en/Denmark (11)	en/Cheese (11)	en/England (7)	en/German cuisine (7)
en/Smørrebrød (11)	en/Heineken Internation(9)	en/Fish and chips (7)	en/Kama (food) (6)
en/Beer (8)	en/Potato (9)	en/Full breakfast (6)	en/Potato (6)
en/Danish pastry (7)	en/Netherlands (9)	en/Sandwich (5)	en/Beer (6)
en/Carlsberg Group (7)	en/Edam cheese (9)	en/Italian cuisine (5)	en/Estonia (6)
Finnish cuisine (16)	**French cuisine (21)**	**German cuisine (16)**	**Hungarian cuisine (15)**
en/Finland (12)	en/France (18)	en/Potato (13)	en/Goulash (13)
en/Reindeer (9)	en/Cheese (15)	en/Germany (13)	en/Tokaji (11)
en/Kalakukko (9)	en/Foie gras (14)	en/Sauerkraut (12)	en/Hungary (10)
en/Milk (8)	en/Wine (14)	en/Bavaria (12)	en/Fisherman's soup (9)
en/Rubus chamaemorus (8)	en/Provence (14)	en/Beer (12)	en/Unicum (9)
Irish cuisine (12)	**Italian cuisine (20)**	**Latvian cuisine (9)**	**Lithuanian cuisine (14)**
en/Guinness (10)	en/Pizza (17)	en/Latvia (7)	en/Potato (12)
en/Republic of Ireland (8)	en/Pasta (16)	en/Potato (6)	en/Cepelinai (10)
en/Potato (8)	en/Parmigiano-Reggiano (14)	en/Cheese (6)	en/Polish cuisine (9)
en/Irish stew (8)	en/Tiramisu (14)	en/Beer (5)	en/Caraway (9)
en/Irish coffee (7)	en/Tortellini (13)	en/Baltic Sea (5)	en/Lithuania (9)
Norwegian cuisine (12)	**Polish cuisine (16)**	**Portuguese cuisine (13)**	**Romanian cuisine (13)**
en/Norway (10)	en/Poland (13)	en/Port wine (12)	en/Romania (9)
en/Lutefisk (8)	en/Bigos (13)	en/Garlic (10)	en/Ciorbă (8)
en/Brunost (7)	en/Vodka (12)	en/Coriander (10)	en/Mămăligă (7)
en/Cheese (7)	en/Sauerkraut (11)	en/Portugal (10)	en/Tuică (7)
en/Atlantic cod (7)	en/Beer (11)	en/Olive oil (10)	en/Transylvania (7)
Russian cuisine (15)	**Serbian cuisine (10)**	**Slovak cuisine (12)**	**Spanish cuisine (16)**
en/Vodka (13)	en/Rakia (9)	en/Bryndzové halušky (11)	en/Chorizo (13)
en/Kvass (13)	en/Serbia (6)	en/Slovakia (7)	en/Gazpacho (13)
en/Honey (12)	en/Baklava (6)	en/Goulash (7)	en/Olive oil (12)
en/Borscht (12)	en/Ajvar (6)	en/Sauerkraut (6)	en/Paella (12)
en/Russia (10)	en/Slivovitz (5)	en/Bryndza (6)	en/Garlic (11)
Swedish cuisine (14)	**Turkish (18)**	**Ukrainian (10)**	
en/Sweden (12)	en/Baklava (16)	en/Borscht (8)	
en/Potato (8)	en/Yogurt (15)	en/Kvass (7)	
en/Bread (8)	en/Ottoman Empire (13)	en/Vodka (6)	
en/Reindeer (7)	en/Turkish delight (13)	en/Ukraine (6)	
en/Crisp bread (7)	en/Kebab (13)	en/Beer (5)	

Table 4.3: **Which concepts are used to describe European cuisines from a global perspective?** Most frequent concepts related to each cuisine and the number of language editions which used them to describe it. The numbers in the headers indicate the number of Wikipedae that describe each cuisine. Interestingly, most of the cuisines with the smallest relative overlap of concepts (Swedish, Romanian, Slovak and Estonian cuisine) are mostly from countries which Wikipedia editions are also rather small. However, the cuisines with the biggest overlap (Bosnian, Russian, Portuguese, Turkish and German cuisine) are not related to the largest Wikipedae (English, Dutch, German and Swedish). Also noteworthy is the fact that even when using the top 10 concepts, the Italian cuisine is the only one were no noise is introduced (only food-related concepts appear).

How does the local perception of a cuisine differ from its global, aggregated one?

Does the Spanish Wikipedia define the Spanish cuisine differently than the other Wikipedae? How are these differences distributed? In order to answer these questions, the two perceptions were identified by using the outlinks of both, the article about each cuisine from the respective Wikipedia as a ”local” perception, and the set of outlinks from the articles on all other language editions. The overlap between each ”external” and ”local” set of concepts was then measured using Jaccard similarity. For instance, all concepts from all but the Spanish Wikipedia describing the Spanish cuisine (the global perception) were compared to the concepts on the Spanish version (the local perception).

The resulting overlap ratios are shown in Table 4.4 and their distribution in Figure 4.5. The data indicates that, different to the similarity relation, there is a generally very low overlap between the global and the local perspective as many different concepts are used. This can be interpreted as each culture having a unique view of their cultural heritage that is not shared by others. For this kind of analysis, the first hop dataset seems to generally be better, as greater overlaps are found (most likely due to the fact that a lot of noise is introduced using the second hop links).

Cuisine	hop1 (mean)	hop1 (std)	hop2 (mean)	hop2 (std)
Austrian	0.0608	0.0576	0.0333	0.0133
Belgian	0.0847	0.0428	0.0426	0.0148
Bosnian	0.1267	0.0393	0.0658	0.0173
British	0.1143	0.0979	0.0326	0.0203
Bulgarian	0.0468	0.0230	0.0504	0.0101
Catalan	0.0696	0.0529	0.0310	0.0192
Croatian	0.0451	0.0401	0.0537	0.0126
Czech	0.0875	0.0587	0.0499	0.0185
Danish	0.0577	0.0451	0.0426	0.0202
Dutch	0.0675	0.0334	0.0506	0.0148
English	0.0397	0.0193	0.0151	0.0084
Estonian	0.0512	0.0297	0.0492	0.0089
Finnish	0.0903	0.0989	0.0459	0.0065
French	0.0751	0.0642	0.0205	0.0117
German	0.0763	0.0668	0.0283	0.0152
Hungarian	0.0489	0.0298	0.0317	0.0128
Irish	0.0687	0.0426	0.0354	0.0132
Italian	0.0626	0.0383	0.0374	0.0174
Latvian	0.0715	0.0371	0.0498	0.0106
Lithuanian	0.0899	0.0478	0.0549	0.0135
Norwegian	0.0703	0.0277	0.0399	0.0053
Polish	0.0720	0.0313	0.0361	0.0125
Portuguese	0.1062	0.0768	0.0489	0.0118
Romanian	0.0481	0.0221	0.0491	0.0134
Russian	0.1137	0.0423	0.0513	0.0097
Serbian	0.0416	0.0370	0.0393	0.0118
Slovak	0.0520	0.0313	0.0448	0.0059
Spanish	0.0478	0.0367	0.0208	0.0124
Swedish	0.0618	0.0451	0.0423	0.0169
Turkish	0.0865	0.0413	0.0581	0.0155
Ukrainian	0.0674	0.0436	0.0577	0.0120

Table 4.4: **How well do other versions of Wikipedia describe cuisines as compared to the description on their "native" language edition?** Jaccard similarity between the concepts defined by the cuisine's representative language edition and the concepts defined globally by all other language editions for the hop1 and hop 2 dataset.

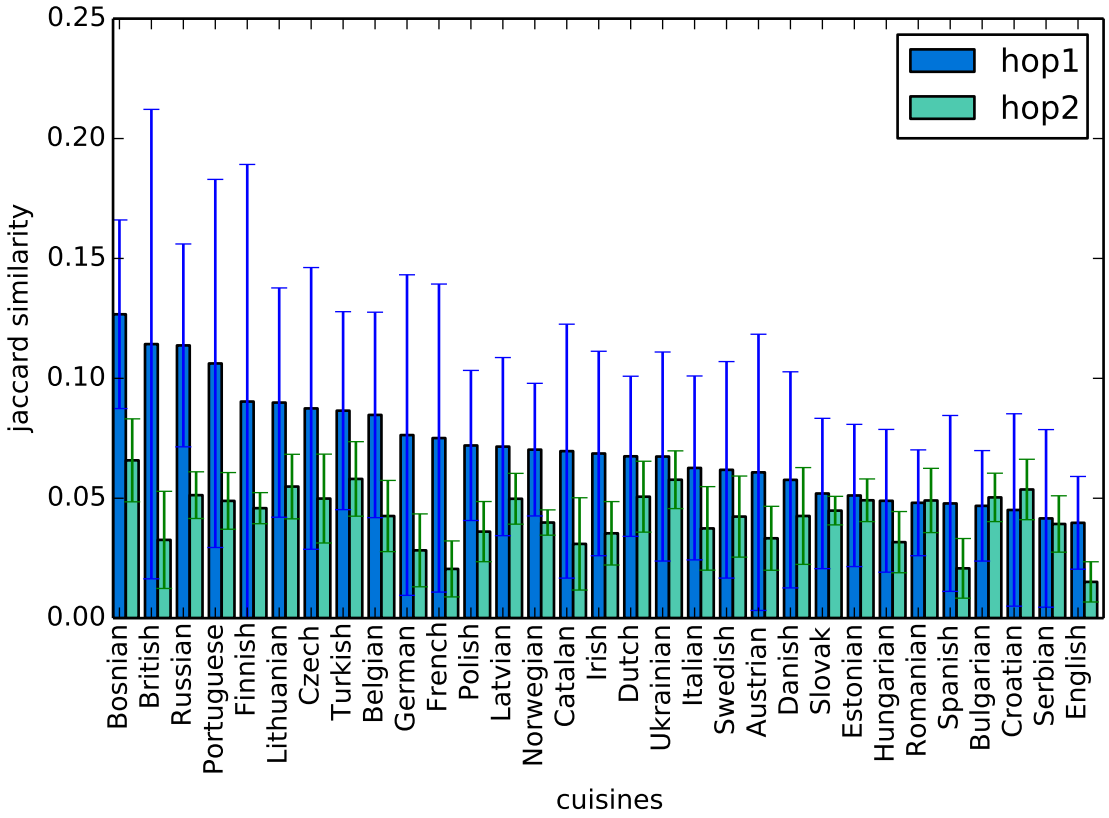


Figure 4.5: **Which cuisine is most accurately described by foreign European cultures?** Distribution of overlaps between the local (internal) perspective of a cuisine and the global (external) perspective as defined by the articles on other Wikipedae for the hop1 and hop2 dataset.

How does the distribution of cultural understanding look like?

Apart from the question whether the internal perception of a cuisine differs from the external perception, it is also interesting to explore the variation of the perception of each cuisine among the different language editions. For all possible language combinations, the overlap between the concepts defined by each language edition for the cuisine was calculated using Jaccard similarity. Figure 4.6 shows the distribution of these similarities. More prominent cuisines such as the Italian, French or Turkish cuisine are more commonly defined than less-known cuisines such as the Bulgarian or Bosnian. The plots also indicate that only a very small set of language editions define a cuisine using more or less the same concepts whereas the majority uses different terms. This supports the idea that there is no global definition of cuisines, but that their descriptions are highly influenced by the culture which defines them.

How heterogeneous is the vocabulary used to describe different cuisines?

Does the Spanish Wikipedia use a more homogenous set of concepts to describe different cuisines than the Italian Wikipedia? Do the used concepts within one language edition differ significantly between different cuisines? Similarly to the previous question, the similarity of cuisine articles is measured. This time, however, not the difference of a single cuisine between different language editions is analyzed, but the difference between multiple cuisines within one language. Figure 4.7 shows the resulting distributions. Very different concepts are used to describe the different cuisines, which explains the small overlaps of less than 0.1. As expected, the bigger Wikipedae show a higher similarity and hence describe different cuisines using partially the same concepts. This might be due to the fact that the average length of a cuisine article on said Wikipedae is longer and therefore naturally leads to a greater overlap. However, certain exceptions of this finding are interesting: The Dutch Wikipedia, for instance, although being the second largest Wikipedia analyzed, seems to describe fewer cuisines, but with a greater variance of concepts. The same holds true for the French Wikipedia, which is the 4^{th} largest encyclopedia.

How well do different Wikipedae describe the cuisine of another culture?

Using the same overlap measure, the cultural understanding that each culture (as represented by the different Wikipedae) has for each cuisine can be modeled. This is an asymmetric measure, as, for example, the French Wikipedia can describe the German cuisine in line with the German view, but the German Wikipedia does not need to know anything about the French cuisine. In this case, one could argue that the French people would have a better understanding of the German cuisine than

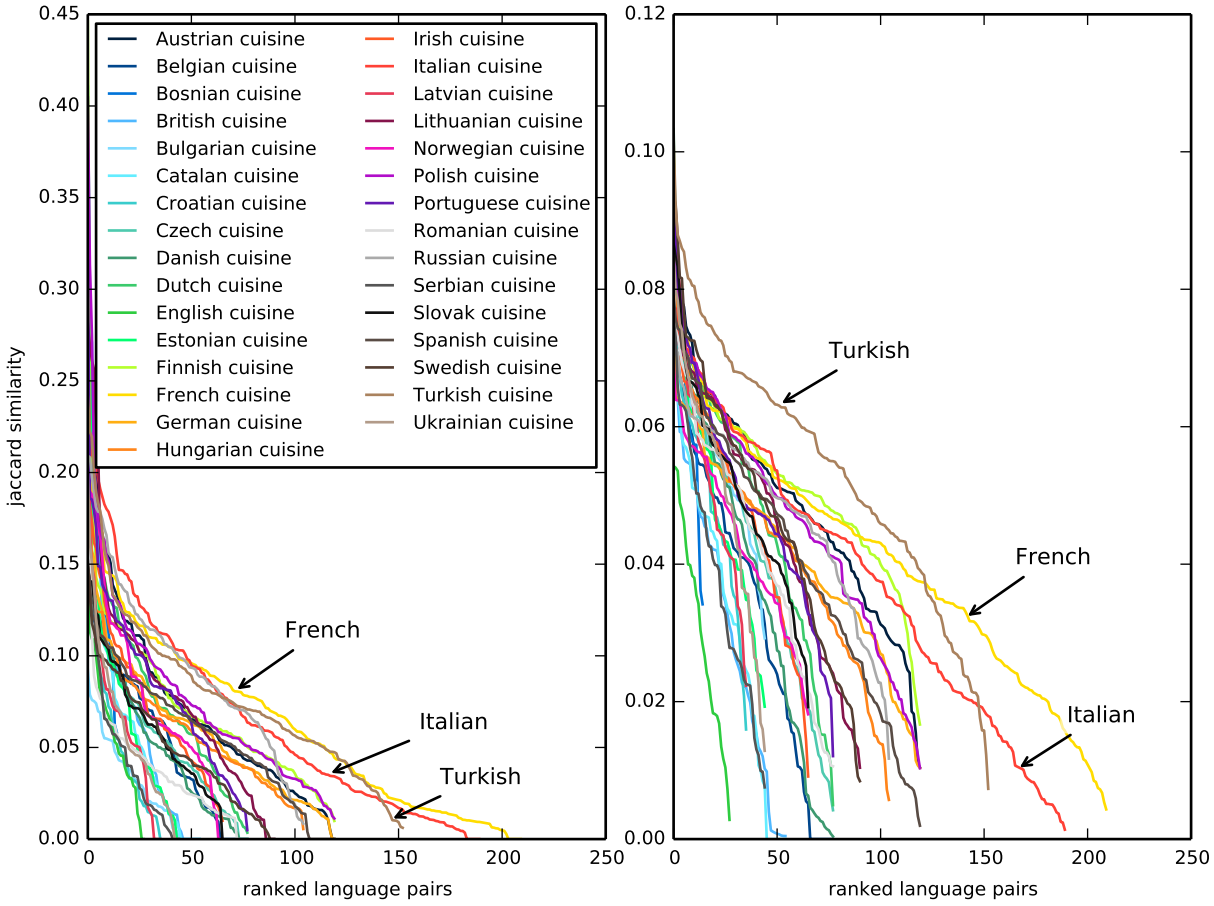
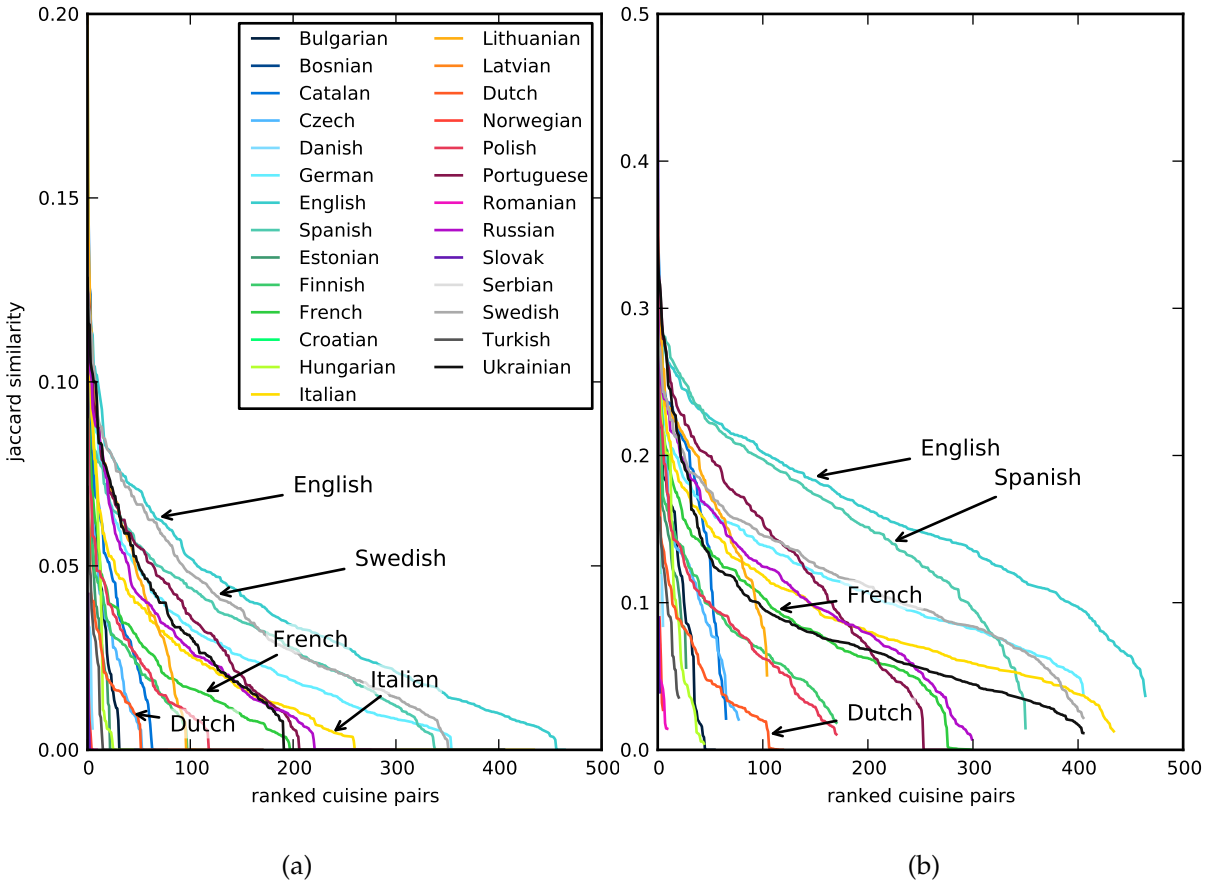


Figure 4.6: **Which are the cuisines with the most uniform description?** Ranked distribution of pairwise similarities between different language editions for each cuisine using the first hop (a) and second hop (b) dataset. The lines end at different points, as the cuisines are covered by a different number of language editions.

vice versa. This is an important point, as the Jaccard similarity is actually symmetric. However, due to defining the "true representation" of a cuisine to be the one presented on its "native" Wikipedia, one can calculate an asymmetric value for the cultural understanding. Figure 4.8 shows a heatmap of the overlap ratios. However, the matrix is rather sparse, which can be interpreted as missing cultural understanding. If, for instance, the Hungarian Wikipedia does not have an article about the French cuisine, one could argue that the Hungarians are either not interested in cuisines altogether (see comparisons in chapter 3.1) or they are not interested in the French cuisine. In both cases, this might be an indicator of missing cultural understanding,



(a) (b)

Figure 4.7: **Which languages on Wikipedia contain the broadest knowledge about different cuisines?** Ranked distribution of pair-wise similarities between different cuisines for each language using the first hop (a) and second hop (b) dataset. The lines end at different points, as every language only defines a different number of cuisines.

at least in the culinary domain.

Furthermore, there are some findings that are noteworthy: For instance, the data suggests that the Catalan Wikipedia represents the Portuguese cuisine far more accurately than the Spanish cuisine. Also the good understanding of the Finnish cuisine by the French Wikipedia is interesting.

Table 4.5 shows the five best understood foreign cuisines per Wikipedia language edition.

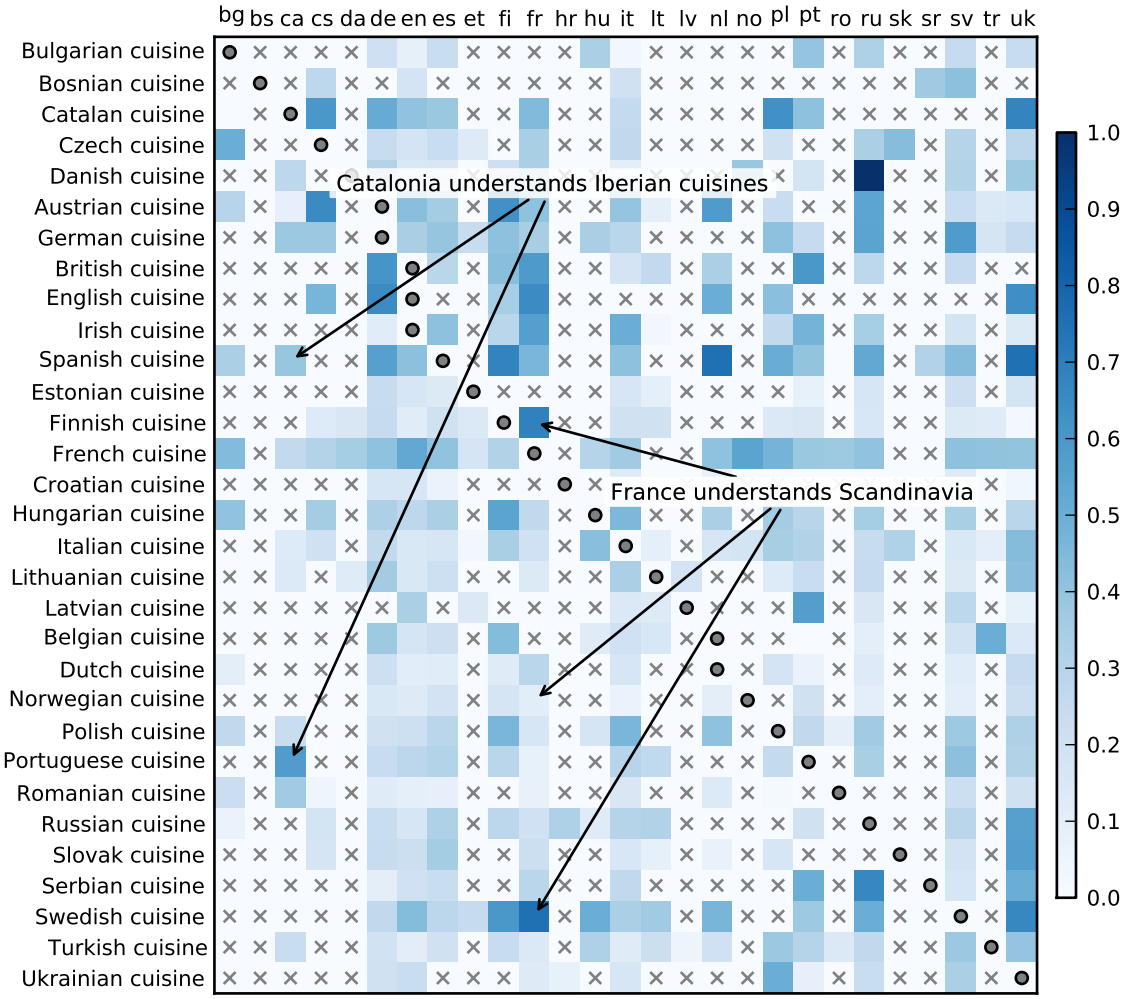


Figure 4.8: **How well do different Wikipedae describe the cuisine of another culture?** The heatmap shows the overlap of concepts used to describe a cuisine for each cultural group with the group that is associated with the cuisine. The dots therefore mark such associations (e.g. the Bulgarian Wikipedia with the Bulgarian cuisine). The crosses represent missing data, i.e. cuisines for which no article exists in the respective language. It has to be noted that the missing of these articles might also be an indicator for (missing) cultural understanding. The plot is based upon the concepts from the first hop dataset.

Bulgarian Wikipedia	**Bosnian Wikipedia**	**Catalan Wikipedia**	**Czech Wikipedia**
Hungarian (0.105)		Portuguese (0.316)	French (0.141)
Romanian (0.076)		Turkish (0.099)	Bosnian (0.127)
Dutch (0.063)		Danish (0.090)	Slovak (0.123)
Polish (0.057)		Spanish (0.086)	Russian (0.094)
Austrian (0.040)		Polish (0.084)	Austrian (0.062)
Danish Wikipedia	**German Wikipedia**	**English Wikipedia**	**Spanish Wikipedia**
Lithuanian (0.087)	British (0.302)	French (0.262)	German (0.264)
French (0.035)	Lithuanian (0.211)	Austrian (0.209)	French (0.184)
Finnish (0.032)	Belgian (0.167)	Portuguese (0.185)	Catalan (0.154)
	Czech (0.159)	German (0.157)	Turkish (0.152)
	Russian (0.156)	Swedish (0.152)	Austrian (0.151)
Estonian Wikipedia	**Finnish Wikipedia**	**French Wikipedia**	**Croatian Wikipedia**
Finnish (0.061)	Swedish (0.133)	Finnish (0.446)	Russian (0.169)
Latvian (0.053)	Belgian (0.109)	British (0.281)	Ukrainian (0.024)
German (0.022)	Russian (0.102)	Irish (0.142)	Bulgarian (0.000)
French (0.016)	Italian (0.080)	Russian (0.135)	
Czech (0.016)	French (0.076)	German (0.132)	
Hungarian Wikipedia	**Italian Wikipedia**	**Lithuanian Wikipedia**	**Latvian Wikipedia**
Turkish (0.179)	Russian (0.137)	Russian (0.139)	Lithuanian (0.050)
German (0.124)	French (0.122)	Turkish (0.116)	Turkish (0.026)
Russian (0.093)	British (0.121)	Portuguese (0.082)	
Polish (0.078)	Dutch (0.108)	British (0.077)	
French (0.048)	Czech (0.102)	Belgian (0.075)	
Dutch Wikipedia	**Norwegian Wikipedia**	**Polish Wikipedia**	**Portuguese Wikipedia**
Austrian (0.109)	Danish (0.143)	Dutch (0.127)	Irish (0.130)
Turkish (0.081)	Italian (0.091)	German (0.099)	Norwegian (0.118)
Swedish (0.075)	French (0.012)	Irish (0.090)	Lithuanian (0.109)
Romanian (0.063)		Finnish (0.084)	Spanish (0.105)
Polish (0.058)		Portuguese (0.081)	Catalan (0.101)
Romanian Wikipedia	**Russian Wikipedia**	**Slovak Wikipedia**	**Serbian Wikipedia**
Polish (0.027)	Ukrainian (0.150)	Czech (0.220)	Bosnian (0.171)
Turkish (0.009)	Lithuanian (0.146)	Italian (0.026)	Spanish (0.014)
French (0.006)	British (0.146)		
	Austrian (0.119)		
	Italian (0.112)		

Table 4.5: **Which foreign cuisines are best understood by each language edition?** The five best understood foreign cuisines per Wikipedia according to the concept overlap (displayed in brackets) of the cuisine description by the respective Wikipedia and the "native" Wikipedia of the cuisine. As some Wikipedae only describe a small number of European cuisines, only fewer entries are shown.

4.2.1 Evaluation

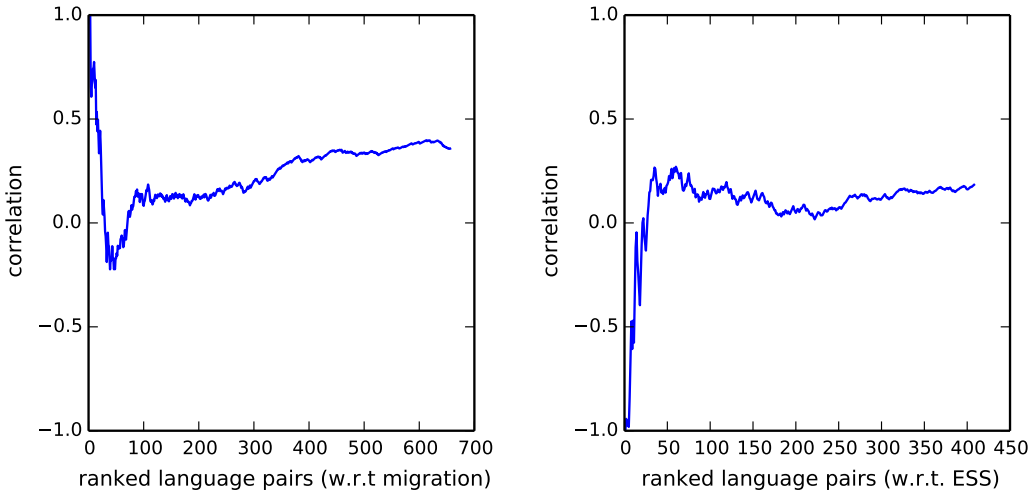
Can one culture's understanding of another culture be approximated by their representation of cuisines on Wikipedia?

In order to answer this question, both an index based on the European Social Survey (ESS) [Roo10] and data from the Global Bilateral Migration Database (migration) were compared to the findings derived from Wikipedia (wiki). These two sources were selected according to the assumption that cultural understanding is influenced by both cultural similarity (i.e. countries that are similar are likely to understand each other) and migration flows (i.e. countries that are the target destination of immigrants are assumed to understand the source country of those immigrants rather well). As both external sources were created using a different set of countries than the dataset based on Wikipedia, only understandings between pairs of countries that exist in both the external source and the findings were considered. The resulting lists of pairs were then ranked by their Jaccard similarity and compared using Spearman rank correlation. The results are shown in Table 4.6 and indicate a low, but significant correlation. This means that Wikipedia is capable of explaining cultural understanding to some extent. Noteworthy is also the low, but significant correlation between the external data sources, which suggests that migration and cultural similarity are, at least to some degree, related. Also, considering these low correlations, it is questionable if a real ground truth dataset for cultural understanding exists. In case none of the compared sources represent such a ground truth, the effectiveness of the approach presented in this thesis is likely to be better than what one would conclude from the evaluation.

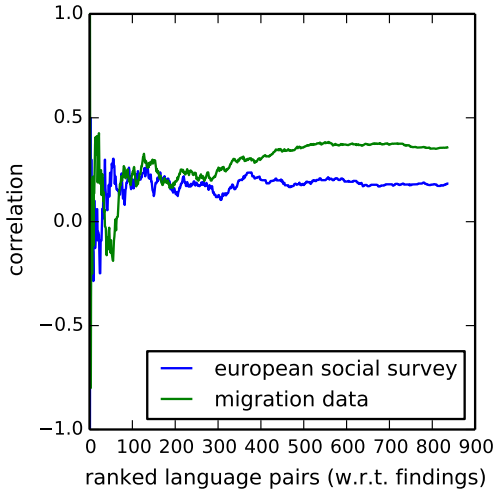
Additionally, the Jaccard index was calculated using only the top k pairs according to all three sources, as presented in Figure 4.9. To gain further insights into the understanding of cultures from to perspective of each individual country, these correlations are listed in Table 4.7.

Pair	ρ	p-value
wiki - ESS	0.18	(0.00019)
wiki - migration	0.36	($1.74e^{-22}$)
ESS - migration	0.22	($8.28e^{-6}$)

Table 4.6: **Does the culinary understanding on Wikipedia correlate with external data?** The correlation values between the findings derived from Wikipedia (wiki), the values of an index derived from the European Social Survey (ESS) and the migration data from the World Bank (migration). As visible, the correlation is significant, but low, and also the two external sources do not highly correlate. This can be interpreted as both migration and culinary understanding to explain cultural understanding only to some degree.



(a) (b)



(c)

Figure 4.9: **A detailed view of the correlation with the external data:** The spearman rank correlation values between the findings derived from Wikipedia (wiki), the values of an index derived from the European Social Survey (ESS) and the migration data from the World Bank (migration) for the top k ranked language pairs with respect to all three sources. Figure (a) shows that the data from Wikipedia highly correlates with the migration data for the biggest migration flows (the language editions of those countries where most people migrate to understand the culture from which the immigrants come from best). Surprisingly, quite the opposite is true for the index of cultural similarity derived from ESS as visible in Figure (b). In fact, the language editions of Wikipedia seem to describe the cuisines of other countries that are highly similar to their origin country according to the index less accurately than those to which they are less similar. Finally, the correlation for the top k language pairs ranked according to the cultural understanding as found on Wikipedia is shown in (c).

Wikipedia Edition	ρ_{ESS}	$\rho_{\text{migration}}$
Bulgarian	-0.08	-0.08
Bosnian		
Catalan		
Czech	-0.09	-0.09
Dansk	+0.15	+0.15
German	-0.04	-0.04
English	-0.08	-0.08
Spanish	+0.29	+0.29
Estonian	+0.17	+0.17
Finnish	-0.17	-0.17
French	+0.06	+0.06
Croatian		
Hungarian	+0.10	+0.10
Italian	+0.24	+0.24
Lithuanian		
Latvian		
Dutch	+0.04	+0.06
Norwegian	-0.08	-0.08
Polish	+0.29	+0.29
Portuguese	-0.11	-0.11
Romanian		
Russian	+0.30	+0.30
Slovak	+0.35	+0.35
Serbian		
Swedish	-0.09	-0.09
Turkish	-0.19	-0.19
Ukrainian	+0.40	+0.40
Average	+0.07	+0.26
Std.Dev.	0.18	0.16

Table 4.7: **For which language edition does the culinary understanding align best with external data?** The correlation values (Spearman rank correlation) between the findings presented in this thesis (wiki), the values of an index derived from the European Social Survey (ESS) and the migration data from the World Bank (migration) per Wikipedia edition. In this case, the correlation with the migration information is significantly higher than with the information of cultural similarity. Not all cells contain values as some countries were not contained in the external datasets or their Wikipedia editions did not contain sufficiently many cuisine articles.

4.3 Affinity and Bias

Cultural affinity is described as the attention that one country pays to another country's cuisine. If one country, for instance Portugal, is highly interested (i.e. pays a lot of attention) in the Spanish cuisine, it is argued that Portugal has a positive affinity towards Spain. Implementing this idea using Wikipedia, the attention is expressed by the length (i.e. number of words or number of outlinks) of the article describing the Spanish cuisine on the Portuguese Wikipedia or the number of times it is viewed. Different formulas, as described in Chapter 3.2.1, account for the global popularity of the Spanish cuisine to correct for a certain expectation of popular cuisines to receive a lot of attention. If, for instance, all language editions show a high interest in the Spanish cuisine, then this would express the importance of the Spanish cuisine concerning its culinary aspects rather than the cultural affinity of other countries towards Spain. Formula $bias_m(l,c)$ therefore subtracts the expected global importance of the cuisine from the attention that it receives by a certain language edition. Formula $bias_s(l,c)$ does not include such a normalization and is therefore expected to also show the general popularity of cuisines. The third formula, $bias_g(l,c)$, is not used, since the results are very similar to the first one, as it only contains a slightly different normalization term.

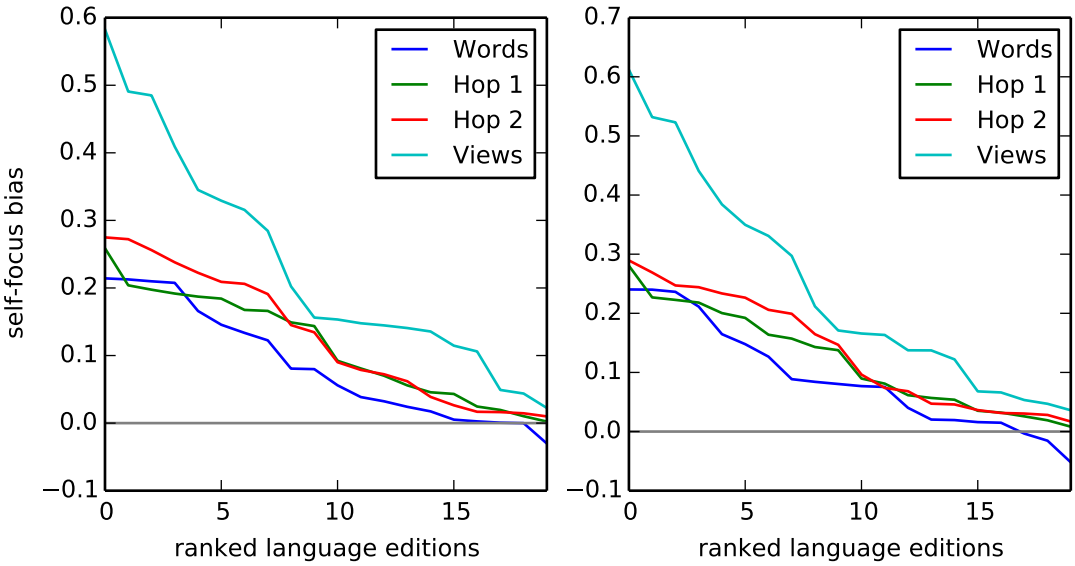
Is there a tendency for a self-focus bias?

Previous work has shown that Wikipedia editions tend to describe objects that are culturally relevant to them in more detail [HG09; OR11; MK06]. Places like cities or monuments that are located in Finland are for instance described much more accurately and in greater detail on the Finnish Wikipedia than any other language edition [HG09]. It is assumed that the same phenomenon holds true for cuisines. In the following, this assumption is tested.

Table 4.8 shows the results of the analysis for equations $bias_m(l,c)$ and $bias_s(l,c)$ and the different datasets, indicating a strong self-focus bias. However, especially for the view dataset, the variance is also rather large. One can see that the differences between the self-focus biases are quite strong, as the English, Russian or Turkish Wikipedia show only a little self-focus bias whereas some cultures, such as the Bulgarian, Catalan or Hungarian, seem to be especially interested in their own cuisines. Figure 4.10 shows the distribution of the biases, where the self-focus bias is clearly visible. It also becomes apparent that the view dataset shows the highest self-focus bias, which means that their own cuisine is more important to consumers of Wikipedia than to its producers.

	$bias_s(l,c)$				$bias_m(l,c)$			
Language	**Words**	**Hop 1**	**Hop 2**	**Views**	**Words**	**Hop 1**	**Hop 2**	**Views**
Bulgarian	+0.122	+0.204	+0.206	+0.583	+0.148	+0.227	+0.234	+0.612
Bosnian								
Catalan	+0.208	+0.167	+0.256	+0.491	+0.236	+0.200	+0.289	+0.523
Czech	-0.030	+0.043	+0.026	+0.148	-0.015	+0.062	+0.046	+0.171
Danish								
German	+0.024	+0.056	+0.072	+0.049	+0.020	+0.054	+0.068	+0.047
English	+0.001	+0.010	+0.017	+0.022	+0.016	+0.026	+0.030	+0.036
Spanish	+0.214	+0.197	+0.209	+0.153	+0.211	+0.192	+0.206	+0.137
Estonian	+0.210	+0.192	+0.238	+0.345	+0.240	+0.223	+0.269	+0.384
Finnish	-0.000	+0.019	+0.090	+0.145	+0.015	+0.032	+0.096	+0.163
French	+0.134	+0.184	+0.275	+0.141	+0.084	+0.138	+0.226	+0.066
Croatian								
Hungarian	+0.213	+0.259	+0.222	+0.409	+0.240	+0.280	+0.247	+0.441
Italian	+0.003	+0.092	+0.079	+0.115	-0.052	+0.035	+0.031	+0.053
Lithuanian	+0.080	+0.025	+0.039	+0.316	+0.080	+0.019	+0.036	+0.331
Latvian								
Dutch	+0.032	+0.046	+0.134	+0.044	+0.040	+0.057	+0.147	+0.068
Norwegian								
Polish	+0.146	+0.166	+0.272	+0.156	+0.127	+0.143	+0.244	+0.166
Portuguese	+0.081	+0.081	+0.062	+0.106	+0.089	+0.090	+0.074	+0.122
Romanian	+0.039	+0.187	+0.010	+0.485	+0.075	+0.218	+0.047	+0.532
Russian	+0.017	+0.002	+0.015	+0.136	+0.019	+0.008	+0.017	+0.137
Slovak								
Serbian								
Swedish	+0.166	+0.149	+0.191	+0.202	+0.165	+0.157	+0.199	+0.212
Turkish	+0.005	+0.070	+0.016	+0.284	-0.003	+0.081	+0.028	+0.297
Ukrainian	+0.056	+0.144	+0.145	+0.329	+0.077	+0.164	+0.165	+0.350
Average	+0.086	+0.115	+0.129	+0.233	+0.091	+0.120	+0.135	+0.242
Std.Dev	0.081	0.076	0.094	0.159	0.088	0.082	0.094	0.175

Table 4.8: **Which self-focus biases are inherent in the descriptions of the different cuisines?** Self-focus bias using two different formulas and four different datasets each. The values are the differences between the affinity towards the own and other cuisines. A value higher than 0.0 indicates a self-focus bias, a value below 0.0 indicates that the Wikipedia edition focuses on other cuisines (the maximum is 1.0 and the minimum -1.0, although extreme values are rare). It becomes apparent that all language editions of Wikipedia do show a self-focus bias, but the variance is rather big and becomes especially noticeable when accounting for the general popularity of cuisines, as done by formula $bias_m(l,c)$. The rows of some languages do not contain any values as these Wikipedia editions do not cover sufficiently many cuisines to allow for a reasonable calculation of the bias.



(a) (b)

Figure 4.10: **How are the self-focus biases distributed?** Distribution of self-focus biases showing the same data as in Table 4.8. The left graph shows the bias for the simple formula $bias_s(l,c)$, the right one for $bias_m(l,c)$. The horizontal gray line indicates no bias (i.e. foreign cuisines are considered as important as the own cuisine). As visible, only few language editions do not show a bias towards their own cuisine (almost all language editions show a self-focus-bias above 0.0). Both formulas lead to similar results.

Is there a tendency for a bias on regions that are geographically close to the country where the respective language is predominantly spoken?

Additionally to the question whether a direct self-focus bias is visible, the data was also analyzed for a focus of each language edition not only towards their own, but also towards the cuisines of their neighboring countries. Again, the three different formulas described in Chapter 3.2.1 were used.

Table 4.9 shows the results of this analysis for equations $bias_m(l, c)$ and $bias_s(l, c)$. Again, the third measure, $bias_g(l, c)$ was omitted. One can see that there are quite strong differences between the different language versions. The Portuguese, Finnish and French Wikipedia seem to focus significantly on their neighboring cuisines. Whereas this can be easily explained by the geographical position for Portugal, as its only neighbor in the dataset is the Spanish cuisine, it is more surprising for the other two languages. Contrary, the Dutch, Bulgarian and Ukrainian cuisine seem to be generally more interested in other European cuisines. This is particularly interesting, as they cover different number of cuisines on their respective Wikipedae. Whereas the Ukrainian Wikipedia has articles about almost all cuisines (29 out of 31), thus showing a high interest in other cuisines, which is also supported by the view statistics, the Bulgarian and Dutch Wikipedae only cover a small number of cuisines (10 and 14 out of 31). Their interest in foreign cuisines might therefore be an actual indicator of their cultural affinity towards other countries. Figure 4.11 shows the distribution of all affinities and Figure 4.12 shows two examples, the French and the Ukrainian Wikipedia, on a geographical map.

	$bias_s(l,c)$				$bias_m(l,c)$			
Language	**Words**	**Hop 1**	**Hop 2**	**Views**	**Words**	**Hop 1**	**Hop 2**	**Views**
Bulgarian	-0.048	-0.038	-0.025	-0.010	-0.036	-0.033	-0.008	-0.004
Bosnian								
Catalan								
Czech	-0.014	-0.015	-0.010	+0.015	-0.001	+0.000	+0.004	+0.038
Danish								
German	-0.000	+0.001	+0.000	+0.010	-0.018	-0.018	-0.018	-0.005
English	+0.030	+0.020	+0.012	+0.079	-0.025	-0.034	-0.047	-0.000
Spanish	+0.032	+0.030	+0.034	+0.076	+0.006	+0.005	+0.008	+0.041
Estonian	-0.030	+0.004	+0.055	-0.027	+0.007	+0.039	+0.086	+0.018
Finnish	+0.099	+0.020	+0.024	+0.020	+0.115	+0.038	+0.041	+0.033
French	+0.049	+0.031	+0.052	+0.060	+0.029	+0.008	+0.030	+0.033
Croatian								
Hungarian								
Italian	+0.013	+0.019	+0.031	+0.032	+0.004	+0.006	+0.018	+0.014
Lithuanian	+0.017	-0.004	+0.008	+0.032	+0.025	+0.005	+0.014	+0.039
Latvian								
Dutch	-0.017	-0.045	-0.028	+0.056	-0.069	-0.096	-0.084	-0.018
Norwegian								
Polish	-0.009	-0.022	-0.015	-0.004	+0.005	-0.009	-0.002	+0.013
Portuguese	+0.077	+0.179	+0.152	+0.142	+0.071	+0.175	+0.149	+0.128
Romanian								
Russian	-0.007	-0.001	-0.003	-0.009	+0.003	+0.007	+0.001	+0.006
Slovak								
Serbian								
Swedish	-0.009	-0.002	-0.006	+0.006	-0.005	+0.004	-0.003	+0.006
Turkish								
Ukrainian	-0.009	-0.013	-0.017	+0.000	-0.009	-0.017	-0.020	+0.004
Average	+0.011	+0.010	+0.016	+0.030	+0.006	+0.005	+0.011	+0.022
Std.Dev	0.038	0.049	0.043	0.042	0.041	0.053	0.051	0.033

Table 4.9: **Which regional biases are inherent in the descriptions of the different cuisines?** Regional bias using two different formulas and four different datasets each. A value of 0.0 indicates that a culture's interest in its neighboring cuisines is equal to its interest in other cuisines. A value higher than 0.0 indicates that the Wikipedia edition focuses more on neighboring cuisines, therefore showing a regional bias (the maximum is 1.0 and the minimum -1.0, although extreme values are rare). Depending on the used dataset, such a regional bias is not universally visible. When accounting for the general popularity of a cuisine as done by formula $bias_m(l,c)$, some language editions seem to focus more on cuisines of distant regions.

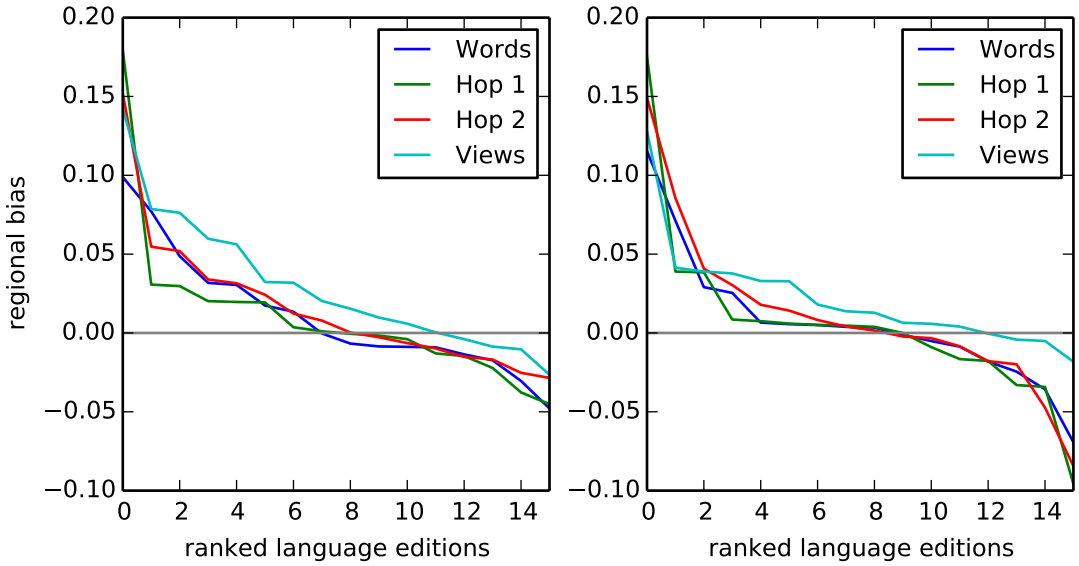
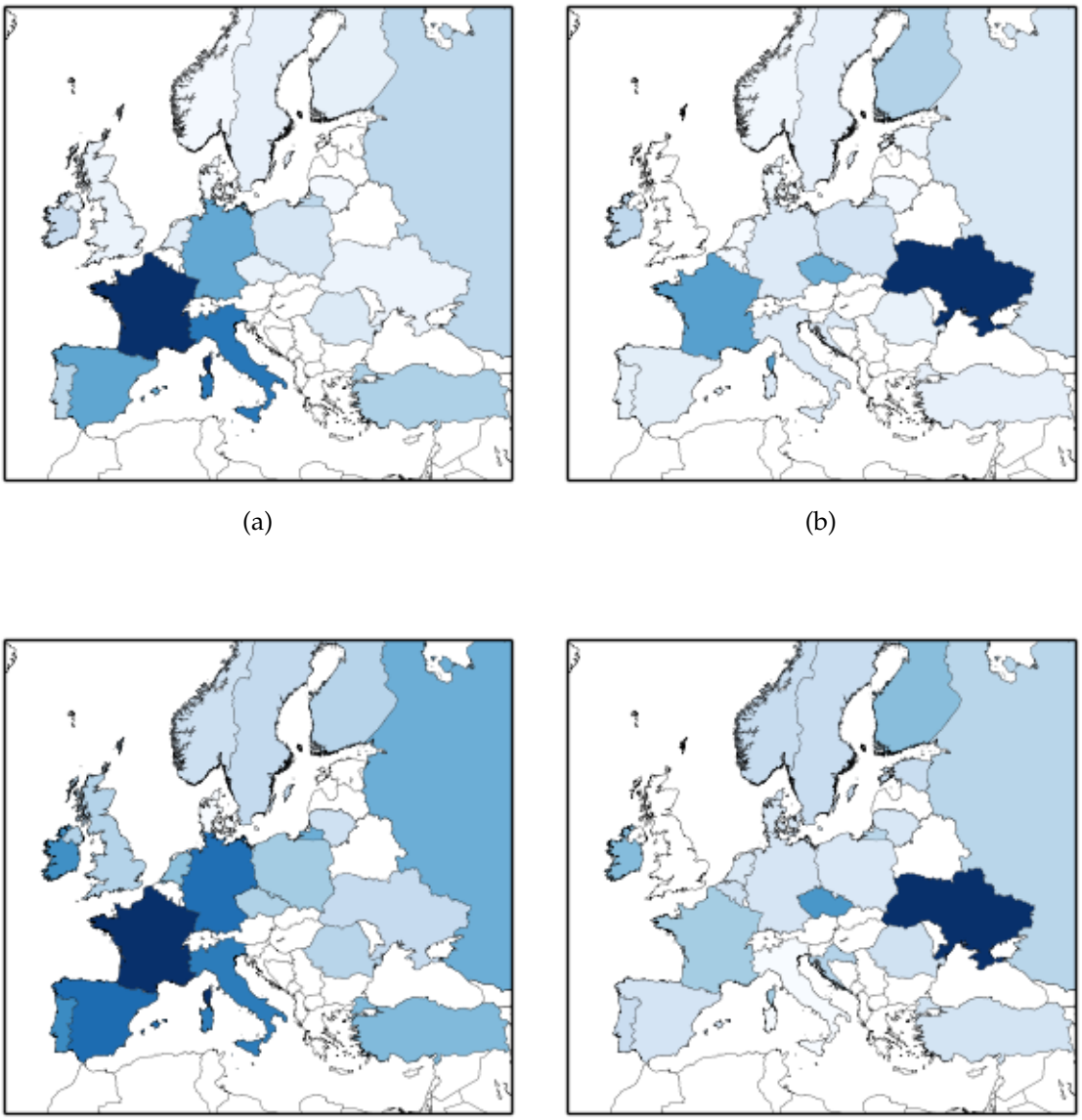


Figure 4.11: **How are the regional biases distributed?** Distribution of regional biases showing the same data as in table 4.9. The left graph shows the bias for the simple formula $bias_s(l, c)$, the right one for $bias_m(l, c)$. The horizontal gray line indicates a neutral bias (i.e. the same interest in both the neighboring and other cuisines). As visible, only a slight tendency towards cuisines of geographically near regions is visible.



(a) (b)

(c) (d)

Figure 4.12: **Does geographical proximity influence culinary biases?** Examples of biases plotted on a geographical map. The top row uses the $bias_s(l, c)$ formula, the bottom row the $bias_m(l, c)$, which accounts for the global popularity of a cuisine. The left column contains the biases of the French Wikipedia, showing a strong regional bias, the right column of the Ukrainian Wikipedia, which in fact expresses a negative regional bias. It has to be noted, that negative affinity was not displayed in red colors on this map, as the analyzed dataset barely contained negative affinities.

4.3.1 Evaluation

Can the affinity of one culture towards another be approximated by the bias of its Wikipedia edition for the other culture's cuisine?

The evaluation data for cultural biases are two external sources: The historical votings of the Eurovision Song Contest since 1975 expressing affinities of countries towards each other and check-ins on Foursquare to restaurants belonging to a certain cuisine. However, the two different sources do not correlate with each other (Spearman rank correlation ρ between -0.089 and 0.21 with p between 0.15 and 0.93 depending on the starting year of the contest data). Therefore it is questionable if they can be used as a ground truth for comparisons.

Before presenting the correlations, the strongest and weakest biases according to the $bias_s$ formula are listed in Table 4.10 in order to summarize the previous findings and give an insight into the actual data. It is noteworthy that especially the last column containing the view dataset shows the self-focus biases as the strongest affinities. This indicates that Wikipedia consumers tend to read about their own cuisines rather than about others. Figure 4.14 shows these biases for the same formula applied on the view dataset. The other formulas are not shown, but, with exception for the view dataset, all three formulas show a significant strong correlation ($\rho > 0.70$, spearman rank correlation). The bias extracted from the view counts correlates highly between the $bias_m$ and $bias_g$ formula ($\rho = 0.97$), however the simple formula $bias_s$ differs a little. As the correlations with Eurovision dataset reveal, the view dataset seems to be the one expressing cultural affinities best, as it is the only one which has a weak, but significant correlation (up to 0.25 with $p = 0.000013$ depending on the covered years and the used formula). The actual correlation values with the Eurovision Song Contest data are shown in Figure 4.15. Furthermore, the distributions of biases of the Eurovision and the Wikipedia data is shown in Figure 4.13. It is visible that less affinity is expressed on Wikipedia compared to the Eurovision Song Contest voting data. That means, the cross-cultural interest on Wikipedia is more balanced in the sense that only few country-pairs show a strong positive affinity. Some countries reveal a slight negative affinity, but no strong negative affinity was found. A comparison with the check-ins at Foursquare is listed in Table 4.11, showing no real correlation. From these findings one can conclude that the data on Wikipedia cuisine pages is only partially capable of explaining cultural biases between countries.

Words		Hop 1		Hop 2		Views	
no/Italian	(+0.8301)	no/Italian	(+0.6463)	no/Italian	(+0.6565)	da/Danish	(+0.7078)
da/Danish	(+0.5272)	ro/Polish	(+0.5665)	ro/Polish	(+0.5856)	bg/Bulgarian	(+0.6205)
tr/German	(+0.4579)	da/Lithuanian	(+0.4267)	da/Danish	(+0.3810)	ro/Romanian	(+0.5880)
ro/Polish	(+0.4000)	tr/German	(+0.4124)	da/Lithuanian	(+0.3569)	ca/Catalan	(+0.5473)
cs/French	(+0.3675)	da/Danish	(+0.4070)	et/Estonian	(+0.3332)	hu/Hungarian	(+0.4682)
...		...		...		...	
es/Finnish	(+0.0023)	pt/Serbian	(+0.0017)	hu/Italian	(+0.0009)	es/Serbian	(+0.0030)
pt/British	(+0.0021)	hu/Swedish	(+0.0015)	nl/British	(+0.0003)	bg/Catalan	(+0.0028)
nl/British	(+0.0020)	fr/Danish	(+0.0011)	bg/Catalan	(+0.0001)	fr/Danish	(+0.0025)
fr/Danish	(+0.0017)	ru/Danish	(+0.0005)	fr/Danish	(+0.0000)	pt/Latvian	(+0.0019)
pt/Belgian	(+0.0001)	pt/Belgian	(+0.0000)	pt/Belgian	(+0.0000)	en/Slovak	(+0.0001)

Table 4.10: **Which are the strongest and weakest affinities expressed on Wikipedia cuisine pages?** Highest and lowest affinity values for each dataset as measured by the $bias_s$ formula. As an example, the Norwegian Wikipedia shows the highest bias towards the Italian cuisine ("no/Italian") in all but the Views dataset. One can also see that, especially in this last dataset, the self-focus biases are reported as the entries with the highest affinity values.

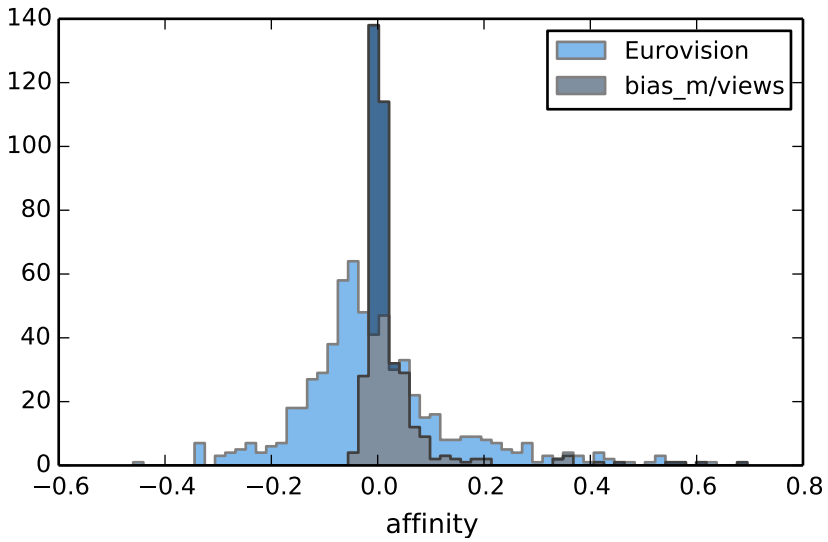


Figure 4.13: **How are the affinities on Wikipedia distributed?** Distribution of affinities according to both the Eurovision Song Contest votings and the bias as extracted from the view data using formula $bias_m$. The histogram shows that the bias on the Wikipedia cuisine pages is less pronounced and contains barely any negative affinity when compared to the findings from the Eurovision Song Contest.

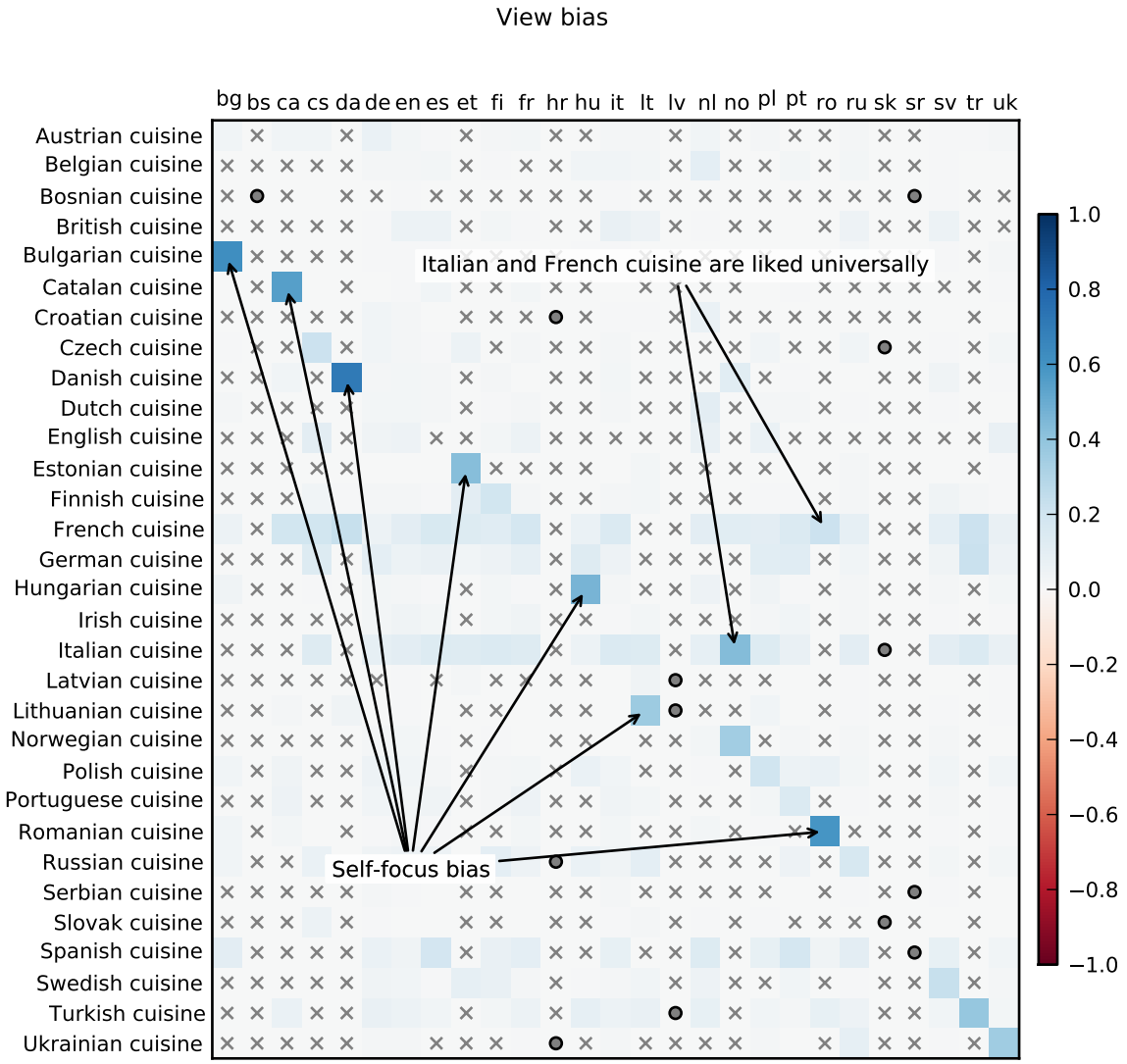


Figure 4.14: **Which biases are present in the description of cuisines on Wikipedia?** The heatmap shows the affinity for the view dataset as calculated by the $bias_s$ formula. The crosses indicate missing cuisine articles and the dots represent invalid values (i.e. for language editions that do not cover a sufficient number of cuisine articles to allow for a reasonable calculation). Some stronger biases like the interest of the Norwegian Wikipedia in the Italian cuisine are noteworthy.

	$bias_g$		$bias_m$		$bias_s$	
Words	-0.2190	(0.115222)	-0.2037	(0.143436)	-0.0882	(0.530169)
Hop 1	-0.1854	(0.183876)	-0.1604	(0.251231)	-0.0507	(0.718626)
Hop 2	-0.1032	(0.462155)	-0.1006	(0.473509)	-0.0216	(0.878125)
Views	-0.1032	(0.462155)	-0.1006	(0.473509)	-0.0017	(0.990607)

Table 4.11: **Do the affinities on Wikipedia correlate with external data collected from Foursquare?** Correlation values for all three formulas using all four datasets with the check-ins on Foursquare. There is only a slightly negative correlation visible, which is not significant. It has to be noted, however, that the Foursquare data only contained 53 country pairs with partially very low counts.

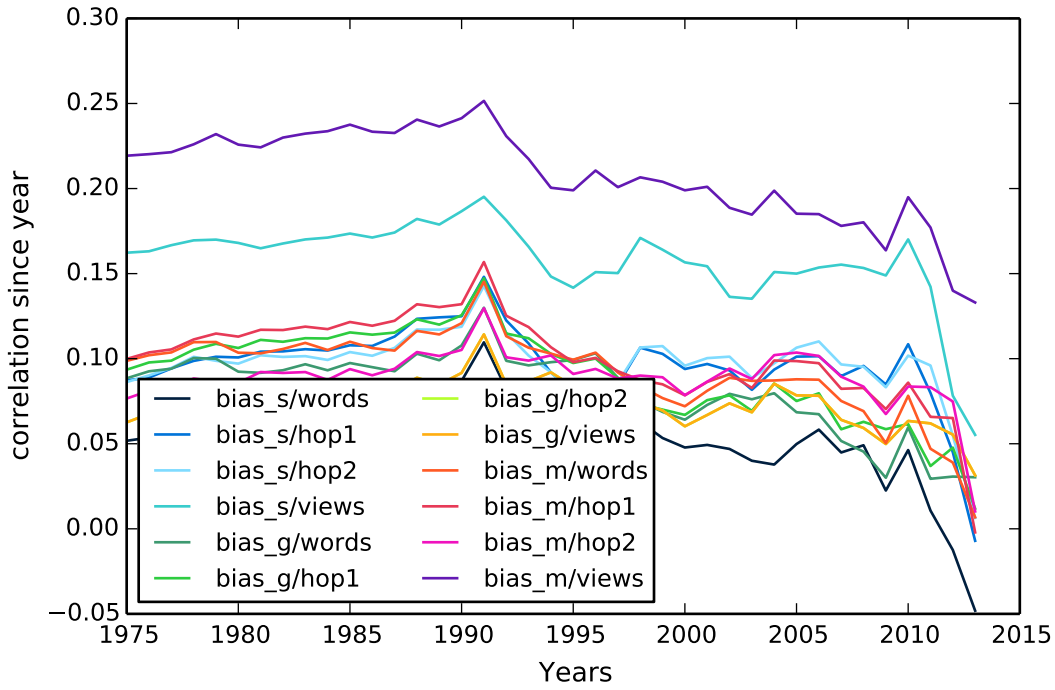


Figure 4.15: **Do the affinities on Wikipedia correlate with the biases expressed in the Eurovision song contest?** The correlation between the average Eurovision votings since different starting years and the affinity as calculated by the different formulas for all four datasets. The view data is the only dataset which shows a significant correlation with the Song Contest votings.

4.4 Summary

In Table 4.12, the three dimensions are summarized for each country. It is noteworthy that the three dimensions do not perfectly align. For instance, the cuisine best understood by Russia is the Danish one, which is also the one the Russians seem to dislike the most. The same is true for the Polish Wikipedia, which describes the Catalonian cuisine relatively well when compared to the locals, but also expresses a great deal of disinterest. Another interesting example is Turkey, which has the strongest affinity towards Germany and the weakest towards Austria, although both countries share a language and presumably some cultural background. In other cases, the three dimensions seem to be related. The Estonian Wikipedia shows strong values in all dimensions for northern countries whereas more southern countries are present in the low ranks. Noteworthy is also the fact that the Latvian, Bosnian, Croatian and Norwegian cuisines are considered to be the least similar cuisines by almost all others. This is an indicator for those cuisines to be either very unique or to have very specific descriptions on their home Wikipedae. Finally, the French cuisine seems to have a special position among European tastes. It is mentioned considerably often as both the most liked and most disliked cuisine.

	SIMILARITY		UNDERSTANDING		AFFINITY	
Viewed By:	**most**	**least**	**best**	**least**	**most**	**least**
Bosnia[0]	Serbia	Norway	-	-	-	-
Bulgaria	Serbia	Norway	Czech Republic	Russia	Spain	France
Catalonia	Spain	Latvia	Portugal	Austria	France	Romania
Croatia[2]	Serbia	Latvia	Russia	Ukraine	-	-
Czech Republic	Slovakia	Turkey	Austria	Romania	England	Romania
Denmark[3]	Sweden	Croatia	France	Lithuania	France	Finland
Estonia	Lithuania	Turkey	Sweden	Italy	Finland	Germany
Finland	Sweden	Croatia	Spain	Denmark	Russia	Denmark
France	Germany	Bosnia	Sweden	Romania	Spain	Denmark
Germany	Hungary	Bosnia	England	Serbia	Croatia	France
Hungary	Serbia	Latvia	Sweden	Belgium	Germany	France
Italy	France	Latvia	Ireland	Bulgaria	France	Catalonia
Latvia[2]	Estonia	Croatia	Lithuania	Turkey	-	-
Lithuania	Ukraine	Latvia	Sweden	Ireland	Russia	Croatia
Norway[3]	Sweden	Bulgaria	France	Italy	Italy	France
Poland	Lithuania	Bosnia	Catalonia	Romania	Germany	Catalonia
Portugal	Spain	Bosnia	Latvia	Denmark	Spain	Turkey
Romania[3]	Hungary	Norway	France	Poland	France	Spain
Russia	Ukraine	Bosnia	Denmark	Norway	Ukraine	Denmark
Slovakia[2]	Czech Republic	Turkey	Czech Republic	Italy	-	-
Serbia[2]	Bosnia	Norway	Bosnia	Spain	-	-
Spain	Catalonia	Latvia	Ireland	Croatia	France	Denmark
Sweden	Finland	Bosnia	Germany	Norway	Spain	Romania
Turkey	Serbia	Latvia	Belgium	Italy	Germany	Austria
Ukraine	Russia	Croatia	Spain	Finland	England	France

Table 4.12: **Which cuisine is most (least) similar, best (least) understood or is most liked (disliked) by European countries?** For each country, the most and least similar cuisine is listed according to the global ranking in the first two columns. All countries with English as primary language were left out and Germany was considered the representative country for the German Wikipedia (Austria was left out). The middle columns show the best and least understood cuisine by each country in the list. As the computation of the understanding required the language edition to contain an article about the target language, the number of articles about foreign cuisines are added in superscript to the countries if less than 4. The Bosnian Wikipedia does not describe any but the Bosnian cuisine, which is why there are no values in the first line. Finally, the last two columns show the cuisines towards which each country has the strongest and weakest affinity (as measured by the view dataset). A minimum of 3 descriptions of foreign cuisines was required in order to allow for a reasonable comparison which leads to 5 of the cells being empty.

5 Discussion

In the previous chapters, three different datasets, word counts, used concepts (outlinks) of the first and second hop network originating from the topical pages and the view statistics were analyzed for their representativeness of cultural similarity, understanding and affinity. For the evaluation, correlations with different external data sources were calculated.

Cultural similarity

Cultural similarity was approximated by culinary similarity by comparing the Wikipedia articles of two different cuisines. Both a global perspective, which considered the descriptions of the two cuisines on all language editions, and a local perspective, where only the descriptions on the 'native' Wikipedae of the two cuisines were considered, were analyzed. The findings suggest that the global perspective is a better indicator of cultural similarity than the local perspective, as it showed a higher correlation with external data retrieved from the European Social Survey (ESS). This may be due to the fact that not every Wikipedia edition contains an article about each cuisine, which leads to sparse data. Naturally, if more articles describing two different cuisines are used for a comparison between the two, the results are more stable. However, the two different perspectives seem to generally align with each other.

Furthermore, a higher similarity with cuisines from neighboring countries was detected. Again, the global perspective showed this phenomenon more clearly and is hence considered the better choice. The similarity between cuisines is, however, rather small, with no country pair exceeding a value of 0.18. This is an indicator for the expressiveness of cuisines and their ability to differentiate from one another.

The comparison with ESS revealed only a small, yet significant correlation ($\rho = 0.25$, $p = 0.0002$) which can be explained by the fact that cultural similarity contains more than just the culinary aspect. If other culturally relevant dimensions (such as music, literature, etc.) would be used to extend the vectors representing each culture, one can expect to have a better approximation of cultural similarity.

Finally, the plausibility of the results was evaluated using an online crowd-sourcing platform. The human workers were presented with two cuisine pairs and had to

choose the most similar pair, confirming the findings in 99.56% of all cases. This evidence is a strong indicator that the approach is generally applicable.

Cultural understanding

Cultural understanding was defined by comparing the external perspective of one culture's cuisine to its internal view. In other words, if, for instance, the Hungarian Wikipedia described the Polish cuisine similarly to the Polish Wikipedia, then the Hungarians were said to understand the Polish culture well.

Firstly, global descriptors of each cuisine were created by counting the concepts that were related to a cuisine on all articles describing it. This approach lead to reasonable descriptors, but also a certain amount of noise (concepts not describing the cuisine itself but possibly its origin or related concepts). However, the results showed that simple counting can be used to accurately detect important concepts for each cuisine.

Then the local, internal perception of each cuisine (as defined on the cuisine's "native" Wikipedia) was compared to each global, external perspective (as defined by all other Wikipedae). Both the first hop (i.e. the concepts directly used in the cuisine articles) and second hop (i.e. extended by concepts used in linked articles) datasets were used for this analysis. The findings suggest that the first hop dataset seems to provide a better indicator of cultural understanding. However, the overlap ratios describing the cultural understanding were relatively small, which supports the idea of each culture having a unique view of their cultural heritage as opposed to how it is perceived by others. These findings were also confirmed when analyzing the cultural understanding that every culture seems to have of each other culture separately. More popular cuisines, such as the Italian, French or Turkish cuisines seem to be better understood than other less popular ones. It, however, remains unclear as to how much the popularity contributes to cultural understanding.

The validity of the results is also influenced by the overlap ratios which seem to be heavily influenced by the length of the articles (i.e. bigger Wikipedae show to have greater overlaps) and the generally sparse matrix (i.e. cuisines are not represented in all Wikipedae).

Finally, when evaluating the findings using two external datasets, the European Social Survey (ESS) and migration data from the Global Bilateral Migration Database, significant, but small correlations are found. This might be partially explained by the external sources themselves which do not correlate very highly, thus indicating that they might not be representative to express cultural understanding. Other datasets, such as language similarity indexes (e.g. one could assume that countries with similar languages can easily communicate and therefore understand each other well)

could be used as an additional dataset for comparison. The definition of cultural understanding itself is, however, rather problematic and probably covers a whole range of aspects, which can not sufficiently be represented by editorial or other activity on Wikipedia.

Affinity and Bias

The affinity of one country towards another was estimated using the attention an article about the target country's cuisine received by the source country. Affinity was hereby defined using several measures, such as the view counts (i.e. the people visiting the Wikipedia page), the number of words describing the cuisine or the number of outlinks it contains. Different formulas were then applied to correct for the general popularity of a cuisine, as one could expect that some cuisines receive a lot of attention, simply because of their culinary aspects.

The first analysis dealt with the issue of a self-focus bias and proved its existence in the available data. The Spanish and Finnish Wikipedia in particular seem to focus on their own cuisines, whereas the bias was still visible, yet not so pronounced on other Wikipedae, for instance on the Turkish or Bosnian one. The findings also showed that the self-focus bias was strongest in the view dataset (i.e. the page view counts), regardless of the applied formula. It is possible to draw the conclusion that the consumers of Wikipedia focus on their own cuisine, whereas the editors strive to maintain a more balanced representation of the different cuisines.

Different to the similarity evaluation, only a small regional bias was visible in the analysis of the countries' affinities towards each other. The variance was also greater, and some Wikipedae even showed a negative regional bias (such as the Dutch, Bulgarian or Ukrainian language editions). This effect can be partially explained by the available data. As only European countries were considered and the number of neighboring countries differs for each country, a certain bias was already introduced by the selection of the dataset. It is therefore not surprising that for example the Portuguese Wikipedia shows a strong regional bias, as it only contains one neighbor in the dataset, namely the Spanish cuisine. However, other strong affinity values towards regionally close cuisines such as those on the Finnish or French Wikipedia, cannot be explained by this factor.

Two different external sources were then used to evaluate the calculated affinities. Comparisons with the Eurovision Song Contest votings showed a small, but significant correlation for the view dataset. As the correlation is rather small, further comparisons have to be evaluated before reasonable interpretations can be made. It is also noteworthy that the affinity values from the Eurovision Song Contest are

much more pronounced, whereas the data from Wikipedia showed barely any negative affinities and seemed to be rather neutral. This is reasonable as Wikipedia is not a competition although one can argue that different articles compete for a limited amount of attention. Nevertheless, views or edit behaviors help to identify positive affinities since they indicate interest, but no explicit indicators for disinterest exist. One can only infer them. In the context of the song contest, every country must vote for every other country. This way countries are forced to make also negative feelings visible. On Wikipedia, negative affinities may result in missing observations.

The second external source consisted of check-ins on Foursquare, where affinity was defined as the number of check-ins into a restaurant offering a specific cuisine (such as Spanish restaurants or Italian restaurants) in a geographical area (e.g. in Germany or France). Correlations with this source did not lead to any results. However, the two different external sources seemed to not be related either. Although they express different affinities (song votings and culinary preferences), this indicates that cultural affinities are a complex issue for which no single ground truth exists.

Finally, the findings suggest that the different metrics used (i.e. view counts, word counts and outlinks) are highly correlated, which means that the affinities are expressed similarly by the structural properties of the articles and its popularity as reflected in page views. It has to be noted though, that the affinities seem to be more pronounced when view dataset is used. The relative ordering, however, remains the same.

Outlook

Possible applications of the presented results are the inferring of cultural values semi-automatically from collaborative online-resources such as Wikipedia. Extending this idea, the cultural development and divergence or convergence of cultures could be directly, and in near real-time, measured and quantified. Using the approaches on single domains, such as food, they could help to build a set of interconnected culturally biased domain ontologies which incorporate cultural differences in a quantifiable manner. In any way, further research is necessary to investigate in the effects of adding further cultural dimensions, such as music, sports or literature to evaluate whether a combination of such dimensions is truly capable of approximating the richness of the different inter-cultural relations. Additionally, the analysis could be re-run including non-European cultures from Asia or South America, as bigger differences might be visible as opposed to only focusing on the European continent. Alternatively, the evolution over time using the history of each Wikipedia article might lead to valuable insights. Finally, the different outliers that potentially encode cultural phenomena such as the detected interest on the Turkish

Wikipedia towards the German cuisine, which could be an indicator for a strong migration flow, have to be interpreted manually.

6 Limitations

Language based comparisons of cultures are limited since language is only one aspect of culture and many different cultures and subcultures may share the same language. Although it is safe to assume that different languages can be directly mapped to the respective countries when only looking at Europe, as done in this analysis, the respective Wikipedia editions may be heavily influenced by other cultural groups outside of Europe. This is especially important when considering the English version of Wikipedia, which is most likely edited by contributors from all over the world. However, using countries is a reasonable approach, as [IW10] showed that despite globalization, nationality is a better predictor for values and beliefs of people than income, education, religion or sex. Also, Hofstede chose nationalities as cultural units as he, although admitting that this grouping is not optimal, considered them to be the only available unit of comparison and thus better than nothing [Hof02]. Nevertheless, it has to be noted that nations as proxies for cultures do have shortcomings, such as ethnic minorities, which may present a cultural group within a country which is different from the cultural group of the majority. To the best of my knowledge, this limitation is present in all current state-of-art studies since no sources for cultural data at different aggregation levels exist so far.

The cultural dimension used in this thesis, a culture's cuisine, was chosen for both the cultural identity it transmits and the reasonable mapping it allows to countries. Although more fine-grained regional cuisines exist, it is common to talk about national cuisines such as the Italian or Turkish cuisine. The problem of only using the culinary aspects of cultures has already been discussed, and further dimensions are needed to draw a comprehensive picture of a culture's preferences and identity. The problem of a varying general interest in the topic of cuisines, as explained in the introduction, may also be overcome by extending the cultural vector beyond the culinary dimension.

Also, the dataset, which is based on Wikipedia, contains an unquestionable bias deriving from especially the editors but also the viewers of Wikipedia, which do not represent the true population of each language group. This problem is though inherent to all online media research and can only be tackled by carefully analyzing the audience of each media and correcting for it (e.g. by combining different sources). Additionally, many missing cuisine articles limit the quantitative approach to a very

small number of samples. One can of course argue that the non-existence of articles contains information about their interestingness, however an alternative explanation is the popularity or size of the selected Wikipedia. If the analysis is performed on other culturally relevant objects, such as those belonging to art or literature, a broader coverage may lead to more valid results.

Finally, the evaluation, although yielding interesting insights into the representations of cultural values on Wikipedia, lacks a true common-ground truth to which it can be compared. Further studies, both qualitative and quantitative, are required to create a dataset on the basis of which the findings of this and oncoming research can reasonably be evaluated. Until then, only proxies such as the migration flows or differently biased online data like Foursquare check-ins can be used as a mere approximation. However, the focus of this thesis was not to create an additional incomplete dataset for comparisons, but to propose a method for automatically extracting cultural relations.

7 Conclusion

In this thesis, a novel approach was introduced to analyze cultural resources in online repositories and derive measures of cultural understanding, affinities and similarities. Overall, the findings reveal that no single definition of global entities, such as cuisines, exists. Instead, multiple differentiated views have to be considered when analyzing user-generated contents. This is especially important for extracting information from these repositories, as no single ontological representation of knowledge is possible. Cultural biases have to be taken into account. This argument is supported by the substantial difference found between the description of a cuisine which is produced by people who speak the language of the country from which the cuisine originates and the definition of the cuisine by others.

To summarize, the findings show that different language editions of Wikipedia reveal a substantial self-focus bias describing culturally relevant resources such as their cuisine. This self-description differs significantly from the description in other language editions. More prominent cuisines, such as the Italian and French cuisines, are slightly better understood by other countries and therefore show less differences. This indicates that a comparison of the descriptions of culturally relevant resources in different Wikipedae may reveal information about the global importance of these resources.

In the analysis of affinities between European countries, only few strong biases were found, most of which are towards geographically close cuisines. Most country-pairs show a neutral relationship when compared to external data. This raises the question whether Wikipedia is a suitable online media for trying to retrieve cultural affinities, as the editing guidelines themselves ask for content to be ”from a neutral point of view” and ”as far as possible, without bias”.

Concerning both cultural similarity and cultural understanding, the findings suggest that they can only be partially explained by views and structural statistics of cuisine pages on Wikipedia. Also, the local perspective gives less insight into cultural similarity than a global, aggregated view of the two analyzed cultural groups. However, a higher correlation with the migration data does indicate that the coverage of content on different Wikipedia editions might be related to migration flows.

Appendix

Bibliography

[Ailo8] Galit Ailon. "Mirror, mirror on the wall: Culture's consequences in a value test of its own design." In: *Academy of Management Review* 33.4 (Oct. 2008), 885–904.

[Ara+12] Pablo Aragon et al. "Biographical Social Networks on Wikipedia: A Cross-Cultural Study of Links That Made History." In: *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*. Linz, Austria: ACM, 2012, 19:1–19:4. ISBN: 978-1-4503-1605-7. DOI: 10.1145/2462932.2462958.

[Bao+12] Patti Bao et al. "Omnipedia: Bridging the Wikipedia Language Gap." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Austin, Texas, USA: ACM, 2012, pp. 1075–1084. ISBN: 978-1-4503-1015-4. DOI: 10.1145/2207676.2208553.

[BBo5] Francesco Bellomi and Roberto Bonato. "Network Analysis for Wikipedia." In: *Proceedings of Wikimania 2005 - The First International Wikimedia Conference*. Jan. 2005. DOI: 10.1.1.178.2233.

[BFB05] Susan L. Bryant, Andrea Forte, and Amy Bruckman. "Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia." In: *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work*. GROUP '05. Sanibel Island, Florida, USA: ACM, 2005, pp. 1–10. ISBN: 1-59593-223-2. DOI: 10.1145/1099203.1099205.

[Bräo5] Andreas Brändle. "Too Many Cooks Don't Spoil the Broth." In: *Proceedings of Wikimania 2005 - The First International Wikimedia Conference*. Frankfurt, Germany, 2005.

[Bro94] Richard Harvey Brown. *Rhetoric, textuality, and the postmodern turn in sociological theory*. Cambridge University Press, 1994, pp. 229–241.

[Car91] Kathleen Carley. "A Theory of Group Stability." In: *American Sociological Review* 56.3 (1991), pp. 331–354.

[CH11] Ewa Callahan and Susan C. Herring. "Cultural bias in Wikipedia content on famous persons." In: *JASIST* 62.10 (2011), pp. 1899–1915.

[Dix+12] Natalie Dixon et al. "FoodMood: Measuring Global Food Sentiment One Tweet at a Time." In: *International AAAI Conference on Weblogs and Social Media (ICWSM)*. The AAAI Press, 2012.

[ES13] Young-Ho Eom and Dima L. Shepelyansky. "Highlighting entanglement of cultures via ranking of multilingual Wikipedia articles." In: *PloS one* 8.10 (2013), e74554. DOI: 10.1371/journal.pone.0074554.

[Filo9a] Elena Filatova. "Directions for Exploiting Asymmetries in Multilingual Wikipedia." In: *Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies*. Boulder, Colorado: Association for Computational Linguistics, 2009, pp. 30–37. ISBN: 978-1-932432-33-6.

[Filo9b] Elene Filatova. "Multilingual Wikipedia, Summarization, and Information Trusworthyness." In: *SIGIR workshop on information access in a multilingual world* (2009).

[Fis88] Claude Fischler. "Food, Self and Identity." In: *Social Science Information* 27 (1988), pp. 275–293.

[GHZ13] Ge Gao, Pamela Hinds, and Chen Zhao. "Closure vs. Structural Holes: How Social Network Information and Culture Affect Choice of Collaborators." In: *Proceedings of the ACM 2013 Conference on Computer Supported Cooperative Work (CSCW)*. San Antonio, Texas, USA: ACM, 2013, pp. 5–18. ISBN: 978-1-4503-1331-5. DOI: 10.1145/2441776.2441781.

[GQJ13] Garcia Gavilanes, Daniele Quercia, and Alejandro Jaimes. "Cultural Dimensions in Twitter: Time, Individualism and Power." In: *International AAAI Conference on Weblogs and Social Media (ICWSM)*. The AAAI Press, 2013. ISBN: 978-1-57735-610-3.

[GT13] David García and Dorian Tanase. "Measuring Cultural Dynamics Through the Eurovision Song Contest." In: *CoRR* abs/1301.2995 (2013). DOI: 10.1142/S0219525913500379.

[HG09] Brent Hecht and Darren Gergle. "Measuring self-focus bias in community-maintained knowledge repositories." In: *Proceedings of the Fourth International Conference on Communities and Technologies*. ACM, 2009, pp. 11–20. ISBN: 978-1-60558-713-4. DOI: 10.1145/1556460.1556463.

[HG10] Brent Hecht and Darren Gergle. "The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2010, pp. 291–300. ISBN: 978-1-60558-929-9. DOI: 10.1145/1753326.1753370.

[HHM10] Geert Hofstede, Gert Jan Hofstede, and Michael Minkov. *Cultures and Organizations: Software of the Mind, Third Edition*. McGraw-Hill Education, 2010. ISBN: 9780071770156.

[Hof02] Geert Hofstede. "Dimensions do not exist: A reply to Brendan McSweeney." In: *Human Relations* 55.11 (2002), pp. 1355–1361.

[Hof80] Geert Hofstede. *Culture's consequences: International differences in work-related values*. Beverly Hills, CA: Sage Publications, 1980, p. 475.

[IB00] Ronald Inglehart and Wayne E Baker. "Modernization, cultural change, and the persistence of traditional values." In: *American sociological review* (2000), pp. 19–51.

[IW10] Ronald Inglehart and Christian Welzel. "Changing Mass Priorities: The Link between Modernization and Democracy." In: *Perspectives on Politics* 8.02 (June 2010), pp. 551–567. ISSN: 1541-0986. DOI: 10.1017/S1537592710001258.

[JL12] David Jurgens and Tsai-Ching Lu. "Temporal Motifs Reveal the Dynamics of Editor Interactions in Wikipedia." In: *International AAAI Conference on Weblogs and Social Media (ICWSM)*. The AAAI Press, 2012.

[Lie+14] Haiko Lietz et al. "When Politicians Talk: Assessing Online Conversational Practices of Political Parties on Twitter." In: *CoRR* abs/1405.6824 (2014).

[LP14] Han-Teng Liao and Thomas Petzold. "Geographic and linguistic normalization: towards a better understanding of the geolinguistic dynamics of knowledge." In: *Proceedings of The International Symposium on Open Collaboration*. 2014. DOI: 10.1145/2641580.2641623.

[MK06] Hermann Maurer and Josef Kolbitsch. "The transformation of the Web: How emerging communities shape the information we consume." In: *Journal of Universal Computing Science* 12.2 (2006), pp. 187–213.

[MS13] Paolo Massa and Federico Scrinzi. "Manypedia: Comparing language points of view of Wikipedia communities." In: *First Monday* 18.1 (2013). DOI: 10.5210/fm.v18i1.3939.

[NF12] Duyen T. Nguyen and Susan R. Fussell. "How Did You Feel During Our Conversation?: Retrospective Analysis of Intercultural and Same-culture Instant Messaging Conversations." In: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW)*. Seattle, Washington, USA: ACM, 2012, pp. 117–126. ISBN: 978-1-4503-1086-4. DOI: 10.1145/2145204.2145225.

[NG11] Keiichi Nemoto and Peter A Gloor. "Analyzing cultural differences in collaborative innovation networks by analyzing editing behavior in different-language Wikipedias." In: *Procedia-Social and Behavioral Sciences* 26 (2011), pp. 180–190.

[Ng12] Pauline Crystal Ng. "What Kobe Bryant and Britney Spears Have in Common: Mining Wikipedia for Characteristics of Notable Individuals." In: *International AAAI Conference on Weblogs and Social Media (ICWSM)*. 2012. DOI: 10.1145/2317956.2318077.

[OGR08] Felipe Ortega, Jesus M. Gonzalez-Barahona, and Gregorio Robles. "On the Inequality of Contributions to Wikipedia." In: *Proceedings of the 41st Annual Hawaii International Conference on System Sciences*. HICSS '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 304–304. ISBN: 0-7695-3075-8. DOI: 10.1109/HICSS.2008.333.

[OR11] Simon E. Overell and Stefan M. Rüger. "View of the world according to Wikipedia: Are we all little Steinbergs?" In: *J. Comput. Science* 2.3 (2011), pp. 193–197.

[PMT14] Katherine Panciera, Mikhil Masli, and Loren Terveen. "Cream of the Crop: Elite Contributors in an Online Community." In: *Proceedings of The International Symposium on Open Collaboration*. OpenSym '14. Berlin, Germany: ACM, 2014, 21:1–21:10. ISBN: 978-1-4503-3016-9. DOI: 10.1145/2641580.2641609.

[PZA06] Ulrike Pfeil, Panayiotis Zaphiris, and Chee Siang Ang. "Cultural Differences in Collaborative Authoring of Wikipedia." In: *J. Computer-Mediated Communication* 12.1 (2006), pp. 88–113. DOI: 10.1111/j.1083-6101.2006.00316.x.

[Ras08] Morten Rask. "The reach and richness of Wikipedia: Is Wikinomics only for rich countries." In: *First Monday* 13.6 (2008).

[Rei+13] Katharina Reinecke et al. "Doodle around the world: online scheduling behavior reflects cultural differences in time perception and group decision-making." In: *Proceedings of the ACM 2013 Conference on Computer Supported Cooperative Work (CSCW)*. San Antonio, Texas, USA: ACM, 2013, pp. 45–54. ISBN: 978-1-4503-1331-5. DOI: 10.1145/2441776.2441784.

[Roo10] Jochen Roose. "Der Index kultureller Ähnlichkeit. Konstruktion und Diskussion." In: *Berliner Studien zur Soziologie Europas* 21 (2010), pp. 3–61.

[Sil+14] Thiago Christiano Silva et al. "You are What you Eat (and Drink): Identifying Cultural Boundaries by Analyzing Food and Drink Habits in Foursquare." In: *CoRR* abs/1404.1009 (2014).

[Ste14] Thomas Steiner. "Bots vs. Wikipedians, Anons vs. Logged-Ins (Redux): A Global Study of Edit Activity on Wikipedia and Wikidata." In: *Proceedings of The International Symposium on Open Collaboration*. 2014. DOI: 10.1145/2641580.2641613.

[SV06] Laura Spierdijk and Michel Vellekoop. *Geography, culture, and religion: Explaining the bias in Eurovision song contest voting*. Feb. 2006.

[War+12] Morten Warncke-Wang et al. "In Search of the ur-Wikipedia: Universality, Similarity, and Translation in the Wikipedia Inter-language Link Network." In: *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*. Linz, Austria: ACM, 2012, 20:1–20:10. ISBN: 978-1-4503-1605-7. DOI: 10.1145/2462932.2462959.

[WFS09] Hao-Chuan Wang, Susan F. Fussell, and Leslie D. Setlock. "Cultural Difference and Adaptation of Communication Styles in Computer-mediated Group Brainstorming." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Boston, MA, USA: ACM, 2009, pp. 669–678. ISBN: 978-1-60558-246-7. DOI: 10.1145/1518701.1518806.

[Xin09] Yu Xintian. *Combining Research on Cultural Theory and International Relations*. 2009. URL: http://www.crvp.org/book/Series03/III-21/chapter-1.htm.

[Yai95] Gad Yair. "'Unite Unite Europe' The political and cultural structures of Europe as reflected in the Eurovision Song Contest." In: *Social Networks* 17.2 (1995), pp. 147–161. ISSN: 0378-8733. DOI: 10.1016/0378-8733(95)00253-K.

[YSK11] Taha Yasseri, Róbert Sumi, and János Kertész. "Circadian patterns of Wikipedia editorial activity: A demographic analysis." In: *CoRR* abs/1109.1746 (2011).

[YYQ09] Keiji Yanai, Keita Yaegashi, and Bingyu Qiu. "Detecting cultural differences using consumer-generated geotagged photos." In: *Proceedings of the 2nd International Workshop on Location and the Web*. ACM. 2009, p. 12. DOI: 10.1145/1507136.1507148.