Identity-based Motivation in Digital Engagement: The Influence of Community and Cultural Identity on Participation in Wikipedia

Marc Miquel i Ribé

TESI DOCTORAL

UNIVERSITAT POMPEU FABRA / ANY 2016

DEPARTAMENT DE COMUNICACIÓ SOCIAL

DIRECTORS

Dr. David Laniado

Eurecat, Centre Tecnològic de Catalunya

Dra. Mari-Carmen Marcos Mora

Universitat Pompeu Fabra

Dedicat a la meva mare, al meu pare, als meus avis i a la Laura.

To my mother, my father, my grandparents and Laura.

Acknowledgements

Even though working on a PhD is a lonely endeavour, I think it would have not been possible without the advice of research experts, reflections out loud with friends, or without the various helping hands that came from many people in very distinct occasions. I would like to publicly acknowledge some of them, although the list of people who deserve my appreciation is much longer.

I would like to acknowledge my thesis director Mari-Carmen Marcos (UPF) for her committed attitude to provide answer to all of my questions. Likewise, without the advice and encouragement from David Laniado (Eurecat), I would have probably set this project aside. The discussions we carried on and the collaborations we embarked in were extremely useful to me and I cannot be more grateful for his support and dedication. Again, thank you for being my co-director.

Other people from the Universitat Pompeu Fabra have provided me wise advice, among them I would like to mention Lluís Codina. Likewise, I want to acknowledge the time and discussions with the data visualisation jedis, Fernando Cucchietti and Luz Calvo (Barcelona Supercomputing).

I want to acknowledge my long-time friend David Morera, who represents my first connection with the Wikipedia editing community. I remember the nice welcome they gave me at the Catalan Wikipedia Community (Amical Wikimedia), their interest in my results and the opportunities granted to disseminate such results. Conversations with Àlex Hinojo, Arnau Duran, Joan R. Gomà, Vicenç Riullop and Laia Benito have been and still are very valuable for this work. Likewise, I was fortunate to have been dedicated time by the experienced members of the Wikimedia Foundation, Pau Giner, Aaron Halfaker and all the people from the labs.

Looking back in time, I want to thank those professors who engaged me in enjoying knowledge and research. Joan Campàs, a referent in academic life, emotional management and critical thinking. Jordi Enrich, for the artistic touch. Both inspired me to mix my technological and humanistic background in this thesis. I am grateful for the various learnings imparted by my professors. Thank you Marcos, Horacio, Eduard, Sebastià, Josep-Maria Gabriel for helping me take the first steps in the rigorous and well-sorted reflections. I want to thank Manuel Baldiz for teaching me the depth of analysis.

This thesis has completely relied on volunteering efforts (just like a Wikipedia), as I have not been funded by any institution, academic or private. Therefore, it is important for me to thank those who made my life easier during these years. I would like to mention Rodrigo Andrade and Joan Enrich, for their practical life lessons and support.

I would also like to thank Ricard Muntané, my friend and partner in the Catalan Games adventure. Together we proved that cultural identity is engaging. I knew it before I could prove it in Wikipedia.

Finally, I want to mention all the institutions that directly or indirectly gave me any support: Universitat Pompeu Fabra, Barcelona Media-Eurecat and Amical Wikimedia-Wikimedia Foundation.

Last but not least, I want to acknowledge my family at the same time I dedicate them this work. I owe a lot to each of them. To my girlfriend Laura, who already knew what writing a PhD thesis means, and was understanding and loving at all times. Now we are finally able to adopt the cat.

Resum

Internet i la tecnologia mòbil s'han consolidat com una esfera pública de la vida, on l'èxit dels llocs web i aplicacions sovint s'equipara a la seva participació. En aquesta tesi s'estudia la influència d'una motivació basada en la identitat en la participació, amb un enfocament especial en l'enciclopèdia col·laborativa Viquipèdia, en la qual les identitats són fonamentals per entendre la dinàmica de la comunitat i la diversitat de temàtica dels continguts. Per mitjà de l'anàlisi de dades de 15 versions lingüístiques, es descobreix que els editors desenvolupen una identitat de comunitat de Viquipèdia i a la vegada creen constantment contingut que representa les seves identitats culturals. Aquest contingut ocupa al voltant d'una quarta part (en nombre d'articles) de cada Viquipèdia, i encara més tenint en compte les d'edicions. Quan els editors augmenten la seva participació o esdevenen administradors, segueixen preferint l'edició de continguts impregnats de significats basats en la identitat cultural, la qual cosa indica una posició central d'aquesta identitat en el procés d'edició. Finalment, d'acord amb aquestes troballes, es destaquen diverses estratègies per fomentar la participació dels editors així com també fomentar l'enriquiment intercultural entre les versions lingüístiques de Viquipèdia.

Abstract

The Internet and mobile technology have consolidated as a public sphere of life, where the success of web sites and applications is often equated to engagement. In this thesis, I study the influence of identity-based motivation on digital engagement, with a special focus on the collaborative encyclopaedia Wikipedia, in which identities are fundamental to understand community dynamics and content diversity. By analysing data from 15 language editions, I find that editors develop a community identity in Wikipedia and at the same time they consistently create content representing their cultural identities. Such content occupies around a quarter of each Wikipedia in number of articles, and even more in terms of edits. When editors increase their participation or become administrators, they still prefer editing content imbued with identity-based meanings, which suggests their centrality in the editing process. Finally, in line with these findings, I highlight different strategies to foster editor participation and increase cross-cultural enrichment across Wikipedia language editions.

Keywords: Engagement, Data Analysis, User Experience, Multiculturalism, Identity, HCI Theory, Identity-based Motivation, Cultural Identity, Wikipedia.

Table of Contents

CONCLUSIONS, FUTURE RESEARCH AND DISSEMINATION

List of Figures

List of Tables

Chapter 1. Introduction

1.1 Motivation

Computers and consumer electronics are no longer seen as tools to simplify complex calculations, neither are they seen as workstations where to store data, but as the background technology sustaining social spaces where people work, develop daily plans, communicate appointments or chat. In the past few years, we have witnessed a displacement of many activities into the web: social networks sites, massive multiplayer online games, massive open online courses are only a few examples of this trend of digitalising social activities which have been encouraged by the expansion of handheld devices with Internet connection, such as Smartphones, to skyrocket into the popularity.

This new way of living in the digital has brought a new idolization to the term *engagement*, which is used to differentiate a good from an exceptional technological design. Research has started to be interested in users' attention and some scholars and business professionals introduced terms like "attention economy" and consider it a currency or an objective in itself. The term engagement is not new in the field of technology, for it has usually been related to user experience and attention. Webster and Ho (1997) used it to describe an immersive user state, which was very close to flow. However, for the digital and social spaces mentioned earlier, engagement is rather referred to as participation.

This other type of engagement finds its origin in Social Sciences, where for instance "civic engagement" has the objective of involving citizens into participating in public issues; or, in a more general view, "social engagement" refers to the degree of participation into a community. In the Internet, this participatory type of engagement is measured by the number of users, along with the intensity of their interactions. Two websites with a similar function are evaluated by these indicators to measure their degree of success, and most importantly, their capacity for thriving in a competitive environment.

Even though what makes an object engaging has been studied by conceptual frameworks (O'Brien & Toms, 2008), these tend to focus only in the user experience or more precisely on attention, with little consideration on understanding participation. Aspects like object's usability, aesthetics, or novelty play an important role in maintaining the user engaged. But, what aspects can explain best a participatory type of engagement? Motivation is a crucial aspect to understand engagement - especially participatory. For instance, T. de Vreede, Nguyen, & de Vreede (2013) presented a theoretical approach to understanding engagement in a collaborative problem-solving environment like crowdsourcing, and indicated that personal interest – such as motivation – is the trigger to further participation.

In this thesis, I propose studying user's identity as the anteroom of motivation in engagement. Identities can be both social and individual, including aspects of the self which are rooted in group memberships, and individual aspects which distinguish one from other people. In the Internet, online communities appeal to a group identity, and

users act motivated by a shared interest, while instead a social network site is based on fostering users' self-disclosure in order to help them build their identity and create relationships. Identities can become salient in certain circumstances, acting as a trigger to action. In this sense, in a participatory type of engagement, it may illuminate both the intensity and the meanings of interactions.

This thesis takes Wikipedia as the focus of its analysis and empirical research, and it aims to study the identities of Wikipedia editors. Wikipedia is generally known for being the most used encyclopaedia in the Internet, but its most remarkable aspect is the fact that editors voluntarily contribute to this public good. More precisely, the quality and quantity of content of the encyclopaedia depends on participation. In fact, the community grew significantly in 2006-2007, but during the past years, it has been characterised by a steady decline (Suh, Convertino, Chi, & Pirolli, 2009). The project's sustained success when it comes to the number of readers contrasts with the impasse of the editors' community. Nevertheless, considering its educative purpose and function in society, Wikipedia could considerably increase the mass of its editors in the future.

In studying editors' identities, it is possible to understand many aspects related to editors' participation, as well as the specific topics they choose to write about. In Wikipedia, disclosing personal content is not encouraged as in social networks sites, since Wikipedia is a place devoted to create an encyclopaedia. Precisely for this reason, studying identities and their influence in Wikipedia makes it an interesting case to inspect whether inadvertently they still foster editors' participation. The insights drawn from this thesis may lead to a better understanding and estimation of participation in future social objects.

1.2 Research Objectives

The goal of this thesis is **to understand the influence of identity-based motivation in digital engagement.** While this is a challenging endeavour, I propose to break it into three specific objectives to be fulfilled in the three parts of the research.

- **Objective 1:** Define digital engagement and create a conceptual model to encompass a participatory type of engagement, based on the current literature.

- **Objective 2:** Understand the aspects which influence Wikipedia editor engagement, by reviewing current research studies.

- **Objective 3:** Investigate the influence of identity-based motivation on Wikipedia editor engagement, by taking into account identities such as cultural identity and Wikipedia community identity.

1.3 Scope

This thesis has a broad scope: Part 1 aims to define and model the concept of digital engagement; Part 2 focuses on its application to the object Wikipedia; finally, Part 3 measures the influence of Wikipedia editors' identities on their engagement as a new aspect of engagement.

By dedicating Part 1 to engagement, I aim to pin down the participatory type among other engagement manifestations, and, at the same time, to set aside the often vague and generic use of the concept. By dedicating Part 2 to Wikipedia, I frame its current problem of engagement, making use of the extensive literature on its different aspects, at the same time I provide a document to raise awareness of the state of the situation and its multiple causes. By dedicating Part 3 to study identity in Wikipedia, I propose using this concept of identity for the first time in its abundant literature, to study behavioural and content aspects, and gain a greater understanding of editors in such a collaborative, voluntary and endless project.

The thesis goes from a theoretical discussion to an empirical study where two cases studies were carried out. The third part is based on the identity-based motivation framework (Oyserman, 2009; Oyserman & Destin, 2010), a Social Psychology theory which sheds light on the dynamic nature of identities and their circumstantial activation in a context. I hypothesise that editors' identities become salient when they choose the content they want to edit, favouring those topics they feel more congruent with. At the same time, editors who develop a community identity based on the project values will feel more aligned with Wikipedia in future actions, and consequently are likely to increase their participation.

Community identity implies sharing values and developing characteristics such as a global view of the project and the predisposition to collaboratively work on its needs, going beyond personal preferences. I have chosen the cultural identity as an example of contextual and shared identity among editors from a Wikipedia language edition. Likewise, cultural identities can also be equated to a set of meanings. By studying and characterising them, it is possible to understand and provide a deeper explanation of a phenomenon already detected in the literature, which is the contextualisation of Wikipedia language editions. This being said, for this empirical part, the election of a quantitative data analysis methodology is very appropriate since in Wikipedia every single aspect of community interaction is stored for all their available languages. This study takes into account the editor population, and only considers the collective of readers for very specific experiments and lateral results.

This study does not include qualitative methods. An approach involving qualitative methods would give insight on how editors perceive their emotions and on how they describe their motivations at the basis of their behaviour. In fact, the desire of studying cultural identity was driven by a survey I run on editors in 2012, while being involved in the Catalan Wikipedia (see Appendix 1). I found that one of the reasons given by the editors to make the encyclopaedia grow is to recreate the cultural heritage of the country, also known in Catalan, as '*fer país*'. While this was a clue for the appropriateness and interest of studying identities in Wikipedia, here I did not consider developing the 2012

survey or expanding it to multiple language editions, but I decided instead to adopt a data analysis methodology.

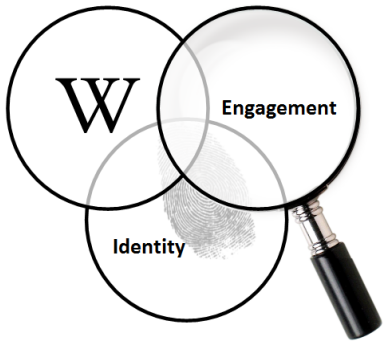
In this regard, engagement research is mostly based on empirical studies, and Wikipedia makes available most of the data from its entire population. A quantitative method allows validating the hypothesis for the entire population of Wikipedia, during its entire history, in several language editions of different scales and cultural backgrounds, hence increasing the robustness of the conclusions. In addition, I believe the use of this quantitative techniques and statistical methods can provide clear characterisation and sound conclusions on specific aspects of engagement and content. In this sense, this thesis has as a side-goal the dissemination of these results in order to be helpful to Wikipedia and its communities. Hence, it is appropiate to make an effort to translate the findings into some design recommendations, and at the same time, to propose solutions to the described problems.

1.4 Thesis Structure

The outline of this thesis is as follows. After this current Chapter 1 dedicated to introducing and motivating the thesis, in Chapter 2 (Part 1), I develop the digital engagement model. In Chapter 3 (Part 2), I present Wikipedia as a socio-technological and multicultural object, and review the current literature according to the aspects related to the digital engagement model.

Part 3 starts in Chapter 4, where I explain identity in the context of Wikipedia and present an identity-based motivation framework in order to understand the influence of editors' identities on engagement. In Chapter 5, I provide the main methods and data used in the empirical research. In Chapter 6, I present the Community Identity as a case study. In Chapter 7 and Chapter 8, I present the Cultural Identity as a case study.

Finally, a conclusions and dissemination part is developed. In Chapter 9, general conclusions and further work are drawn. In Chapter 10, I discuss the relationship of this research and society, its ethical considerations and possible implications and opportunities driven from it (from design recommendations to community activities). I included Appendixes at the end of the thesis to present complementary results.



"It's not computer literacy that we should be working on, but sort of human-literacy. Computers have to become human-literate" Nicholas Negroponte

PART 1: DIGITAL ENGAGEMENT

Chapter 2. Defining and Modelling Digital Engagement

2.1 Introduction

Understanding engagement has become the ultimate challenge for any designer or technology researcher, a sort of deep knowledge they all aim to. An engaging object is not just preferred over a similar one, but it will be more intense in any possible given use. Engagement means more. The term is employed in very different contexts, from games (E. A. Boyle, Connolly, Hainey, & Boyle, 2012; Cheung, Zimmermann, & Nagappan, 2014) to social networks sites (Freyne, Jacovi, Guy, & Geyer, 2009) and educational multimedia presentations (Jacques, 1995), among many others. In the web sphere, research has been particularly prolific while the industry has put analytics methodologies at the service of marketing objectives (Peterson & Carrabis, 2008). Engagement occurs in the highest complexity of virtual worlds, but also in the simplicity of a text-based communication. It has become popular and synonymous of desirable.

Because of this, during the past years, empirical research has reached maturity and a great range of methods have been detailed. The user has been analysed in its cognitive, emotional and behavioural dimensions, by means of both objective and subjective measures (Lalmas, O'Brien, & Yom-Tov, 2014). Likewise, the study of scenarios like multitasking (Lehmann, Lalmas, Dupret, & Baeza-Yates, 2013) or the use of multiple portable devices (Giang, Hoekstra-Atwood, & Donmez, 2014) have provided valuable insights on how people relate with technology. However, despite its soundness, empirical research has appeared dispersed and unable find a common ground for the studies. Paradoxically, although engagement gained momentum in empirical research, it remained vague at a conceptual level.

In fact, the broad use of the concept is at risk of overlapping with previous terms from the Human-Computer Interaction field. For instance, positive psychology Flow theory explains a mental state of a long and sustained use of an object with a focused attention, which is sometimes equated with engagement, but this is not necessarily the only way of engaging with objects. A more narrative-explorative use of the term engagement is explained by Activity Theory (Marsh & Nardi, 2014). However, in the past years, engagement has been used in Social Media websites such as online news to imply participation (Ksiazek, Peer, & Lessard, 2014; Liikkanen & Salovaara, 2015).

This participatory type of engagement has a long tradition in the field of Social Sciences, where civic engagement refers to the objective of involving citizens into participating in public affairs, or employee engagement focuses on worker performance. Even though with the emergence of Social Media this participatory type of engagement has become very popular, there is no model which explains how it occurs. The current framework for engagement with everyday websites is useful in order to explore the user experience (O'Brien & Toms, 2008), but does not include any concept dedicated to the intensity of interaction, namely the user's participation.

In this Part 1, I pursue **Thesis Objective 1** of defining digital engagement and creating a conceptual model to encompass participation, based on the current, which I believe it can also be helpful to researchers, designers and users in reaching a common understanding. To study engagement, I propose and explain several essential aspects of the object (composition, design, logics and content) and of the user (emotion, motivation, understanding and attention). The main contribution of this chapter is a theoretical discussion leading to a preliminary model, which ultimately bridges theoretical concepts with current empirical research. In general, I see this preliminary model as one more step towards a better understanding of how people engage with technology. This chapter is organized as follows:

In Section 2.2, I review the main definitions and background of the different uses of the concept 'engagement' in Social Sciences and Human-Computer Interaction. Then, in Section 2.3, I propose a definition of digital engagement as a meta-construct with a focus on aspects of both user and object. I integrate such aspects into the model. Consequently, in Section 2.4, I review and synthesize the broad variety of methods to measure engagement and classify them into user-centered and object-centered. I review how these methods were used in studies published during recent years. Finally, in Section 2.5 I conclude with a discussion. As an addendum, in Section 2.6, I review the different objects where identity may play a role in engagement.

2.2 Previous Definitions and Applications

In this section, I review the definitions and applications of engagement to better understand how the concept has been used and what challenges it may involve.

By definition, to *engage in* is "to attract and hold fast", while *to be engaged* stands for, among other possible meanings, "occupying the attention of someone", either with an activity or with a commitment[1]. From these multiple meanings, two separate streams of research on engagement arise, with a growing cross-fertilization between them: one in the interdisciplinary field of Human-Computer Interaction (HCI) which usually approaches engagement focusing on the psychological aspects of a person performing an activity with technology; the other in the broad field of Social Sciences, which remarks the person's commitment and social actions such as participation.

2.2.1 Human-Computer Interaction Tradition

The concept engagement was first employed at the beginning of the 1990s to characterize the user's psychological state while interacting with all kinds of technological interfaces. Therefore, it covered meanings similar to being attentive and absorbed while enjoying technology. First, Laurel (1991) studied software interfaces and referred to engagement as the feeling of being in direct manipulation with a physical object. Laurel (1991) considered that when a system is working properly, the user entails "sustained belief" that it will respond as if it is alive, even bringing "playfulness". Further on, in the

[1] http://www.dictionary.com/browse/engage

context of educative technology, Jacques et al. (1995) referred to engagement as the effect of a system which ultimately attracts the user's attention by arousing his emotions. For Webster and Ahuja (2006), engagement with a website was similar to a flow state of mind in which the user enjoys a very focused attention, and its satisfaction could trigger a future intention to return to the website. Engagement was considered mainly an emotional or attentional component, but with a sense of amusement.

As technology evolved and applications were designed for many more objectives, the term engagement incorporated "user". For instance, in video games, user engagement was considered a prior phase to immersion and presence (Brown & Cairns, 2004), two states in which the player abandons himself in a virtual world and identifies himself with the character. Still in videogames, it was correlated with entertainment and it was very influenced by usability (E. A. Boyle et al., 2012; C. M. Karat, Karat, Vergo, & Pinhanez, 2002). Later, van Vugt et al. (2007) analysed engagement with virtual reality by measuring it as a concept between involvement and distance. In other distant fields like the creation and use of information systems, user engagement also comprised a sense of involvement (Hwang & Thorn, 1998; Kappelman & McLean, 1992). All in all, the term comprised different psychological attributes depending on the context and application, and it overlapped with other concepts in the same field of Human-Computer Interaction.

One of the overlapping concepts is the term user experience (UX), which appeared several years after engagement, to cover the emotional and exciting side of technology. The advantage of UX over previous concepts is that it allowed introducing a discourse which was not centred on efficiency (like usability or the more classic HCI). By taking into account the user psychological state, and also by emphasizing positive emotional outcomes such as joy, fun and pride (Hassenzahl & Tractinsky, 2006), UX found its place and dominated the field in the development of services, products and digital objects – both in the academia and especially in the web industry. However, the original sense of playfulness initially explained by engagement was then better covered and generally assumed to belong to UX. It became the popular term and the general catch-all term to refer to user needs, feelings, thoughts, expectations in order to improve the design process (Hassenzahl & Tractinsky, 2006).

Engagement needed to be redefined in order to avoid repeating the same debates in a parallel research line with UX. A possible solution was given by O'Brien and Toms (2008) whose strategy was to define user engagement as "a quality of the User Experience" (p. 949). By embedding engagement into the newer, more popular and studied concept of UX, O'Brien and Toms (2008) would limit the concept to a range of positive experiences. The above-mentioned authors developed a framework for the web research where user engagement is characterized by "challenge, aesthetic and sensory appeal, feedback, novelty, interactivity, perceived control and time, awareness, motivation, interest, and affect". This extensive list of attributes was very common to the UX studies - e.g., aesthetic appeal (Lavie & Tractinsky, 2004) and emotion (Forlizzi & Battarbee, 2004) -, and it would explain engagement modelled as a process, with a "point of engagement", an "engagement period", a "disengagement moment" and maybe a "re-engagement".

Even though the use of attributes could explain how a time-based process develops, I consider it presents several problems. My critique to this perspective is three-fold:

- First of all, the framework has varied along the years and the authors included usability as secondary when applying the framework to news portal (O'Brien, 2011), while it disappeared in later versions to include trust (Lalmas et al., 2014). It is necessary to clarify which are the essential attributes to explain engagement and which attributes are instead secondary to better understand specific scenarios.

- Second, the UX perspective of engagement solely considers the user, thus the attributes are often rewritten from this point of view even though they do not emanate from it. Usability becomes 'perceived usability' and aesthetics 'aesthetic and sensory appeal'. This relegates the object in a passive secondary plan. Hence, certain aspects of the object like 'content' or 'meaning' cannot be incorporated in the framework.

- Third and most importantly, possibly due to the embedding of engagement into User Experience, there is no attribute related to the external dimensions of engagement such as user behaviour (interaction, or participation). Nevertheless, empirical research based on this framework ends up measuring user behaviour by using metrics and data analysis techniques (Attfield, Kazai, & Lalmas, 2011; Lalmas et al., 2014). If the intensity of the user behaviour is considered engagement, then the relationship with its causing factors should be explored. In other words, for a more comprehensive model of engagement, a participatory type should be explained.

2.2.2 Social Sciences Tradition

When engagement is applied to Social Sciences, it emphasizes the participation and a sense of social relatedness. Examples are varied from all areas of public life. For instance, in civic or political engagement (Ball, 2005), engagement implies an orientation or predisposition towards action. To engage citizens means helping them become members of the political process through discussions and debates which influence them. Any kind of community engagement refers to the way an individual integrates into a group, whether if it is a public government, an education system or a research group (Ahmed & Palermo, 2010). Engagement is desirable in order to improve social dynamics, give value to the relationships and achieve their group goals.

In addition, by engagement is also meant an individual process where the individual progresses in a specific activity or environment. In the education field, it is connected to intensity of behaviour and emotional involvement during the task (Appleton, Christenson, Kim, & Reschly, 2006). In a work environment, employee engagement is seen in terms of the relationship with the organization, of the commitment with group values, and it is aimed at improving group performance (Reeves & Read, 2009). Likewise, sport engagement is more focused on the path of achieving autonomy and improving the quality of its practice (Alvarez, Balaguer, Castillo, & Duda, 2009).

All kind of groups and individuals are interested in having engaged people, whether these people assume their activity consciously or the purpose is not publicized and goes

unnoticed. This is especially interesting for all the fields related to business. In marketing, brand engagement refers to the relationship of a customer with the image of a product or company, encompassing aspects from the regularity of use, involvement, or even recommendation to others (Arcas, 2014; McWilliams, 2013). Similarly, customer engagement discusses how users co-create value around a company, purchases and interactions (Brodie, Hollebeek, Juric, & Ilic, 2011).

Until recently, any activity would include the term engagement and associate it to participatory values, while technology use would refer to engagement as a matter of attention and emotion. However, the advent of Social Media and all the new technological and portable devices has led to a wide sense of the term engagement which overpasses frameworks and past definitions like the one from O'Brien and Toms (2008), which is narrowed to attributes from the cognitive and emotional dimensions of the user.

Especially in the web sphere, examples of new forms of engagement based on this social sense are abundant. For Peterson and Carrabis (2008), visitor engagement in websites implied reaching some objectives throughout the measurement of user behaviour with metrics (e.g. number of pages visited, loyalty or recency). In social networks, engagement is mainly considered and measured in terms of social interaction - i.e. number of votes, comments or shares - (Smith & Gallicano, 2015), while in content repositories like Wikipedia, the editor engagement is linked to the editing activity in articles and in policies (Halfaker, Geiger, Morgan, & Riedl, 2013a), as their success is totally dependent on it.

2.2.3 Challenges for a Shared Model

In summary, research on engagement has been conducted both on individuals and on groups (in the latter case in organizations with collective goals) interacting with all kind of objects. Engagement happens 'to be everywhere' because it is implied in the sense of relating to something. When applied to the current technology, I remark the following shared conclusions about engagement from both Human-Computer Interaction and Social Sciences traditions:

- Engagement is an objective of the researcher or the designer, who all have an expectation set on the user to act in a particular way. Therefore, the measurement of metrics (Peterson & Carrabis, 2008) confirm an object is properly designed for its goals.

- Engagement is multidimensional in the emotional, cognitive and behavioural aspects of the user (Attfield et al., 2011; Lalmas et al., 2014), and also takes into account the design aspects of the object such as usability.

- Engagement is considered positive with no clear absolute value. It has a positive sense which emphasizes the positive aspects between both object and user interactions (Lehmann, Lalmas, Yom-Tov, & Dupret, 2012).

As seen, the obstacles for obtaining a universal definition of engagement reside in the slightly different uses of the concept in non-related fields, in the technological advances and their socialization, in addition to the interferences from non-academic uses of the word. The most important challenge is that current models do not explain the participatory type of engagement. In order to conciliate this weakness, I attempt to provide a clear definition to study engagement. This will be the object of the next section.

2.3 Digital Engagement

In this section I propose a new working definition for the concept 'digital engagement' and I discuss the aspects that influence it.

Consistently with the aforementioned conclusions from past section, I define digital engagement as ***the quality which guarantees that the connection between a user and a digital object remains active.*** In doing so, the concept of engagement becomes inclusive of the previous uses in both tradition Human-Computer Interaction and Social Sciences, and fits best with the available evidence from research.

Engagement exists as long as the connection is active. In order to do so, the digital object necessitates the user's response; the minimal expression of such response is attention. An engaged behaviour can be either the user passively absorbed or participating frenetically, but in both cases, it guarantees the connection remains active. This way, the user participation becomes one specific manifestation of the connection.

An engaging object or an engaging experience are desirable or alluring, because they keep the connection alive. Checking the e-mail, updating a profile in a social network or browsing the Internet in the search for a particular piece of information can imply connections at different levels of engagement intensity. Thus, the expressions of engagement are the outer manifestations of the connection, in other words, the interaction between the user and the object which keep the connection active. Engagement is a concept to understand such connection in its multiple configurations.

Each connection may manifest itself in a different way (longer or shorter duration, and more or less interaction). These manifestations are measurable and can be explained by studying each part of the connection. Engagement needs to be holistic and embrace complexity, as all user and object aspects are interrelated and may influence one another. For a full understanding of the engagement quality, one has to consider both user dimensions (emotional, cognitive and behavioural) and object characteristics. These dimensions will be discussed in the following sections.

2.3.1 Study of the Connection

User, object and agency. Drawing upon this definition, I view the connection between user and object as the unit of analysis, where editor and object are equally important. This

is in sharp contrast to the user-centred paradigm prevailing in Human-Computer Interaction, which considers the object as a passive tool. The first computer applications designed for massive use had a practical goal. For instance, the spreadsheets allowed companies or families perform accountability calculations in an easier way. Later on, computers enabled the creation of objects simulating a place where users could engage in activities; this was called *media* or *medium* (Laurel, 1991). Nowadays, websites and software create digital spaces in which the user learns, plays, competes or communicates with others. And most remarkably, the last fronteer in digital objects is their capacity to perform communicative actions, which makes them convert into *social actors* - for instance, a personal assistant which can figuratively encourage users to achieve goals or change habits in their daily life (Fogg, 2003). This is why in the advent of a more sophisticated artificial intelligence, in order to understand the connection between a user and a digital object, engagement should not be exclusively centred on the user's perceptions, needs or behaviours, but it should also consider the object.

Digital objects can be programmed to constantly change in their design and content to attract and maintain interaction with the user, and their behaviour can be totally unexpected as if they were beings (Suchman, 2007). Yet, they cannot respond to the notion of 'agent', or the "one who initiates the action" (Laurel, 1991, p. 4). They are designed with an active purpose, but when it comes to establishing a new connection, they are conditioned by a user's previous acceptance (e.g. a smartphone is able to receive app notifications but only if the user turns it on). Even though the object is not considered equal to the user, the former is designed to engage and hence to continue the connection by fulfilling its purpose. Objects have no consciousness, but their intentionality is delegated by designers to its content (and meaning), design (both aesthetics and interaction), and behaviour (sustained by the logics of their algorithms). From the designer's perspective, the user has to confirm the quality of his creation with its use. Engagement is a measure of success.

Object composition. Every digital object can be composed of smaller ones - sometimes to provide a functionality or a new piece of information. Therefore, each inner object has at least one action to engage a user into a pattern of interaction and influence behaviour. The composition property is the key for simple and complex websites, video games and all kinds of digital objects. Any connection with a compound object can lead to several sequential connections with different inner objects. This is known as multitasking, and it can happen either with multiple objects or with compound objects.

A social network website is a clear example of a compound object. This often includes inner objects such as a synchronous communication channel (i.e. "chats"), photographs and news all at once. Banhawi, Ali, & Judi (2012) analysed the use of the social network site Facebook and found that the novel content, appearing constantly as new inner objects, drives people to be eager to see more. The users' preferred activity was writing to other users' personal spaces, followed by watching photographs, status updates, social investigation and content surfing. Users engage with Facebook in unlimited combinations with inner objects which disappear or are substituted. Likewise, this can happen in a multiobject context, in which the user has to respond to notifications from a social network, the e-mail, an opened document, the phone and a control panel with a connected smart home (see Figure 1).

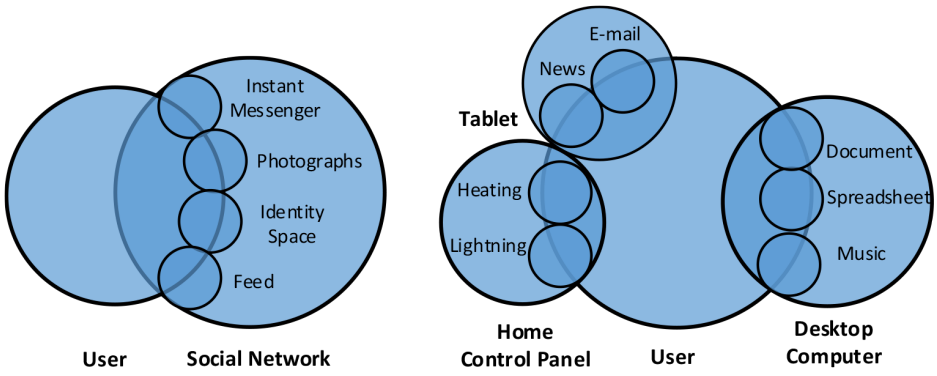


Figure 1. Two scenarios of engagement with multiple digital objects. One with several inner objects from a Social Network Site and another using different electronic devices

Aspects of engagement. When developing any engagement study, it is mandatory to define the two parts (user and object) and the precise context in which the connection takes place. The holistic view of engagement implies that all elements must be taken into account for their interrelations. For example, when a group of users connects to a single object, this will be called community engagement. In a digital object mediated communication the engagement of each individual will be determined by the interactions of his peers. Likewise, a single user can engage with multiple independent digital objects in order to reach a specific goal, to understand a story or simply for the sake of entertainment. Hence, the study of engagement can get beyond the limits of a single object and include multiple objects in the same scenario. As a result, focusing on only one object (e.g. a website) without considering the rest would lead to wrong conclusions (Lehmann et al., 2013).

Once the compositional parts are specified, it is necessary to understand which are the inner aspects which drive them to constitute in a temporary relationship and maintain it. On the wake of O'Brien and Toms (2008) I aim to appeal at several aspects of the user and the object to explain the reason why the user and object stay connected. As far as the user is concerned, I propose *emotion, motivation, cognition* and *attention*, as related concepts. When it comes to the object, I take into consideration *design, composition, content* and *logics* (Figure 2). The connection between both parts will generate a fluent dialogue, which is an aspect dependent on both parts. Each of these aspects will be developed and explained in the following sections, in order to set a proper setting, hypothesis, and the variables to perform an experiment.

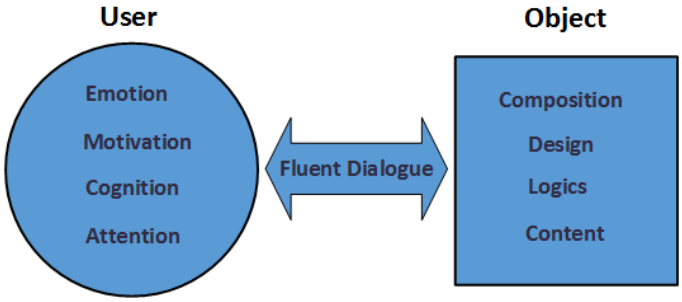


Figure 2. Main aspects of the user and of the object influencing digital engagement.

2.3.2 User's Emotion and Motivation

User's *agency* or *drive* to act has been widely explained by the user's emotional and cognitive dimensions. Concepts such as emotion and motivation are fundamental in understanding why a user gets involved in an activity with a digital object. In psychology, an emotion is seen as a set of internal processes of self-maintenance and self-regulation (Markus & Kitayama, 1991). The introduction of this concept in the study of technology use has contributed to understanding the centrality of emotion in the user's experience. Strong emotions and pleasure alter our perception of products (Forlizzi & Battarbee, 2004). User's positive emotions assure the connection is maintained and guarantee user's satisfaction at the end of the task; in the same way as user's good performance predicts a future intention to return (Chung & Tan, 2004; Webster & Ahuja, 2006).

Motivation is a complex construct linked to both emotion and cognition. Recent studies appealed to motivation to study the depth of engagement (Ainley, 2006; Bouvier, Lavoue, & Sehaba, 2015; Chapman & Selvarajah, 1999; T. de Vreede et al., 2013; O'Brien & Toms, 2008). Motivation is the key factor for users to initiate, persist in or resume an action. It is related to energy, direction, persistence, all aspects from activation to goal reaching (Deci & Ryan, 2012; Ryan & Deci, 2000). Every connection with a digital object has a motivation behind, whether it is random and unique access to a website, or routine and regular use of a phone App. The principle 'The more positive the experience, the more driving force will the object have' does not always apply. As a matter of fact, some experiences can be unpleasant or arouse negative emotions in the user and still motivate the user to engage with the object.

Since motivation is central to the user's behaviour, understanding it is at the very basis of understanding engagement. In other words, the study of motivation allows the researcher to explain how to make connections last longer or be more intense in terms of interaction, namely the specific design changes he would implement.

There is a great variety of models of motivation; it is not a unitary phenomenon. For instance Self-Determination Theory relates motivation to psychological needs, such as relatedness, competence and autonomy (Ryan & Deci, 2000). The same theory proposed the distinction and generalization of motives into intrinsic and extrinsic, according to the user's locus of control towards action. Intrinsic motivation is independent from any valuation and is induced by the inherent satisfaction derived from performing an activity. On the contrary, extrinsic motivation is triggered by activities which imply an outcome of any kind, either a reward or ego involvement.

Concerning immersive experiences, it has been argued that an intrinsic motivation can easily lead to focused states of attention. Theories like Cognitive Evaluation Theory (Ryan & Deci, 2000) or Flow Theory (Csikszentmihalyi, 1991; Nakamura & Csikszentmihalyi, 2009) have explored which factors could facilitate an intrinsic motivation, focusing on user's competence and autonomy. Flow is achieved in an activity with challenges of all kind (mental or physical) but which does not exceed the user's existing skills, in such a way that the user is in control of the situation, never bored but neither anxious, between control and arousal. The user is capable of dealing with every challenge according to his skills, while new challenges appear continuously. With this

mindset, a user performs a task driven by intrinsic motivation, and he does so with such a joy and intense concentration that he loses reflective self-consciousness and sense of time (Nakamura & Csikszentmihalyi, 2009). This is a so-called "optimal experience" because the user does the best performance and the sense of absorption in the activity is complete - a loop in experience.

Intrinsic motivation can be a source of very joyful experiences (Chapman & Selvarajah, 1999) and, its related flow state can be very beneficial to keep a user engaged and therefore a connection active. However, active connections can also exist in contexts that do not provide the suitable challenge-control structure necessary for Flow to happen (O'Brien & Toms, 2008; Webster & Ho, 1997), or can be driven by extrinsic motives such as social or physical rewards. The importance of motivation lies in that it sets the direction for user's action and the reason behind it, which depending on the degree of motivation may act with more or less intensity. In consequence, a strong motivation will lead to long and sustained object use or, in other words, to an intense interaction. However, aspects regarding the object design, content and purpose can be as determinant as motivation on how the interaction unfolds in a connection.

2.3.3 Object's Design, Content and Logics

Facilitating Flow and Zone. The most significant difference between physical and digital reality is that the latter can be designed up to its minimal details. Design implies both aesthetics and functioning of the object. Changes in object design can be tailored to respond to the different kinds of motivation and improve engagement by providing interaction. Ever since their appearance, video games have been considered the closest expression to a complete digital and active reality. Przybylski, Rigby and Ryan (2010) applied the Self-Determination Theory to videogames and found out that they induced a feeling of well-being into players due to their addressing and fulfilling basic psychological needs of the user, such as competence, autonomy, and relatedness.

Motivation can be reinforced by interaction and design, and therefore both aspects contribute to maintaining the connection. One example can be found in Cheung, Zimmermann and Nagappan (2014), who evaluated the impact of different video game design elements by means of self-reported comments. They advocated the idea that design was crucial for engagement (especially during the first hour) and that it influenced how players perceived the rest of the game. Namely, according to the abovementioned authors, "the first hour must provide the right balance of challenge and skill to put players on the right track to enter a flow state" (p. 59). In this first hour, the player learns the control keys, the mechanics and the consistency of the scenario, which allows him to progress and gain control at the same time, satisfying his motivation. Among the players' comments collected by the study, several users were asking for trainings to be provided at the beginning of the game in order to avoid frustration. Cheung et al. (2014) concluded that it was the rapid figuring out of how to control interaction (clarity in the interaction controls), a curve of challenges as well as allowing the user to set further goals which kept motivation and interaction stimulated.

In a slightly different environment, Schüll (2012) studied a scenario involving no challenge, namely the digital games of chance. Schüll (2012) noted that the use of videogambling machines in Casinos induce a psychological state called 'Zone'. This state was very comparable to flow in terms of absorbed attention, but instead of stimulating activity and high performance, the player remains in an idle and desubjectified position for hours. Paradoxically, in the zone the player seeks and feels a sense of control while actually being out of control (Schüll, 2012). Schüll attributes this state to the way the design conduces the player triggering extrinsic motivations (with reward structure). The odds of wining are low and even the frequency depends on optimized algorithms aimed to trap specific player preferences and styles. In addition, the interface with numerous buttons creates a false control sensation and every option is disposed to keep the zone going. When the player enters the game, he delegates the action to the machine.

In both experiences of Flow and the Zone, interaction design and motivation are crucial to reinforcing the connection, which remains active in a loop structured experience. The two cases present some structural differences and similarities (see Figure 3). The main difference consists in the necessary challenge and sense of progress, which in well designed videogames or in any other digital object may lead to flow while the user is trying to accomplish a goal. The need to be in control of the situation is a requirement for Flow, and it is based on a cause-effect sensation in each user action. The same does not occur in videogambling machines mainly because their game goals are set externally to user's will and are not controllable. Flow implies an emotional self-reward and intrinsic motivation, while the Zone is provoked by the videogambling rewards which entail monetary prizes (extrinsic motivation). Instead, the commonality they share is a sense of rapid feedback and a continuity. The feedback provided by the videogambling design is a key factor in maintaining the player's desire to continue. The fact of being shown the next gambling round triggers a passive acceptance in the player.

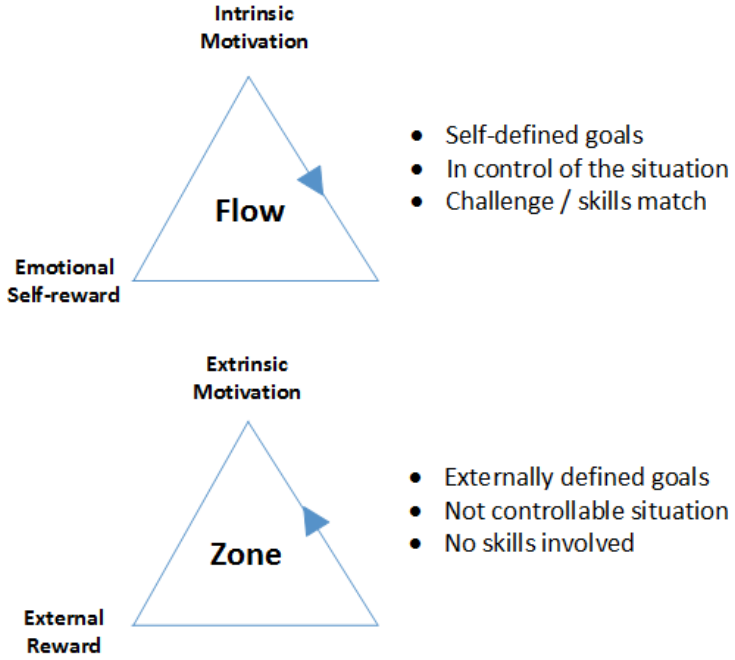


Figure 3. Flow and Zone motivation loops and their structural characteristics.

As already stated, in both cases design is the counterpart for motivation. When the user's skills and decisions play a determining role in the interaction, the continuity is totally *user-directed*, when instead this is not the case, the continuity is totally *object-directed*. As long as object's interaction design presents continuity and reinforcement for user's motivation, the motivation type can become secondary. In other words, the object, in order to feed the user's needs and motivation, can present feedback and affordances.

Flow and the zone are two clear and delimited types of immersive experience with an emotional and motivational loop structure, but other more mixed digital objects can produce similar effects. Mauri et al. (2011) investigated the psychophysiological effects of Facebook to find out why it is so successful. They noticed that the measures described a core state (between valence and arousal) very similar to Flow but in an environment of no challenge. However, in such a social network site, the positive affect is associated with a recreational activity which addresses the social needs of the user by presenting multiple inner objects related to the user (Mauri, Cipresso, Balgera, Villamira, & Riva, 2011).

All in all, user's motivation is central to engagement, but no less than the way the object design anticipates the interaction. In fact, a user can remain engaged with an object, switching between inner objects, as long as there is a motive for interaction (Marsh & Nardi, 2014). In this sense, some digital objects may be designed with strategies using a rich variety of characteristics and functionalities. When studying engagement, it may be interesting to ask: what contributes more to connection continuity, the user's motivation, or the interaction provided by the object's design? In some cases, it is the object that is more influent in keeping the connection going, while in other cases it is the user's motivation.

Object design strategies for continuity. Engaging with digital objects is very similar to engaging in the physical world with people, processes, places or groups. However, the digital reality can count on strategies to encourage continuity by anticipating steps totally tailored to the user's motivations. These strategies can use content and meaning (the "what"), but also different design components and available actions (the "how").

In fact, content is the object's property which triggers the user's interest, a kind of intrinsic motivation with positive emotional valence (Ainley, 2006). In consequence, users can possibly be engaged because a specific content is interesting to them. As an example, in an online news website the specific content was a key factor engaging users in reading (Arapakis, Lalmas, Cambazoglu, Marcos, & Jose, 2014). The higher the users' interest, the more comments they posted, in parallel, the more enjoyment they draw from watching the video, the higher the possibilities to take an active role and comment. In a similar manner, de Vreede et al. (2013) considered that engagement in a crowdsourcing community was determined by how it enabled developing personal topic interest - stimulating the user to go from a passive user to an active contributor.

Other content strategies to maintain the user emotionally aroused aim at providing novelty or structuring information in the form of a story. Laurel (1991) studied the different canons of drama theory and showed that meaning could be a driver of interest in keeping attention high. The different phases of a linear story raise or lower the emotional arousal the same way in a narrative video-game as in a theatre play. The interaction between the

elements of the story, if well written, cause in the user excitement and interest to see what comes next. In a digital object, these elements can be combined and varied according to inputs, while the experience can be personalized to keep the motivation high. This is the case of social networks sites: the content continuity is provided in a central channel of information (feed), while the social continuity is ensured by means of a synchronous communication channel (chat). In fact, putting the accent on computer-mediated communication in order to convert websites in social and foster engagement has been a common strategy.

In addition, technology can be used for persuasive purposes such as increasing engagement (Fogg, 2003). Sophisticated algorithms allow digital objects to perform actions humans would not be able to. For instance, digital objects can be persistent in presenting actions repeatedly and in an impersonal way (e.g. a software sending e-mail to customers informing them of an unfinished purchase and what they left in the basket). To this same purpose, digital objects can use rich design components based on video and sound. Not to mention their easiness of transport which grants ubiquity.

In general, the more technology evolved, the more engagement has become critically dependent on design aspects such as rapid feedback. I believe that with the development of artificial intelligence technologies and the abundance of data, designers' efforts will focus more on personalization, in order to achieve a greater symbiosis between the user and the digital objects. An example of this is the *filter bubble* algorithm used in web searches or social networks. This strategy exclusively provides results or information tailored according to previous results, avoiding cognitive dissonance and therefore reinforcing the user's point of view and expectations (Pariser, 2011).

2.3.4 Cognition, Usability and the Fluent Dialogue

Properly designed digital objects can entice interaction by providing new goals to keep the user motivated. In some cases, for instance in activities where the user's creativity is stimulated, motivation alone can be sufficient. However, any sort of feedback is useful to inform on the user's progress in attaining a specific goal. For Laurel (1991), feedback was a necessary part in order to sense a "direct manipulation" with the interface, because it reinforced the interaction with immediate response.

On the contrary, lack of feedback leads to frustration, because it leaves the user with no indication on how to continue the interaction, and can be as detrimental for the connection as the lack of interest or of motivation. The user can hypothesize the next step in an interaction either thanks to the design or to his own prior experience and skills. Even though most of the times the user can learn and become tech-savvy, the object is also expected to support the user by means of an understandable, self-explanatory design. It is only after a repeated use of an object that the user internalizes actions and achieves autonomy to perform the activity with little effort or conscious thought (Marsh & Nardi, 2014).

Fluent dialogue depends both on the user and on the object. It can be looked at in terms of a trade-off between the user's skills and cognition and the object's design with its

affordances for interaction. While the user must identify the next step, and understand how to reach it, the object must provide clear affordance and feedback in order to facilitate the user in doing so. One of the object's usability goals is avoiding user disorientation by offering clear affordances on any possible further action.

Usability is the object's design property responsible for providing feedback, visual cues and information in order to facilitate the performance of the user and make it satisfying and memorable (Nielsen, 1999). In early studies feedback quality and speed were considered related to usability. This is why usability is a central aspect of the object, and it has sometimes been taken for granted in relation to engagement in previous models (Lalmas et al., 2014; O'Brien & Toms, 2008).

Fluent dialogue may exist in the different design components and physical channels in which a digital object is represented. Depending on the type of object, there may be varied design components available for interaction (audio, visual, touch, space, etcetera.). For instance, a game can imply a 3D immersive experience with a whole range of audiovisual features; an instant messenger may only involve text and few pictographic images. In some cases, these components may even allow the user to modify the object (e.g. comments in a website or uploading a video) or communicate with other users.

If an object encompasses several inner objects, the design should consider the overall perception of these objects in order to avoid confusing the user. A fluent dialogue between a user and a compound object can take place with multiple channels and inner objects at the same time, in a similar way to multimodal communication (Klein, 2015; Norris, 2004). This could be the scenario of home automation, in which temperature, lighting and music are controlled coordinatively. For instance, song selection could be manipulated using a screen interface, while temperature change could be simultaneously activated by voice. Very importantly, in order to attain a fluent dialogue and keep connections active the diversity of components does not have to exceed the user's cognitive abilities.

2.3.5 The Connection is Reciprocity

Previously, I assumed that for a connection to be active there must be reciprocity between the object and the user. While the object is able to create and manage multiple connections with different users independently, the user can *only* respond to one connection to the exclusion of others - giving it his attention in a precise situation or in repeated moments along time. Attention is the cognitive process of selecting information by allocating limited resources of processing (Anderson, 2009). The complex process of paying attention has been depicted as a continuum with different levels of attention, going from unconsciousness (total lack of awareness) to focal attention (vivid awareness) (Norris, 2004). In this section I explain why the management of attention is a key aspect of engagement, closely linked to emotion, motivation and interaction.

Attention and multitasking. Connecting to multiple objects at once is known as multitasking. Switching tasks can the result of external interruptions (Mark, Iqbal, Czerwinski, & Johns, 2015) or of self interruptions such as internal decisions (Benbunan-Fich, Adler, & Mavlanova, 2011). Since multitasking depends on the management of

thoughts and notifications, the variety of possibilities of attending to multiple stimuli in a short period of time is high. Users do not cope with several connections simultaneously but experience them sequentially, in a process of fast engaging and disengaging. Typically, each of the old and new connections can be explained by motivation. However, when studying the reason why a user engages into a new object, one also needs to take into account the user's emotional and attentional state prior to it.

Mark et al. (2015) studied states of attention in a work environment. In order to understand how people multitask while they perform their job tasks, the authors tracked thirty-two employees by means of different metrics. They found that, in any time of the day, the choice of a particular object was related to the one object used just a moment before. For instance, rote or routine work was followed by more Facebook or face-to-face interaction, while focused and aroused states lead to more e-mail. Mark et al. (2015) concluded that users choose some objects and create connections as 'short breaks' in their on-going tasks, breaks aimed at emotional relief (also known as emotional homeostasis) and at keeping the balance. Furthermore, even though attention is linked to the activities' degree of challenge, the availability of the other objects is an influent, possibly distracting factor.

Prior to engaging with an object, the user's attention is already susceptible to be distracted. This means that dividing the phases into "point of engagement", "engagement", "disengagement" and "re-engagement" as depicted by O'Brien and Toms (2008) would be over-simplistic. Users are potentially already unconsciously connected to a new object before it actually happens. Therefore, each connection must be explained by the context where other objects come into play, by the previous object the user has connected with, and by the previous interactions with the same object (if any). They can all be indicative of the reason why the user engages in a connection with an object.

The beginning and the end of the process of engaging with an object tend to be blurry and fragmented. However, the interactions and the elements the user identifies as emotional rewards are able to lead to higher states of attention. For example, a user can feel positive emotions after achieving the proposed challenges in a video game, which in turn would stimulate him to continue and set more difficult challenges, until perhaps reaching a Flow state. This may depend on many variables such as the user's skills or object design (i.e. challenges), which makes the Flow outcome – loss of sense of time – a very unique guarantee of a long-lasting connection.

Most digital objects are used in a noisy environment with multiple objects sending notifications (e.g. e-mail or Social Media), and therefore to study their connections one has to consider multiple periods of time. Each connection is dependent on the previous connections, their interactions and sketched situations. Likewise, different connections held over time between the same user and object can be analysed as a longer connection or, in other words, as a relationship.

Multiple connections and transitioning states of attention. User attention states are a reflection of how connections are developed with an object and with its composition, or even in a broader context of multiple objects. Remaining in a connection or transitioning to others will depend on how the user discriminates the different stimuli provided by

single or composed digital object. I delimit four different states of attention - flow, focus, distracted and background - taking into account the continuum from unconsciousness (total lack of awareness) to focal attention (vivid awareness) (Norris, 2004). For a better understanding of the interrelation between the user's attention and object's composition, the four states of attention are depicted in Figure 4.

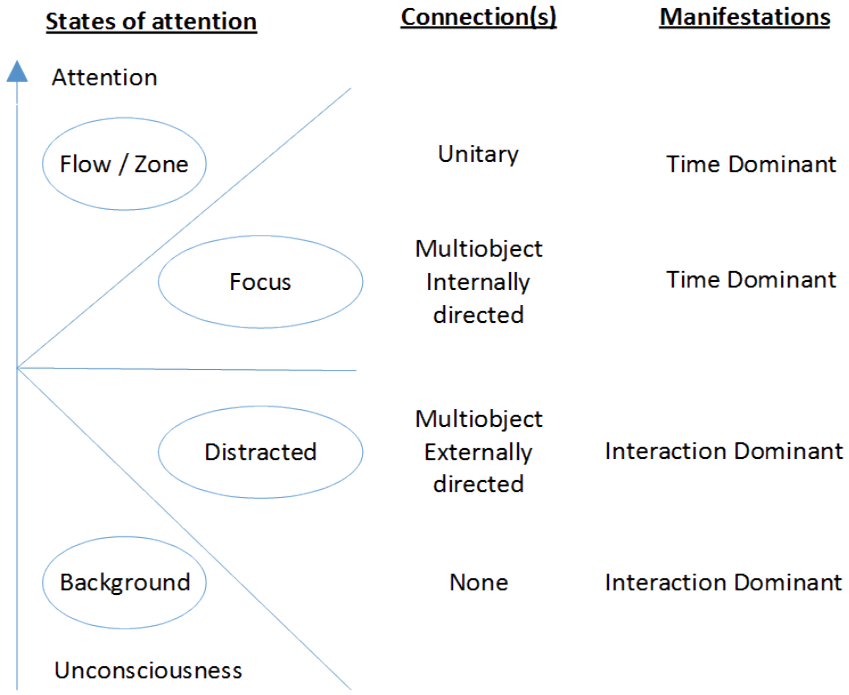


Figure 4. States of attention and their manifestations on the user-object connection.

Notice that in one end there is the Flow state (where the user is interacting with the digital object as a whole), while in the other end there is the background state (where the user knows there is a connection opened or the possibility to start one but has not engaged in it yet). Each state of attention is a diffuse division to help understanding the experiences with digital objects. Depending on the user and object aspects, as well as on the overall context described, the user can transition from one state to another, maintaining or switching between active connections. Furthermore, a connection with a user in a certain attention state may be more inclined to one type of manifestation than another: longer connection or more interaction.

- **Flow or Zone state** manifests when the user feels a sense of direction in the experience with a digital object, a connection totally excluding the other objects. The user completely abandons himself in the connection. Either the object or the user takes total control of the interaction, in a challenging progression or a repetition stimulated by the design itself. As already mentioned, Flow can be experienced in many situations, for instance while working on problem with a software tool or while playing a video-game. Instead, the Zone is reserved to structures of external reward such as videogambling. A connection involving a user in the Flow or Zone state of attention tends to last more than others.

- **Focus state** appears when the user's attention is occupied by several connections in a coordinated experience. The focus state does not limit to a single connection like the Flow/Zone, but it allows progressing in one direction towards a goal. In such a state, the user directs the interaction with one or more digital objects by avoiding other objects external to the experience. User can reorganise his priorities to maintain focus. In addition, the user can use several objects in multiple connections as long as they serve a general activity. This is a common state while working, playing or performing any other activity in which a task is constricted by rules and one or more goals. A connection involving a user in a focus state tends to manifests, first of all, in an increased time, and secondly, during the interaction, because of the reorganization of the multiple objects in use.

- **Distracted state** appears when multiple objects pop-up resulting in new connections starting while other possible connections are left for a later stage. Generally, distracted state implies pursuing several goals at a time and if instead a single goal is wanted, the user has to struggle to maintain attention on it. Between explorative and curious, the user is motivated to change the object or to explore the different inner objects which are likely to emerge from a bigger object (Marsh & Nardi, 2014). In any case, the user is externally directed by multiple objects. This state commonly manifests when surfing the Internet while working, or in a social network site. A connection involving a user in the distracted state tends to manifest in a larger number of interactions with a shorter duration.

- **Background state** manifests either when a user is aware of a new object but chooses not to focus his attention on it, or when he remembers that an active connection has been left open and could possibly be resumed. In a background state, the user can unconsciously resume a connection in order to draw further information, hence interrupting an on-going activity in focus state. Smartphones and smartwatches are a clear example of devices with digital objects in reach, likely to stimulate the user to start a connection. With a user in a background state, the most important issue to consider is the time it takes for the user to react and interact with the object.

User's attention states are related to the composition of the object the user is interacting with (for instance it is hardly possible to stay focused in social netwoks sites). Manifestations of the connection will be as varied as the wide range of digital objects. Hence, depending on each object's purpose, success in terms of engagement can be either better represented by time duration, or by the number of interactions or multiple accesses. Some objects may only be used by a user in a Flow state, while others will be used in a distracted state.

In an object that aims at a participatory type of engagement, it will be equally useful to have connections with users during multiple periods of time with interaction (e.g. logging into Twitter several times a day for a tweet) or multiple connections with several inner objects in a period of time (e.g. making several tweets directed to different Twitter users in one single access). This is very common in objects such as Social Media or Online Communities, where there is a bigger purpose as well as different inner objects which encourage different sorts of interactions.

When studying connections, defining the object composition is fundamental. Objects may appear differently depending on the scale of analysis. Namely, a social network site can be seen as a single object, although at a greater level of detail it implies different inner objects connected in a background state or a focus state depending on user's attention. To take an example which rests my case, a car race simulator can be looked at as the vehicle and the road representing the primary object, and the trees and road signs as secondary objects which appear, gain focus and disappear.

2.3.6 External Facets of the Connection

Since people live potentially attached to digital objects, there is a huge interest in measuring the connections and tracking their activity. For a designer or a manufacturer, success depends on how engaging the product is. The way the engagement of an object is rated differs. Online marketing companies and some researchers have somewhat intuitively assessed the value of each manifestation in relation to the object (Lehmann et al., 2012). Namely, certain websites measure their success in terms of short visits followed by frequent returns of visitors, while others in terms of long visits. It is only relevant to compare objects with a similar composition, purpose and functionalities, or similar groups or types of users, to see how they vary in the connection's manifestations.

In studying digital engagement, one needs to consider the aspects related to both user and object to understand which part is more determining in keeping the connection active. Yet, due to the variety of digital objects, their causal relationships cannot be determined in a single way. Engagement is multi-causal. Aspects in the user (motivation, emotion, attention and cognition) and in the object (design, content, logics and composition) are the causes of the manifestations, while the latter can be taken as the consequences. I integrate them into a conceptual model suitable for analysing engagement in a user-object connection. Nevertheless, it has to be borne in mind that this model is not presented as "comprehensive model" for every object. I only included the most common aspects, additional ones should be introduced for particular object studies (e.g. challenge, interest or aesthetic pleasure) to obtain complementary insights.

In this model, I propose *four descriptive facets* to operationalize the connection manifestations, by focusing on time, interaction or a combination of them, in order to characterize the manifestations of any connection (Figure 5). They show that an engaging user-object connection can enhance either a *faster appeal*, a *longer duration*, a *higher interaction* or a *frequent return*. Differently put, each manifestation makes it possible to assess the success of each object. I propose broad facets to encompass all the manifestations of an active connection and previous research engagement studies. This is precisely the solution in order to integrate the two different types of engagement, the one closer to user experience in the Human-Computer Interaction tradition, and the participatory type from the Social Sciences tradition. I resort to the facets to review the specific metrics employed by the current literature to assess the level of engagement.

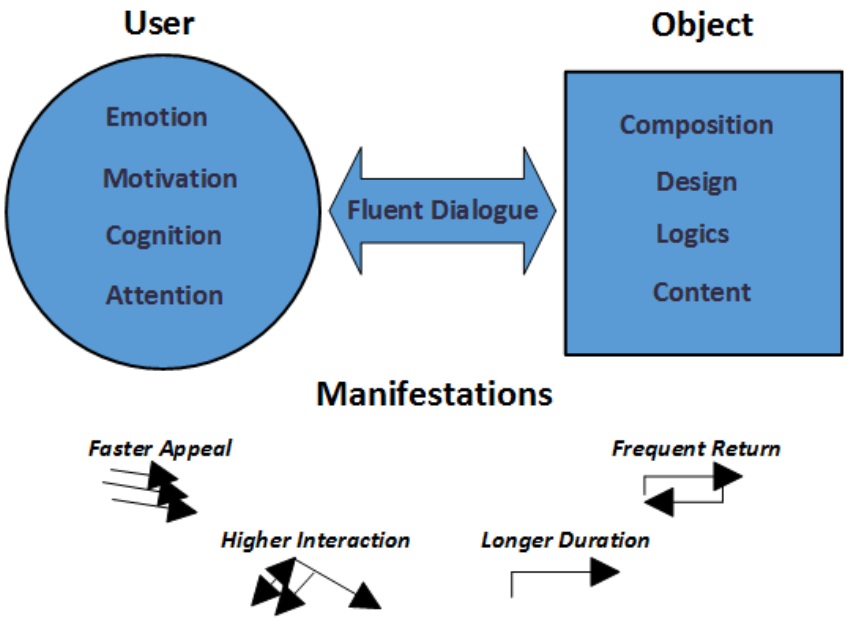


Figure 5. Model of digital engagement with aspects and manifestations.

Faster Appeal stresses the importance of the beginning of a connection. Since an engaging object catches and captivates the user's interest (Jacques, 1995), faster appeal refers to this initial period. Hence, measuring faster appeal is tantamount to assessing the time it takes from an initial point of the connection to a more advanced one, or to quantifying the number of connections initiated with an object. Faster appeal can be established either with an object or with its different inner objects (i.e. the time it would take to click on a picture on a social network site, or the number of clicks a picture receives would be measures of this inner object faster appeal). For instance, faster appeal can be used to understand the first hour of video game playing (Cheung et al., 2014) or the first days as a Wikipedia editor (Panciera, Halfaker, & Terveen, 2009). In order to consider certain objects as successful, the user should not disconnect at an initial stage.

Higher Interaction pays attention to the number of interactions in a particular connection or in an aggregation of connections. Digital marketing mostly measures the interactions of a user with an object, but higher interaction also encompasses the notifications an object sends to the user. Hence, to measure higher interaction it is important to define whether it is within a single connection or within the sum of various connections. For instance, in online news or videos websites, a higher interaction in terms of user comments or contributions has been considered a positive sign of engagement, and is also reffered to as "participation" (Ksiazek et al., 2014). In a social network site, Freyne et al. (2009) proposed the use of a recommendation tool in charge of sending messages aimed at increasing user interaction, which eventually led the user to make more contributions.

Longer Duration stresses the importance of the time spent in a connection or in an aggregation of connections. While longer duration can be measured between the engagement point and disengagement, some studies also consider the "perceived time" by the user (Arapakis et al., 2014). Time spent navigating in a website is very indicative of the type of site (Lehmann et al., 2012). For instance, in the context of video playing,

Dobrian et al. (2011) proved that the video quality had an effect on playtime and such effect was more or less intense depending on the type of video (e.g. sports or a TV show). Configurations such as a lower bitrate or buffering rate decreased viewing time.

Frequent Return pays attention to the resumption of previous connections. Some objects may not necessarily be used during a long period of time but be continuously accessed instead. If an object is engaging it will create endurability (Lalmas et al., 2014; O'Brien & Toms, 2008). This facet is usually implemented by metrics which measure the time between sessions as well as the number of times a connection has been resumed. For instance, the intersession time (also known as 'absence time') has been measured in users consulting search websites such as Questions & Answers (Dupret & Lalmas, 2013). In a way, absence time and return rate metrics can perfectly complement the metrics from the facet faster appeal. Depending on the object, these metrics are also referred to as "loyalty", "retention" or "survival" metrics. They are especially important, for instance, in measuring a customer in an e-commerce website, or an editor in Wikipedia.

2.4 Summary Review of Methods

In this section I do a brief overview of the methods most frequently employed in the measurement of engagement; I classify them and provide some case studies as examples from the current literature.

The methods or approaches to engagement are as varied as the disciplines in the Human-Computer Interaction field. The difficulty in studying engagement lies in the variety of approaches to user aspects such as motivation and attention, as well as in the measurement of the connection facets. In an ideal case study, the researcher would assess the level of engagement using all the facets of a user-object connection, and additionally investigating its cause in the aspects of both user and object (in other words, he would study whether a connection was triggered by a button in the interface design, or by a motivational trait of the user). Nevertheless, most studies limit themselves to investigating but a few manifestations of engagement and a restricted number of aspects concerning either the user or the object.

Lalmas et al. (2014) proposed a clear classification of engagement measurements based on objectivity-subjectivity and on the time of measurement. A method which obtains its data, for instance, by means of self-reported questionnaires is considered subjective, while a method relying on external measurements (behavioural measures such as Skin Conductance Activity (EDA) and Heart rate (EKG)) is considered objective. A subjective method such as self-reported questionnaires is a method employed *a posteriori*, i.e. **after** the connection took place; while another subjective method, such as for instance the think-aloud protocol, is employed **during** the connection.

Such a distinction is useful to introduce Lalmas' et al. (2014) second dimension: time, employed to distinguish between 'real-time' measurements (which measure **the process**) and *a posteriori* measurements (which measure **the product**). I broaden Lalmas' et al. (2014) classification by adding a further dimension: **measurement place**, where I

distinguish between **user-centred** and **object-centred**. Table 1 presents a joint classification of measurement methods, building on the work from Lalmas et al. (2014) and adding the measurement place dimension. The user-centred and the object-centred measures are treated in detail in sections 2.4.1 and 2.4.2 respectively.

Table 1. Classification of methods according to time approach and measurement place. *Cells present different methodologies and approaches which are aimed at obtaining data from the user or the object.*

Time Approach / Measurement Place	***A posteriori* (PRODUCT)**	**Real-time (PROCESS)**
USER-centred	**User (Subjective)** - Interview / Survey **a) USER RECALL**	**User (Both)** - Physiological **(Objective)** - Think-aloud **(Subjective)** **b) REAL TIME USER EXPERIENCE**
OBJECT-centred	**User/Object (Objective)** - Object Design aggregated changes - User Activity aggregated **c) OBJECT CHANGES - OBJECT USE**	**User/Object (Objective)** - Object Design changes - User-Object Interactions **d) USER-OBJECT INTERACTION**

2.4.1 User-centred Measures

User-centred measures are a commonplace in the study of user experience, either during the connection or *a posteriori*, as the user recalls it after it took place. The ones measuring the connection in real-time include a broad number of methods to approach user aspects such as emotion, attention, while the ones measuring it *a posteriori,* assess user aspects such as motivation.

User Recall (Product). Such measures are applied *a posteriori* and they evaluate the Product, i.e. the result of the user-object connection. The user recall measures assess what the user remembers about his experience with the object. They are self-reported measures such as questionnaires, interviews or diaries. Such measures scrutinize users' perceptions, motivations and emotions, by means of eliciting users' comments concerning the connection. An advantage of these methods is that they do not interfere with the experience itself, hence they do not introduce any bias in the users' line of reasoning or in his feelings (O'Brien & Toms, 2008). A possible drawback consists instead in the difficulty of constructing a survey without introducing any preconception in the options. A well established survey is User Engagement Scale (UES) by O'Brien (2011), which has been applied to different objects like news, e-shopping or video games (Wiebe, Lamb, Hardy, & Sharek, 2014). In the already mentioned case study (Cheung et al., 2014) on the importance of the first hour when playing videogames, users' comments posted online helped identifying which design elements were frustrating, a key information in

improving the fluent dialogue and overall experience. Some comments explicitly pointed at simulated scenes which were impossible to skip, while others stressed the importance of a good story in maintaining the attention focused. Anything reducing fluency was reported, like bad controls, low quality feedback or not showing clearly what actions could be performed next.

Real-time User Experience (Process). Such measures are applied in real-time and they evaluate the Process, i.e. the user-object connection while it takes place. User experience measures assess the user's cognitive activity and emotional arousal during the connection, in particular the emotion, attention, cognition and motivation aspects. Objective and subjective methods complement each other in rating the above-mentioned user aspects. For instance, Think Aloud Protocol is a subjective method in which users are asked to verbalize during the interaction - what they are doing or what they are thinking. This method is helpful to reveal user's anxiety provoked by some dubious aspects in design-usability. Unfortunately, it adds a reflective layer in the measurement which questions the validity of such measure (Lalmas et al., 2014). On the other hand, objective methods like psychophysiological responses in the skin, facial expressions or eye movements are very valuable as they provide even unconscious information from the user, but they can be at the same time obtrusive (Lalmas et al., 2014).

For instance, McCay-Peet et al. (2012) took a mixed approach of subjective and objective methods to the study of attention in online news, in particular, how the visual catchiness of relevant information impacts engagement. In this context, they initially used an Eye Tracker to measure the duration of first fixations, and secondly, they used a scale to measure the perceived level of attention. They found that saliency of certain objects was not a guarantee of focused attention, but that focused attention was more related to interest on the topic. In the study, when the proposed tasks were carried out with focused attention, they enhanced positive emotions in the user. In addition, the study suggested that featuring content interesting to the user could lead to a faster appeal.

2.4.2 Object-centred Measures

Object-centred measures are the core of data analysis. They consist in implementing underlying codes in digital objects which allow tracking and analysing the users' activity. These measures are able to discriminate between a passive and an active use of the object (and even modifying it).

Object changes / Object use (Product). By object changes / object use methods I refer to the methods assesing the amount of user activity within an object or the amount of changes he produced on the object. This latter aspect has not been covered enough in the engagement studies and is particularly important for Social Networks, Online Communities and more generally in User-Generated Content sites, where users post comments to videos or news articles. Much more common is the aggregated data provided by analytical tools, which show the most visited places in the website or the number of clicks, among other metrics. Aggregated data can be determinant in establishing whether in a connection there has been a higher interaction or a faster appeal.

Ksiazek et al. (2014) studied online news videos and user comments in the online video repository site Youtube. They conceptualized engagement as both user-content interactions and user-user interactions. After assessing and analysing engagement with several metrics using number of views, ratings and rankings, their main conclusion was that comments on popular videos were mostly directed to the content of the video. while less popular videos had a higher number of user-user interactions. When content was very specialized, the degree of interaction was higher among those users with common topics of discussion (Ksiazek et al., 2014).

User-Object Interaction (Process). By user-object interaction methods I refer to methods extracting time-related data from the user's behaviour. This approach in the web allows obtaining accurate data on the connection's facets such as faster appeal, longer duration and frequent return by means of the metrics suggested in the previous section. The advantage it provides is that it allows a clear comparison of two different scenarios in the same object or in two different objects. Nevertheless, quantitative data requires a contextualization and complementary insights in order to allow for a proper interpretation of its underlying reasons and causes determining the connection to stay active.

In the web, Lehmann et al. (2012), using an add-on installed in users' browsers, collected data during a year. They modelled such data discover patterns in the use of 80 online sites. They found it useful to classify the websites by popularity, activity and loyalty metrics, which would correspond to faster appeal, higher interaction and frequent return. As expected, its use was variable depending on the content - if it was news it had more appeal, while search sites had a shorter dwelling time than entertainment sites. Also, high popularity did not imply high interaction. After having found an answer to a question on a search site, or after having checked the forecast on a weather site, most users usually left. The popular websites were the websites users often returned to. These findings provide evidence that analytical approach can characterize very well the connections and it is indispensable for assessing the level of engagement in the facets.

2.5 Summary of Conclusions

Research on digital engagement sheds light on several topics of key interest in technology use. Since the late 80's, engagement with technology was based on the psychological aspects of the user. Nonetheless, the spread of different Internet applications has shaken the way and the contexts in which people use technology, ether to play games, to learn or to buy any product, and consequently require to revise the current models to study engagement. Because a different use of the term, rooted in the Social Sciences, implies a participatory sense which is now indispensable to understand these social and digital objects.

In order to conciliate these various meanings, I proposed a working definition with engagement as the quality which ensures a user-digital object connection stays active. Hence, an engaged behaviour can be either the user absorbed or participating frenetically, but in both cases, it guarantees the connection remains active. Each connection may manifest itself in a different way (longer or shorter duration, and more or less interaction).

These manifestations are measurable and can be explained by studying each part of the connection.

Hence, digital engagement model takes the connection as the unit of analysis. This view evolves from a user-centred perspective in studying HCI, which has been dominant first, since usability studies appeared, focusing on the object properties which enable task efficiency, and second, with user experience aiming at explaining the user's range of emotions and needs in relation to a product or a service. As said, the user-centred perspective is useful for designing as it helps in understanding aspects of cognition and needs, but assumes most of theories assume that the object is passive. This is not what happens nowadays or will happen in the future.

Digital engagement integrates both perspectives and considers the object and the user are intertwined in a connection by and for different reasons. This is because the study of engagement needs to go beyond motivation or a user-centred perspective, since digital objects present an active reality. I advocate that this paradigm shift will take more relevance when objects become more complex, in terms of design, both in the audiovisuals and the behaviour encoded in advanced Artificial Intelligence algorithms.

I discussed the role of several aspects from the user and from the object and their influence on the connection. In the first, psychological aspects like motivation, emotion and cognition. In the second, design, either by providing rapid feedback, usability, and by anticipating interaction. I explained how the user's attention and its different states proved to be the key factors to understand how focus is related to the type of object in terms of composition and purpose. Depending on the object's composition, purpose and the user's attention, the connection will manifest towards longer time or a higher interaction. All in all, these aspects were unified in a model, along four facets to explore the manifestations of any digital object.

For some authors, engagement has been considered a *science* (Attfield et al., 2011), due to the extensive interdisciplinary literature and the increasing complexity of measuring it. The conceptual model presented in this paper is expected to help stimulate research on both direction and impetus. Once a clear definition and model are set, the real challenge lies on measurement. Different methods and approaches have been sketched to show how to operationalise its facets. Current research examples have been selected to show the usefulness and appropriateness of the explained concepts. Studying engagement may lead to improve any object's design, whose use can be tracked and support specific changes in an iterative process of design. Technological progress is taking place at a great pace, and some exciting avenues for future research into the connections between users and digital objects lie ahead.

2.6 Identity in Digital Engagement

Participation is the manifestation of digital engagement which can best explain the way we use social and digital objects. Yet, the success of these objects lies in how they accommodate their user's interactions, but also in how they motivate and emphasize both the individuality and commonalities of users - the two first meanings in the definition of 'identity'[2]. I suspect that identity and motivation have an influence on increasing the users' participation in social objects - namely, the interaction manifestations of digital engagement. In this sense, I propose a brief overview of several digital objects with different purposes, to see how they allow their users to represent their identities:

Massive Multiplayer Online Games (MMOG). These games allow players to complete with each other simultaneously in the same instance. The role-playing type (Massive Multiplyaer Role-Playing Games) encourage players to construct their identities largely by using social practices (e.g., behaviours, communication styles), virtual objects and roles (e.g. be a princess, elf, etc.) hence promoting an identity-leveraged experience. Steinkuehler (2006) studied MMRPOG in the context of education in order to understand the opportunities for learning, and concluded that identity can be a factor to balance between learning and playing, the two activities supporting each other in the narrative.

Massive Open Online Courses (MOOC). These courses allow students to take free lessons online. Cassidy, Breakwell and Bailey (2014) studied student engagement in relation to workload, tasks and facilitation. Results showed that the level of participation was directly proportional to the workload. While observing participant feedback, Cassidy et al. (2004) point out the links between personal identity and intrinsic motivation.

Social Media. These digital objects "employ mobile and web-based technologies to create highly interactive platforms via which individuals and communities share, co-create, discuss and modify user-generated content" (Kietzmann, Hermkens, McCarthy, & Silvestre, 2011). According to Kietzmann et al. (2011), identity is the first and central block. Users are motivated to consciously or unconsciously self-disclose personal information such as thoughts, feelings, preferences, consistently with the image they want to give of themselves. This is a step in the development of relationships (Kaplan).

Online Community. These digital objects are defined as platforms where a group of people can interact sharing a goal, need or topic (Porter, 2006; Preece, 2000; Yuqing Ren, Kraut, & Kiesler, 2007). In this context, users tend to develop a common identity, which makes them be more attached to the platform purpose. Ren (2007) studied the importance of building a common identity in online communities, along with the importance of creating bonds with other users by disclosing other identity information, concluding that both were relevant for engagement.

Taking all this into account, identity appears as a relevant concept in social objects, being an integral part of the user's motivation and related to the many aspects of the social interactions.

---

[2] http://www.dictionary.com/browse/identity

I consider Wikipedia is a suitable object to study the influence of identity on participation and this is the reason why I chose it for this research. Wikipedia is often referred to as an online community (Halfaker, Geiger, Morgan, & Riedl, 2013a; Kittur, Chi, Pendleton, & Suh, 2007) with an educational purpose, where participation is key for its development. To my knowledge, no research has applied the concept of identity to the study of Wikipedia. Despite the abundant research on the object Wikipedia (Mesgari, Okoli, Mehdi, Nielsen, & Lanamäki, 2015; Okoli, Mehdi, Mesgari, Nielsen, & Lanamäki, 2012), no study has yet presented an holistic view on all the possible aspects which contribute to how Wikipedia editors are engaged.

In the following Part 2, I apply the model of engagement to Wikipedia. Then, in Part 3, I conduct empirical research to study the influence of identity on Wikipedia editor engagement.

"Given enough eyeballs, all bugs are shallow" (Linus's Law; The Cathedral and the Bazaar) Eric S. Raymond

PART 2: WIKIPEDIA EDITOR ENGAGEMENT

Chapter 3. Past and Present of Wikipedia Editor Engagement

3.1 Introduction

Wikipedia is the most popular general reference site in the Internet. Millions of users from all over the world access this online encyclopaedia to obtain information, and they do so through all kind of devices. For more than five years it has been among the seven most visited websites in the Internet, and it represents an educational resource. Not surprisingly, Wikipedia also covers information about current news and events, which are far more read than other topics (Keegan, Gergle, & Contractor, 2013; Miljesic & Ricchiuti, 2016).

Wikipedia's most striking characteristic is the fact that it is a collaborative project: everybody can become a volunteer contributor and join the project. In its beginnings, co-founder Jimmy Wales thought of it as an experiment, and many doubted it would succeed (Lih, 2009). At present, there are 291 Wikipedia language editions[3], English being the largest with more than 5 million articles (and a total of 40 million articles counting all the languages). Wikipedia's goal is to provide the "sum of human knowledge", available to everyone for free. This makes Wikipedia unique and difficult to be encapsulated in a single definition, something in between an online community and an encyclopaedia.

In spite of this undeniable achievement, Wikipedia is not a "finished product", but an on-going process which focuses on content; it is the nexus between editors and readers, who access different parts of the site with different needs and motivations. In its beginnings, this duality of users has acted similarly to a feedback loop system: the availability of new content contributed to popularise the encyclopaedia and improved the position for searchers, which in turn increased the use of Wikipedia and of its editing community who created new articles. This could be seen as a self-reinforcing mechanism (Suh et al., 2009): the more valuable Wikipedia became, the more contributors joined it to bring even more value.

In Wikipedia, engagement is mainly referred to as participation, which is essential as it makes the project grow. Furthermore, Wikipedia's success depends on editor participation, but also on attracting newcomers and retaining them. Lowering the barriers to entry by accepting everyone - even let users act anonymously – contributed in 2006 to a great increase in the number of editors. However, during the last few years there has been a clear decline in the number of editors. A minority is in charge for most of the activities (Ortega, González-Barahona, & Robles, 2008; Voss, 2005), namely the editors who joined in 2006. At the same time, attracting new editors is difficult (Halfaker, Geiger, Morgan, & Riedl, 2013a).

[3] Throughout this thesis, constructions such as English Wikipedia, English Wikipedia language edition, English language edition and English language are used interchangeably.

How to stimulate the participation of consolidated editors and how to retain the new ones are usual concerns in online communities (Yuqing Ren et al., 2007). In Wikipedia, many measurements have been applied to identify the seriousness of editor decline, as well as the causes which refrain new editors to settle in the community. Nonetheless, no study investigates in a holistic and extensive way all the aspects and manifestations of engagement. This is paradoxical, considering that Wikipedia's success and unique characteristics have favoured the publication of numerous peer-reviewed academic articles investigating its social dynamics, content and readership (Mesgari et al., 2015; Okoli, 2009; 2014).

Wikimedia Foundation, the non-profit organisation developing and maintaining the technological side of the project, has dedicated efforts to study and improve editor engagement and even set a team in 2012. Although increasing editor engagement is listed among its goals, a great range of other initiatives are being developed: new editor tools, events and movement promotion. Hence, while the project is still growing in terms of articles, and each language edition community proposes new milestones to be achieved, the issue of editor decline finds no solution.

In Part 1 I claimed that Wikipedia could be a suitable object to study the influence of identity on editor engagement. I try to prove this claim in Part 3. Before that, In Part 2 I propose deepening into Wikipedia, its functioning and its organisational values, and as **Thesis Objective 2,** I pursue the objective of understanding the aspects which influence Wikipedia editor engagement, by reviewing current research studies. This part is composed by a single chapter, which is organised as follows:

In Section 3.2, I first present what Wikipedia is and how it was created according to its origins and cultural values. Then, in Section 3.3, I describe the main rules which sustain the content creation practice as well as the community governance structure. In Section 3.4, I evaluate each of the aspects of engagement that have been individually studied in Wikipedia. In Section 3.5 I review the different measurements of engagement, taking into account participation and retention. In Section 3.6, I discuss the role of the different actors involved in Wikipedia community and foundation, in order to understand their priorities in the technology design process and how this impacts on the retention of new editors. Finally, in Section 3.7 I conclude with a discussion.

3.2 What is Wikipedia?

Hacker Ethics and the 'wiki'. Wikipedia did not appear out of the blue in 2001. When Jimmy Wales and Larry Sanger started the project Wikipedia, they were already involved in a free encyclopaedia called *Nupedia*, which responded to the exact same purpose: "a free-access, free-content Internet encyclopaedia"[4]. The idea of creating a free content encyclopaedia was inspired by GNU licenses[5], used by software and usually developed

[4] https://en.wikipedia.org/wiki/Wikipedia
[5] https://www.gnu.org/licenses/licenses.html

by experts in the field. In that same year, free culture started to find a structure and free licenses "Creative Commons" were also created.

Wikipedia arose in this context, surrounded by other projects in the free software and culture movement, which are based on a hacker culture and ethics. Hackers enjoy overcoming challenges to prove their worth to themselves; their culture is one of self-empowerment (Gehring, 2004). In their view, everyone should share their knowledge so that the others can benefit from it. In fact, principles of the Hacker Ethics (Levy, 1984) such as sharing, openness, free access and world improvement are present in Wikipedia. The only lacking principle is 'decentralization': Wikipedia centralizes all the content in its website, as encyclopaedias do.

But the key to Wikipedia's success – as compared to *Nupedia* – was the implementation of the wiki technology: a type of website which allows collaborative modifications directly from the browser. Even though at that time Wikipedia was a secondary experiment to help *Nupedia*, they realized the project was benefiting from an influx of thousands of volunteers, and rapidly surpassed in value the original encyclopaedia (Lih, 2009). It became an encyclopaedia that anyone can edit, at any time and evolved into the biggest example of User-Generated Content (UGC) site in the Internet, governed by its own editors and ready for mass consumption.

Even though the 'wiki' interface helped users become contributors of the encyclopaedia, the website had to gradually evolve to accommodate a community revolving around its altruistic free knowledge dissemination scope. For instance, each encyclopaedic article page had attached a talk page where editors could discuss the appropriateness of the content displayed. Later, user pages were created for each registered user, in order to allow them set a description and introduce themselves to the community. Other types of pages like the ones called 'Wikipedia' were dedicated to house policies, essays, among other community aspects. The wiki system transformed the site into a big hypertext, where every page could be linked to any other page. In a way, all these non-article spaces served specific communicative and management purposes so that editors could work towards the encyclopaedic goal of "gathering the sum of human knowledge".

Encyclopaedia, Online Community and Social Network. Since its creation, Wikipedia has remained loyal to its principles and to its free content goal, close to free culture and to open source initiatives like Firefox Foundation and Open Street Maps. Based on these values, Wikipedia editors organise themselves as a community and use specific pages from the site to develop the project. Instead, readers are users who only access the final content, and may not be aware of the community and of the possibility to contribute (Halfaker, Keyes, & Taraborelli, 2013b). Because of this, Wikipedia is a *sui generis* object, difficult to locate into a single category; it is defined as an encyclopaedia outwards, and an open community inwards[6].

Indeed, in past literature, Wikipedia has been referred to as an online community (Halfaker, Geiger, Morgan, & Riedl, 2013a; Kittur et al., 2007). The case of Wikipedia fits the definition of an Online Community, often defined as a group of people who

[6] https://en.wikipedia.org/wiki/Wikipedia_community

interact through the Internet in order to work towards a shared goal, need, thematic interest or purpose (Porter, 2006; Preece, 2000; Yuqing Ren et al., 2007). Editors share the ethical purpose of providing a free repository of encyclopaedic content. *Vice versa*, specific pages within Wikipedia discuss why it should not be considered a social network[7]. Yet, Wikipedia would probably fit one of the most common definitions of Social Media, defined as "a site which employs mobile and web-based technologies to create highly interactive platforms via which individuals and communities share, co-create, discuss and modify user-generated content" (Kietzmann et al., 2011).

In particular, the main reason against considering Wikipedia a Social Media is that every space must be focused on creating a high-quality encyclopaedia. In fact, according to Kietzmann, Hermkens, McCarthy & Silvestre (2011), Social Media types are the sum of the following characteristics: sharing, conversations, groups, reputation, relationships, presence, sharing and identity. All of them exist in Wikipedia (see Figure 6) but, while Kietzmann et al. (2011) consider identity the first and most core aspect of a Social Media, in Wikipedia the most fundamental is the shared content. Content is the focus of Wikipedia, and its centralization in the site is the main goal. However, leaving identity in a secondary position and even holding contradictions does not mean that Wikipedia is not affected. Identity is a fundamental concept to understand editors' behaviours, as I will explain in Part 3.

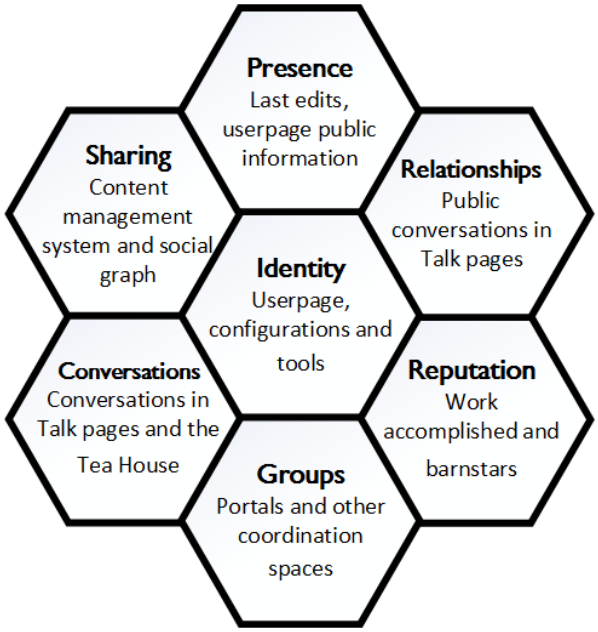


Figure 6. Wikipedia spaces in agreement with the Social Media functional blocks (Kietzmann et al. 2011).

[7] https://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not

Regarding the rest of Social Media aspects in Wikipedia, editors' presence can be inferred from their last edits and user page; relationships are maintained through talk pages, and groups are constantly created around specific themes in inner pages called portals. The rest of functional blocks can be deduced from tools and services editors use; some functionalities are even decentralized and have not been implemented in the site, as for instance the conversations and groups in social media sites such as Facebook or synchronous conversation channels in IRC (Internet Relay Chat)[8].

Another remarkable difference from Social Media sites is the fact that Wikipedia adds one more layer of complexity with the governance of its content and community (see the next section). All in all, the only social network to consider in Wikipedia is the structure of editors' interactions (Kane, 2009) - not the intent or scope of a social network site.

3.3 How Does Wikipedia Work?

How does Wikipedia work? Unsurprisingly, not very different from other organisations, namely, with bureaucracy. The technology wiki supports the development of content in a flexible and open way. However, the need for a governance system has been solved by a gradual implementation of rules, policies, guidelines and different types of roles - some of them are general to all language editions. However, language editions also maintain a certain degree of autonomy to create specific rules according to their particularities.

The document containing the fundamental principles of Wikipedia is called 'Five pillars'[9]. Such pillars stand for the project's scope ('Wikipedia is an encyclopaedia'), the golden rule of project content ('Wikipedia is written from a neutral point of view'), the project's main ethical characteristic ('Wikipedia is free content that anyone can use, edit, and distribute'), and a conduct recommendation ('Editors should treat each other with respect and civility'). The fifth pillar states that 'Wikipedia has no firm rules', hence softening the importance of rules and encouraging editors to be bold and act in favour of the project.

Probably the most important rule is the 'Neutral Point of View' (NPOV)[10]. This central content policy roughly means that all the different editors' positions must be represented in the text. This way, all points of view (also opposite) have a place in the text. This contrasts with the idea of objectivity, because neutrality only asks for a fair weighted representation of the current points of view. The only requirement is that every piece of information needs verifiable source (the rule 'Verifiability'[11]). This invalidates personal opinions or full articles with authorship.

Two other core content rules delimit what is accepted in the encyclopaedia: 'Notability'[12] and 'No Original Research'[13]. For 'Notability', editors judge whether a specific topic

[8] https://meta.wikimedia.org/wiki/Category:Communication
[9] https://en.wikipedia.org/wiki/Wikipedia:Five_pillars
[10] https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view
[11] https://en.wikipedia.org/wiki/Wikipedia:Verifiability
[12] https://en.wikipedia.org/wiki/Wikipedia:Notability
[13] https://en.wikipedia.org/wiki/Wikipedia:No_original_research

deserves an article. In some cases, content may be interesting to exist, although in a section of another article. In others, it is not notable because it does not fit a criterion of sources (or it responds to commercial purposes). 'No Original Research' simply states that Wikipedia should only accept material already published in other sources.

By asking civility, editors are expected to show good manners and respect when considering others' points of view displayed in the content of an article; the same goes for any other Wikipedia space like discussions and help pages. Editors should assume good faith, before expressing any difference. Before making any decision in Wikipedia, there is a prior debate aiming at achieving consensus on what is the best resolution. This conduct rule stays as a fundamental value of the Wikipedia culture, and dominates all kinds of decisions: from granting a right to a user, to changing content or updating a tool used by the entire community.

Consensus is implemented by Wikipedians (the volunteers who edit and contribute to the project; they are called Wikipedians to differentiate them from readers). However, some specific functional roles are assigned to some community members in order to preserve content, solve disputes and welcome new editors. These editors hold a flag, which grant them special permissions or specializes them into specific tasks. Among them, the administrators are the most relevant group of privileged editors, as they can perform special actions to pages (like deletion or protection) or to editors (like blocking them from editing).

All in all, despite the brave claim encouraging to 'ignore all rules', Wikipedia has increased its documental complexity of roles, documents, policies and guidelines of all kind, which specify in detail how to perform specific tasks (Butler, Joyce, & Pike, 2008). Several studies even affirm that Wikipedia, as a socio-technological artefact, has found a maturity in these structures and "norm networks", obstructing the incorporation of new rules from new editors (Butler et al., 2008; Halfaker, Geiger, Morgan, & Riedl, 2013a; Heaberlin & DeDeo, 2016).

3.4 Literature Review of Aspects of Wikipedia Editor Engagement

In this section, I apply the digital engagement model (see Chapter 2) to Wikipedia, I discuss its aspects and review the literature. I will focus on those aspects which have a direct effect on engagement: I dedicate a section to each aspect and succinctly explain the available peer-reviewed studies to date, in a similar way to the various scholarly research reviews on Wikipedia (Mesgari et al., 2015; Okoli, 2009).

Some of the studies only describe the aspect without evaluating its consequences on engagement. Likewise, some aspects have different or even contrary implications for editors, depending on whether they are experienced or newcomers. I first describe the components that allow a fluent dialogue between an editor and Wikipedia, and afterwards I focus on the editor's emotions and motivations. Finally, I describe the object strategies employed to stimulate continuity.

3.4.1 The Components of the Fluent Dialogue

The fluent dialogue is a requirement for the Wikipedia - editor connection to continue. It is a trade-off between editor's cognition – necessary for knowing how to act – and the object's usability to ease its use. This means every editor needs to understand what can be done next and how to do it. This depends both on his cognition and current knowledge and on the object's design characteristics. I distinguish three different types of literacy which need to be acquired: *technical design, norms* and *community*. It would be possible to argue a fourth, *knowledge and writing literacy*, but given the wide variety of tasks available in Wikipedia which go from fact correction to proofreading, I consider that this can be acquired while being engaged in Wikipedia.

These literacy types imply a learning for any reader who wants to become a contributor. In fact, Antin & Cheshire (2010) studied the differences in readers' knowledge and their predisposition to contribute, and found that they could become Wikipedians, if they better understood how. According to Antin and Cheshire (2010, p. 130), readers "are deliberately cautious individuals, dipping their toes in to passively participate while learning more about a complex system". By means of a survey, Antin & Cheshire, (2010) found out that readers become familiar with the editing interface, policies and roles of the encyclopaedia before registering, and only a 10% of the participants knew of the existence of the policy "No Original Research".

Firstly, **the technical-design literacy** is a long-debated issue, since several studies considered at their time that the usability of the MediaWiki technology has significant room for improvement. For instance, in the field of education, Raitman et al. (2005) used a wiki and concluded that it had a poor interface and was cluttered. In higher education, Ebner et al. (2008) experimented with the use of wikis to engage students. Even though many pedagogical factors influence on the students' performance, the study concluded that a wiki was not a proper tool for assignments. Among the survey answers, bad usability appears as one of the barriers and therefore a potential reason for students' few editing.

In 2009, a remote testing organized by Bolt Peters and the Wikimedia Foundation identified the obstacles in creating new articles by evaluating how novices interact with Wikipedia[14]. One of their findings showed that many contributors did not notice the edit button and found unintuitive the use of the wiki-markup (a similar language to html with specific tags native from the MediaWiki system). Cowan (2011) identified the syntax as a hurdle that often overwhelms editors. To solve some of these issues, a MediaWiki extension named VisualEditor was released in 2012 to provide a What You See Is What You Get Editor, which allows editing the same way as writing in a word processor. Other technical aspects pointed out by Cowan (2011) were the formatting and the templates[15, 16, 17], among others.

[14] https://usability.wikimedia.org/wiki/Usability_and_Experience_Study
[15] https://en.wikipedia.org/wiki/Wikipedia:Templates
[16] https://en.wikipedia.org/wiki/Wikipedia:Tutorial/Formatting
[17] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Infoboxes

Secondly, **the rule literacy** to become a fully operative editor can be considered challenging and stimulating. The growth in number of policies, guidelines, and documentation has been reported by several studies (Butler et al., 2008; Heaberlin & DeDeo, 2016). This increased complexity is considered a cost with negative impact on production (Suh et al., 2009). In this sense, the Wikimedia Foundation has created a guided tour in which they teach you how to edit according to the most important rules - unfortunately it is available in few more than ten languages[18].

Thirdly and finally, **the community or social literacy** is also a requirement for editors to understand the community dynamics and find a place in it. In one of the previous studies on usability, Cowan (2011) states that being judged by other peers in a wiki creates a concern for new editors, who may be anxious about the quality of their contributions (accurateness or validity) in front of the entire community, or even in front of the readers. In addition, not all new editors assume the idea of co-ownership in content creation, and they also feel anxious about how other editors delete or amend what they consider their content.

Taken all together, the different types of literacy an editor must acquire in order to contribute to Wikipedia imply that the initial period of time after registering is a key period, and it can ease this process. Community spaces like 'Teahouse' in the English Wikipedia and their equivalents in other languages serve as a place where newcomers can ask more experienced editors questions on any topic, from the process of contributing content, to the use of their personal User Pages (Morgan, Bouterse, Walls, & Stierch, 2013).

Even though in theory the required literacy to edit in Wikipedia can be taught to everyone, an analysis of the current community of editors shows that the most common contributor profile is the high-skilled male (Hargittai et al. 2015). Hargittai et al. (2015) found that the gender gap in editing, or the lack of women in the community, is worsen by a similarly important Internet skills gap. This means that people's background and knowledge prior to start learning about the Wikipedia literacies are important factors which determine whether they will be able to succeed in this process or the frustration will determine them to abandon.

3.4.2 Editors' Emotions

Interacting in a Wikipedia editing process make editors experience a wide range of emotions. Some of the most positive emotions are often related to the motivation type (which are covered in the next section). As already mentioned, Cowan (2011) argued about the anxiety produced by learning how to act in the community, in such a transparent environment like a wiki. Nonetheless, the most important studies about emotions are few, and specifically focus on understanding the emotional dimension of communication.

[18] https://en.wikipedia.org/wiki/Wikipedia:The_Wikipedia_Adventure

In this sense, Laniado, Kaltenbrunner, Castillo & Morell (2012) studied emotions within Wikipedia discussion pages. They employed the ratings of Affective Norms for English Words (ANEW) in order to quantify the emotional tone of the different conversations. The abovementioned authors found evidence that female editors tend to use a more positive tone, and that administrators were more positive than non-administrators. Generally, editors tended to reply with a more positive tone in other editors' talk pages than in their own user page. Inspired by these results, Laniado et al. (2012) recommended an appropriate wording, as well as providing new ways of channelling negative feelings. For new editors who are still discovering how to operate within a system, receiving positive messages is very important.

In a similar vein, Iosub et al. (2014) explored the relationship between gender, status and communication, by applying some lexicon-based computational methods. Some of their findings insist that female editors' communication style create stronger emotional connections than that of male editors; in fact, they insist that their values are not dependent on their position or status in the community. Female editors make more relationship-oriented choices, while male administrators are less oriented towards building relationships. Another result which confirms (Laniado et al., 2012) was that editors interact more with other peers with a similar emotional style (e.g. editors showing higher levels of anger communicate more among themselves).

Also on women editors, Menking and Erickson (2015) proposed an interview-based study on the editors' emotional management in communication. They observed how women, when in a marginalized situation, could change or suppress some feelings (known as 'emotional work') in order to continue contributing. For this, they reviewed the factors at the basis of the gender gap in editors' population, and interviewed 20 women editors. Some of the current female editors preferred to deal with conflict by not to argue and avoid wasting energy. Others revealed that they reached the conclusion that they should ignore harassment. All in all, Menking and Erickson (2015) conclude that the way Wikipedia is constructed requires emotional labour as a cost in the project, and suggests solutions to raise awareness to the present state of affairs and end the gender gap. The study points that suppressing emotions and facing harassment is not only a women problem. In fact, harassment is a problem already identified in Wikipedia community and there are campaigns to ban its manifestations[19].

3.4.3 Motivations for Editor Continuity

Given that Wikipedia is made by volunteers, it is very interesting to find out what motivates them to contribute. Motivation studies try to understand what are the reasons that push somebody into an action, and in Wikipedia, there is abundant literature on the topic which reveals how these reasons emerge from the object characteristics. Studies usually have a preference for certain methodologies such as surveys and qualitative research.

[19] https://en.wikipedia.org/wiki/Wikipedia:Harassment

One of the first studies surveyed editors from English Wikipedia in order to discover their main motivations for editing and found out that fun, ideology and values were the most usual and significant answers from editors (Nov, 2007). Other reasons were the same process of learning skills, socializing with other peers, and developing a writing career. Kuznetsov (2006) identified the main reasons by which Wikipedians are motivated to contribute using an iterative methodology named Value Sensitive Design (VSD) which consists in an empirical, conceptual and technical investigation. This author noticed that a prevailing majority would edit for ideological reasons, and as an exchange with the Wikipedia community. In fact, these reasons are similar to the ones found by Nov (2007).

Forte and Bruckman (2008) took a qualitative approach and interviewed 22 volunteer Wikipedians in order to understand their main motivations. By applying a Latour concept of 'cycle of credit', they saw that making a name in the community and gaining authority was an incentive to continue contributing. In Wikipedia, as compared to other online communities, acquiring technological power or skills do not represent important factors when it comes to contributing to the project, but it is more important to obtain credibility and gain recognition.

Yang et al. (2010) surveyed editors and found that an internal self-concept is the most important motivation, even more than reputation, accomplishment and gaining autonomy, which were still relevant. Their approach differs from the others by modelling internal and external self-concept, in addition to intrinsic and extrinsic motivation. According to Yang et al. (2010), the internal self-concept only depends on self-evaluation, but is not linked to on enjoyment like intrinsic motivation. Instead, the internal self-concept motivation refers to making decisions consistent with the personal standards and the self-evaluation of achievement.

Other studies applied the Self-Determination Theory, distinguishing extrinsic and intrinsic motivation. For instance, Xu and Li (2015) classified different motivational factors into content contribution and community participation. By means of a survey on Chinese Wikipedia, they found that content contribution was driven by extrinsically oriented motivations such as reciprocity and self-development; while community participation was enhanced by altruism and sense of belonging. Zhang and Zhu (2011) analysed contributions in Chinese Wikipedia and found that editors are mostly moved by intrinsic motives - that is, editors find pleasure in the writing task itself without the need for external evaluations or rewards.

All in all, studies mostly agree on the list of motives that push Wikipedians to contribute, but disagree when it comes to assessing which are most important. This may be due to different methodologies, different editor communities' populations sizes and, also, their cultural background, which according to Pfeil et al. (2006) influences several aspects of behaviour. Therefore, it is possible to affirm that the diversity of studies shows the daunting complexity in determining the composition and the importance of each motivation type in Wikipedia.

I want to highlight that I see reasons to think of identity as a possible source of motivation in Wikipedia. Besides some motives focused on self-development (such as learning how to write) and the underlying free knowledge ideology in Wikipedia, I see there are several

other motivation types based on the social aspects of Wikipedia. Building a reputation or acting as an autonomous editor within a community implies that there is a sense of relatedness with the other peers. Therefore, this goes in the direction of building an identity within the community - even one author stressed the importance in the self-concept. However, the specific ways this identity is shaped by the community and by the Wikipedia project, and its representation are different matters which are related to but go beyond motivation. Such insight will be further developed in Part 3.

3.4.4 Design, Content and Social Continuity

Motivation studies explain the reasons why editors want to continue in a connection with Wikipedia. A different aspect is to understand how the interaction and the continual choices presented by the object also foster continuity. In the case of Wikipedia, continuity can be driven by the technological design, continual content changes and the social aspects of communication between peers. This often implies that the effects of these interactions are different for newcomers and for experienced editors.

Design continuity exists when changes in the object encourage editors to continue interacting. In Wikipedia, one of the few and most popular tools for this is the "Watchlist", which is self-managed from the editor interface and it enables tracking changes in Wikipedia articles, and is often used by experienced editors (Forte & Bruckman, 2008). In a similar way, external tools such as the GapFinder designed by the Wikimedia Foundation recommends articles present in a certain language to editors from different languages where such articles do not exist, encouraging this way the creation of the lacking articles in new language editions. This article recommendation (or discovery) tool is based on a research study (Wulczyn, West, Zia, & Leskovec, 2016), which models different content features, along with editor topical preferences and editing history in order to provide both popular and interesting articles for the editor. After testing it with 12000 editors from French Wikipedia, Wulczyn et al. (2016) found that personalizing recommendations increased editors' engagement in terms of article creation by a factor of two.

A different source of object-directed continuity comes from messages sent by bots, which are usually a way of preventing bad behaviours and vandalism (Halfaker, Geiger, Morgan, & Riedl, 2013a). Even though these messages respond to specific purposes, and have been referred to as an 'immune system', they also produce undesired effects.
Halfaker et al. (2013) tested the use of bot warnings and rejections in newcomers' contributions by means of regression analysis, and results showed a significant increase over desirable new editors. The authors suspect that these algorithmic automatized tools have an undesired effect over these new editors' contribution. To prevent such outcome, the Wikimedia Foundation, together with authors from the above-mentioned study (Halfaker, Geiger, Morgan, & Riedl, 2013a) are developing a system called ORES (Objective Revision Evaluation Service)[20], which attempts to incorporate more sophisticated algorithms not to mislabel contributions from new editors, and encourage sending appropriate informational messages.

[20] https://meta.wikimedia.org/wiki/Objective_Revision_Evaluation_Service

Content continuity exists when the changes in the articles' content encourage editors to continue interacting. Keegan et al. 2012 (2012) explored editing patterns with the aim of understanding whether contributing to an article was more influenced by the editor's experience, by the editing history of the article or by the demands of other editors and their characteristics. They applied a statistical method called p*/exponential random graph models (p*/ERGMs) in order to make a multi-level network analysis. Their results showed that the previous editors' experience and the article editing history were more important than any other factor (as for instance the experience of contributors to that article) for an article to obtain more edits. Similarly, Aaltonen and Seiler (2015) studied the progressive growing of content to understand a possible cumulative growth effect, in other words, whether articles which are heavily edited and reach a higher length are more likely to be edited more in the future. They modelled factors such as articles' topics popularity and general growth trends. Results suggested that articles would have been 45% shorter without the cumulative effect of years. Therefore, Aaltonen and Seiler (2015) concluded that editors are more encouraged to edit an already edited article since it lowers the editing costs.

While these studies focused on the effect of content and its changes in new interactions, other studies focused on the effect of reverts, a particular action which allows undoing one or more edits and returning to a previous state of the article. This type of action is usually done by a more experienced editor, and it could be considered a social interaction mediated by content. For instance, Suh et al. (2009) studied the number of reverts-per-edits (or new contributions rejected) in English Wikipedia and found out that they doubled from 2005 to 2008 (2.9% to 6%). Halfaker et al. (2011a) studied the effects of reverts on the quantity and the quality of newcomers' contributions. After applying regressions on different activity indicators, they found out that reverts drastically affect newcomers' future activity; in some cases, a revert only questioned the quality of the work, as if being reverted could be part of the learning experience. In this sense, Halfaker et al. (2011b), proposed an interface change to inform editors about the reasons of the revert. After testing it in a trial group, Halfaker et al. (2011b) found that a simple warning message could improve the involvement and content quality of editors with different degrees of experience.

In a similar way to reverts, Halfaker et al. (2013a) studied the newcomers (i.e. new editors) contributions to norms. Results showed that while norms had been revised and expanded, new ones did not emerge at the same pace since 2006. This was mainly due to the fact that newer editors were finding their policy propositions mostly rejected as compared to those of editors from earlier times. Contrarily to what happened in the case of reverts, this did not stop newer editors from contributing to essays and other spaces of community governance, although essays are not as official as policies and therefore were not applied in the same way.

Other studies (H. Zhu, Zhang, He, Kraut, & Kittur, 2013) analysed peer feedback in a more broad way. They characterized feedback (positive, negative, directive and social) according to the tone and measured its effect on contributions. Their results showed that positive feedback and social feedback influenced the quantity of work but did not have effects on focal tasks. And, unlike Halfaker et al. (2011a), they found that negative

feedback on newcomers did not decrease their future interactions but encouraged them to work harder.

It is worth mentioning that the leaders of the Wikipedian community employ different content strategies to encourage continuity on the part of their members. Wikiprojects and Challenges constitute a usual and effective mean of community coordination to work on specific topics. The former, Wikiprojects, are to be found in the Wikipedia pages dedicated to topics which should be turned into articles; in there, editors often list and organise the articles they plan to write. Regarding Wikiprojects, Kittur et al. (2009b) studied how participating to such projects affected diverse aspects of editors' future behaviour. One of them was a significant but very moderate increase in their participation (1.6% in total edits). Instead, participation to Wikiprojects greatly influences increasing the interactions on discussion pages and with other editors.

Even though no research studies explore the success of Community Challenges[21], these are very similar to WikiProjects in that they are about a topic, although they are more time-restrictive projects, and set specific goals and prizes for their participants. These Challenges also have their specific Wikipedia page, which is often advertised through the newsletter – this is why it is often difficult for newcomers to take note of the challenges. Therefore, the two types of content-social continuity are effective, although their visibility is often not obvious outside the community.

Social continuity exists when peer communication encourages new interaction. Unlike communication through content, personal communication has not been much examined in relation to editors' continuity. In one study, Morgan et al. (2013) measured the influence of interacting in the Teahouse (a Wikipedia space self-defined as "a friendly place to learn about editing Wikipedia") on newcomers. Among the editors who visited this social space, the study took into account those editors with less than 100 edits and proposed them a survey to assess their satisfaction. The study also measured the future edits. Results showed that they increased their participation in number of edits, and especially in discussion pages. Therefore, this type of social interactions is useful to help newcomers settle in the community, and encourages new interaction.

Tsikerdekis (2015) modelled personal messages in user's talk pages by using a p* model with the goal of understanding its effect on consolidated editors. Results showed a link between being in the personal network and contributing with high quality content. Even though very engaged editors also contribute to more high quality articles, they do not necessarily participate on the personal communication network. Tsikerdekis (2015) emphasizes the essential role played by communication channels in a collaborative project such as Wikipedia.

Lastly, Biuk-Aghai and Hong Lei (2010) discussed the possibility of introducing instant messaging (i.e. synchronous communication) in a wiki, also considering the benefits it would have on coordinating efforts while editing articles. At present, Wikipedia only employs asynchronous forms of communication, and it has always been very cautious when implementing channels where editors could discuss and socialize (Lih, 2009). As a

[21] https://en.wikipedia.org/wiki/Wikipedia:Catalan_culture_challenge

consequence, Wikipedians often access open-standard chats like IRC (Internet Relay Chat) and Facebook in order to have synchronous communication channels (i.e. chat). They have different windows simultaneously opened; one in Wikipedia to edit the articles, and another with a IRC client or Facebook.

3.5 Literature Review of Measurements and Experiments on Wikipedia Editor Engagement

In this section, I deepen into the studies measuring Wikipedia editor engagement and I present the main experiments with interface changes or tool proposals intended to foster engagement. While the measurements were mostly undertaken by researchers in the Academia, most of these experiments were either directed by members of the Wikimedia Research Team, or in collaboration with them.

In 2012 the Wikimedia Foundation launched a project of "Editor Engagement Experiments"[22], undertaken by the Growth and Core Features teams, and with the aim of providing a better infrastructure. Some of the specific metrics[23, 24, 25] and definitions employed by this particular project became standards of research for Wikipedia.

Firstly, I review the studies dedicated to characterise the participation of the entire community and explain its current situation; secondly, I pay attention to the studies proposing experiments to foster the retention of editors.

3.5.1 Participation and the State of the Community

Wikipedia has become a kind of “living laboratory” ideal for empirical research (Schroeder & Taylor, 2015). Understanding the community composition has been a general concern for scholars ever since Wikipedia started having a considerable mass of editors. Studies that statistically quantified how contributions are spread among editors started to appear. In the first study with this aim, Voss (2005) quantified and examined the distribution of distinct authors per article in the German Wikipedia, and found out that they were following a general power law, in particular, the number of distinct articles per author followed a Lotka’s Law. These statistical distributions explained that a minority of editors created the great majority of the content.

When Wikipedia had already reached great popularity, Ortega and Baharona (2007) widely validated these results using the top-ten Wikipedia language editions. In order to calculate the level of inequality in the contributions, they used the Gini coefficient and found that more than 90% of the content can be attributed to less than 10% of the community. However, a more alarming result was that the community started decreasing

[22] https://en.wikipedia.org/wiki/Wikipedia:Editor_engagement
[23] https://meta.wikimedia.org/wiki/Research:Metrics
[24] https://meta.wikimedia.org/wiki/Category:Metrics
[25] https://www.mediawiki.org/wiki/Analytics/Metric_definitions

in number of active editors - that is, editors who edit at least five times every thirty days. Other studies measured growth, population shifts and patterns of editor activities, and found that the growth had declined (Suh et al., 2009). During those years of impasse, the slow growth of Wikipedia was explained by an increased activity on the part of the very active users and a diminished activity on the part of the middle group of editors. Later, it was demonstrated that editors who joined in 2006 were still more active than any other annual group and the editors who were leaving were the new ones (Geiger & Halfaker, 2013).

Results had shown this decline for several years, and they are well-known by scholars both outside and within the community[26, 27, 28]. The Wikimedia Foundation with the Editor Engagement Team started a project called Vital Signs Dashboard[29, 30] (2013) in order to provide tools for the measurement of on-site activity based on standardized metrics. This tool is still in development, and it is mainly intended for product and program managers. Nonetheless, a user-friendly and easily accessible version of it would guarantee awareness in the entire community.

3.5.2 Retention and New Editors' Experiments

Since the discovery of active editors' decline, researchers have started conducting work on engagement. For some researchers, this was a motivation to develop concepts, metrics and hypotheses. In fact, current research has conceptualized all possible phases for a reader to turn into an editor. Far from considering readers 'lurkers' (someone who takes advantages without giving anything in exchange), they have been considered as someone who may be interested but does not know how to actively participate in the project. Some studies aim at understanding how to transform readers into newcomers. Others studies focused on how to retain the newcomer, in order to consider it a surviving editor[31] (a first-time editor who continues editing in the project once the agreed amount of time – around 60 days - expires). Measurements like the editing frequency or the faster appeal (such as the time required to produce a certain amount of contributions) are employed to understand retention[32, 33].

Considering that the experiments on new editors' concentrate, first of all, on how to engage them in the project, and second, on how to retain them, Halfaker et al. (2013b) discussed the implications of receiving new contributions and the necessity to 'patrol' them. In his study, to help bridging the transition from reader to editor, the authors introduced a new tool called "The Article Feedback" Experiment. It was implemented as a new UI layer on the Wikipedia article interface with the tag "Improve this article". After testing different tag prominence based scenarios, Halfaker et al. (2013b) could see

[26] https://strategy.wikimedia.org/wiki/Editor_Trends_Study/Results
[27] https://www.technologyreview.com/s/520446/the-decline-of-wikipedia/
[28] http://www.nytimes.com/2015/06/21/opinion/can-wikipedia-survive.html
[29] https://analytics.wikimedia.org/dashboards/vital-signs
[30] https://metrics.wmflabs.org
[31] https://meta.wikimedia.org/wiki/Research:Surviving_new_editor
[32] https://meta.wikimedia.org/wiki/Research:Editor_retention
[33] https://strategy.wikimedia.org/wiki/Editor_Trends_Study

that many readers used the tool to give their impressions, and the possibility of editing went mostly unnoticed. The abovementioned authors concluded that although unproductive edits and comments may appear, the proportion of good new edits still benefits the development of Wikipedia.

In a similar way, Ciampaglia and Taraborelli (2015) tested an interface prototype called Moodbar which consisted in a lightweight socialization tool. It allowed new editors to send feedback about their first experience in Wikipedia. After measuring its degree of use and their effects, Ciampaglia and Taraborelli (2015) concluded that the fact of being able to express doubts and of receiving mentoring at an early stage improved retention and the likelihood of new editors turning into long-term editors. However, the study also suggests that socialization might have a cost in the efforts invested by the experienced editors.

As seen, these studies have put a lot of effort into both designing and analysing new ways of introducing changes to Wikipedia meant to produce a higher retention and participation as an effect. They work on retention in the idea that little changes oriented in the right way can turn new editors into very experienced Wikipedians. However, Panciera et al. (2009) studied Wikipedians (considered as only the very participative editors) in the first days after the moment they register to see if they were already different than other less participative editors prior to this learning processes. They found a recurrent pattern in every future high activity editor; if new Wikipedians promptly made a large number of edits, the probability of them becoming highly active editors increased by 18%. Even though the suggestive title of 'Wikipedians Are Born, Not Made' indicates a determinist view, I believe Panciera et al. (2009)'s results show that there is a certain kind of new editors who adapt to the bureaucracy and the project with ease, and who keep a strong motivation. These are very few, as results indicate that most editors struggle to survive these initial days (unfortunately 60% never made another edit 24 hours after registration). However, I consider that these findings should not bring a fatalist view, instead it should foster research and development aimed at pursuing the right design changes to enable new editors to enter and renew the community.

3.6 Actors in the Infrastructure Governance

After having reviewed the different aspects and manifestations of Wikipedia Editor Engagement, in this last section I propose a discussion to understand the decision-making mechanisms in the Wikipedia infrastructure governance. I want to describe the technology design process in order to identify the obstacles in the way of improving engagement.

Intuitively, I expect the lack of response to the problem of engagement could be explained by three different factors or stages: *1) degree of knowledge on the problem*, *2) awareness or interpretation of the problem*, *3) decision-making mechanisms and governance of technological changes*. To discuss them, I will base my exposition on the available documentation, discussions and material in the Wikimedia movement website, since one of its core values is transparency.

I first describe the actors involved in the creation and governance of Wikipedia. Second, I revise the current state of awareness of the problem of engagement and how the technological background culture can shape its interpretation. Third and finally, I expose the nature of the governance of the infrastructure by proving specific examples in the technology design process and identifying its possible effects on new editors.

3.6.1 Organization and Governance in the Technology Design Process

Wikimedia Foundation is the non-profit organization in charge for the Wikimedia projects, including Wikipedia. As of October 2016, Wikimedia Foundation has 279 employees[34] (staff and contractors), and it is organized into different teams dealing from technical aspects such as software development to administration and legal issues, among others. They are in charge of the budget administration and distribution to fulfil the movement goals. Their governance is based on a board of trustees, some of which are openly elected by the community. Jimmy Wales, as a founder, has the absolute power but only in theory, since in practice the decisions are taken by the board. He has moral authority to give his opinion on the path the movement should take. Although at the same time, anyone from the community could openly try to convince him, in a similar way to Wikipedia article discussions.

The community are the language communities of editors who contribute to Wikipedia and any other Wikimedia project. They are the ones that make the project grow and maintain it, but they hold no individual or group ownership towards it. They manage the tools, content, their discourse in order to follow the most important rules, and when they consider necessary, they create new policies and guidelines. Besides the community, there exist some Wikimedia chapters, which are organizations founded to support the Wikimedia projects in specific geographical places. They are mostly funded by the Wikimedia Foundation and organized by members of the language communities they represent. Chapters are a useful tool to work on the territory, to spread ideas and share debates with the Wikimedia Foundation. The cohesion between chapter and community depends on a scale factor; for instance, in the Catalan and the Dutch Wikipedia they can easily represent a language, and members can know each other. This is more difficult in large languages where communities are scattered and complex. In comparison to communities that are undefined entities whose members appear and disappear at certain moments depending on the topic or subject treated, chapters are more organized.

The relationship between the community and the Wikimedia Foundation can be explained by the different rights and duties, as well as by their organizational characteristics and capacities. For instance, while in some online communities the contributors are also in charge of providing the infrastructure (Morell, 2010), in Wikipedia contributors mostly focus in the definition and distribution of tasks and policies, and the foundation acts as a technological provider. In fact, this is the way it is defined in its mission statement[35]: "In collaboration with a network of chapters, the Foundation provides the essential infrastructure and an organizational framework for the support and development of

[34] https://wikimediafoundation.org/wiki/Staff_and_contractors

[35] https://wikimediafoundation.org/wiki/Mission_statement

multilingual wiki projects and other endeavours which serve this mission". In Morell (2010), this distribution of ownership, functions and roles between the two actors has been described as co-governance, because they share a mutual dependency, and their proper functioning is necessary for each other to continue growing the project.

3.6.2 Awareness and Technological Background Culture

In large organizations, problem awareness is often a first and difficult necessary step, with strong efforts in communication and coordination, in order to provide widely agreed solution. In Wikipedia, awareness must come from both the Wikimedia Foundation and the communities. The problem - synthesized as a decrease in number of active editors and the difficulty in renewing its long-committed members - has been detected in the academia as early as 2008, with important contributions in the following years. As explained in Section 3.5, research studies – some directed or co-authored by current members of the Wikimedia Foundation – explain the effects of bureaucracy, editor communication and bot messages, on the decrease in new editor retention. Likewise, metrics have been defined in order to quantify the degree of decline, and their application in tools facilitates measurements and real-time evaluation of engagement.

Hence, the Wikimedia Foundation has plenty recognition of the engagement difficulties. In fact, the 2016 Wikimedia strategy documentation[36] lists engagement as one of the very specific needs and priorities of the annual plan, which is generated within their teams and throughout communities' feedback. The total list of priorities is mainly structured as: finding a way to encourage more traffic while serving free content, improving the content and facilitating readership by adapting it to readers' needs, and nonetheless, growing the community and helping it to be more welcoming towards new editors. However, although different projects are linked directly to these priorities, the goal of engagement has been presented by the foundation ever since 2012, when entire teams were created with the aim of improving it.

At community and chapters level, there is no possibility of obtaining such a clear declaration of intentions. Chapters that organize conferences dedicated to promote the movement often include in their panels questions on how make tools user-friendly for non-tech users, on how to help non-tech contributors to participate[37]. However, I have reasons to believe that the communities' awareness on the current situation of engagement is only partial. First, there is no specific page dedicated to inform editors of the current state of their community engagement; second, not all editors follow the Wikimedia pages dedicated to strategy or to the evaluation of the current situation, neither do they consult the scientific literature on this topic, but prefer focusing on creating content instead.

By definition, Wikipedians[38] are more concerned with editing Wikipedia articles, than with making part of the community. Nonetheless, the abundant media coverage of the current situation of engagement implies that most, if not all, editors must hold some

[36] https://upload.wikimedia.org/wikipedia/commons/1/16/2016_Strategic_Approaches_Report.pdf
[37] https://wikiconference.org/wiki/Submissions
[38] https://en.wikipedia.org/wiki/Wikipedia:Wikipedians

degree of awareness about the decline in number of editors. Some editors in Wikipedia language edition community created a page dedicated to coordinate efforts to improve editor retention[39], however this was expanded only to Spanish and Ukrainian. In general, there is no consensual strategy to support the changes that could improve the situation. Being aware of having a problem is the first step, implementing the necessary actions to solve it is the next and most difficult step.

Engagement is the design goal for any technological object. Products need to first attract and later engage (Sutcliffe, 2010). With this aim in mind, designers focus on the user's needs in order to create the best tailored experience (McCarthy & Wright, 2004). Design frameworks like user-centered design precisely focus on understanding users' needs and motivations in order to make technology usable and pleasurable. These are usually structured in an iterative directed process or cycle in which there are phases dedicated to investigation, design, creation of a prototype and its evaluation.

Interestingly, Wikipedia and online communities put the emphasis on the content they collaboratively create. This does not imply the design process cannot focus on their users (namely on editors), as their tasks and activities could be facilitated by usable tools for all kind of editors. Nevertheless, the interpretation of the need for usability is dependent on the technological background culture. Wikipedia was created amidst a wave of free software and free culture initiatives, in a tech-savvy environment. It shares many of the hacker ethics principles (Levy, 1984) (sharing, openness, free access, world improvement), and many of its current editors are self-motivated editors to whom the ideology and learning skills plays an important role (Nov, 2007). The goal of creating Wikipedia is currently seen by many as a reason to self-empower, and, instead of feeling the need for a better technology usability, many editors prefer learning new ways to solve problems and continue contributing to the encyclopaedia.

In fact, in most of the projects that share a hacker ethics, such self-empowerment often goes along with a technological decentralization, since contributors do not consider the *means* as important as the *goal* to be achieved (whether if it is sharing knowledge, code or any other greater good). Contributors make an effort, and their ingenious solutions to contribute are valued by their peers. This decentralization is perhaps the only hacker ethics principle that does not apply to Wikipedia and Wikimedia projects, since editors want to put together all the available knowledge of the entire humanity in one site. Still, the decentralization tendency would explain the abundant use of external communication tools and spaces - at the expense of the project's usability. Taken together, it is clear that the values from design culture and hacker ethics are on quite opposite extremes. Nonetheless, if the current object presents some difficulties for new editors to engage in, the ultimate responsible is its design (from interface to guidelines, and from community structure to the way it channels the communication). Hence, to understand why and how the object does not present the required changes for engagement to happen, it is necessary to examine the infrastructure governance and the implementation of its design process.

[39] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Editor_Retention

3.6.3 Governance in the Technology Design Process

First of all, it is necessary to refer to the Wikimedia Foundation as the provider to the community. Both actors maintain a co-governance relationship that could obey a representation *versus* participation with their different dynamics and logics (Lovink, Tkacz, Reagle, O'Sullivan, & Liang, 2012). In some occasions tension emerged, in one case even leading to the executive director's resignation in February 2016[40]. Differences between Wikimedia Foundation and the community are also visible when it comes to their technological development capacities and planning. More precisely, the Foundation tends to support the communities' requests for minor changes[41], at the same it develops strategic plans and projects, while communities focus on the daily work and are only able to make minor technological changes[42] (e.g. configurations in the MediaWiki software or a new tool using the JavaScript language), which require community consensus. For more complex changes, editors open a Request for Comment or for task in the Phabricator tool[43], for Wikimedia Foundation staff to implement it.

When a change proposed by the community is not evident, Wikimedia Foundation staff check the community consensus. Then, they check if this change can cause a security flaw such as lowering the server performance, and finally verify it is not against any global policy. For instance, the Catalan Wikipedia community changed the logo in order to advertise specific milestones like for instance when they reached 50,000 articles. But the Wikimedia Foundation established a stricter brand policy, and further on, when the Catalan Wikipedia reached 500,000 articles, the Catalan Wikipedia community did not even discuss the possibility of creating a special logo for it. Other changes requested by the Catalan Wikipedia community were the inclusion of the feminine word for user (Cat. *usuària*) in the User Pages address, so that it could be used additionally to "user" by female editors. In some Wikipedia language edition, there is a small group of users called 'technical ambassadors'[44] who can act as a bridge between developers and editors.

When the Wikimedia Foundation proposes a software change, its implementation is discussed in each Wikipedia language edition community and decided by consensus. Regarding this issue, founder Jimmy Wales wrote in his statement of principles[45] in the foundational year 2001: "any changes to the software must be gradual and reversible. We need to make sure that any changes contribute positively to the community, as ultimately determined by the Wikimedia Foundation, in full consultation with the community consensus." This consensus decision-making follows the principle "rough consensus and running code" which is very known in FOSS (Free and Open Source Software) (Morell, 2010, p. 172). Community editors do not vote but explain their reasons and discuss until they find a solution that reasonably satisfies the majority.

As an example, Flow is a project that started in 2013 and aimed at implementing more modern discussion pages for any Wikimedia project. Its development included all design

[40] https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_Signpost/2016-03-02/News_and_notes
[41] https://en.wikipedia.org/wiki/Wikipedia:Requests_for_comment
[42] https://meta.wikimedia.org/wiki/Limits_to_configuration_changes
[43] https://phabricator.wikimedia.org
[44] https://meta.wikimedia.org/wiki/Tech/Ambassadors
[45] https://en.wikipedia.org/wiki/User:Jimbo_Wales/Statement_of_principles

phases such as prototyping and user testing in order to make pages simpler and user-friendly. The Catalan Wikipedia community participated in the project giving feedback and as soon as it was available, they first implemented it in a trial period and further on they decided by full consensus to keep it. In the English Wikipedia community, instead, Flow was rejected after the initial testing, and the reason for rejection was the need to better address particular aspects from long discussions, such as the use of templates, bots' interactions and mediators. Consequently, Flow developers had to nearly reboot the project to address these issues.

A very different case is the MediaViewer, an extension aimed at improving the display of pictures and multimedia. It was initially deactivated in the German Wikipedia, and even though the majority was against it, the Foundation forced its implementation[46].

Let us take another example. One of the most desired software updates, the VisualEditor, has not yet found a complete implementation in all of the language editions. This project has the aim of providing a WYSYWG editor (What You See Is What You Get) similar to a word processor, especially useful for those editors who do not know how to edit with the wiki-markup language code. In fact, this type of editor has been reported to solve some key usability problems since 2006 (Cowan, 2011). The project started in 2011, and today it is finally implemented in most Wikipedias[47]; nevertheless, it is not the default editing style in all of the languages in which it was implemented (e.g. at the moment of writing this thesis, the English Wikipedia is one of the languages where VisualEditor is not the default editing tool).

All in all, there exist some specific forums and annual meetings where communities make their petitions, but generally the Wikimedia Foundation decides where to put the efforts in the software development. Likewise, in a following phase, the Wikimedia Foundation provides documentation and asks the community members to get involved in giving feedback during the different stages of development considering a relationship of provider-client, but also of software partners who need to mutually help each other (Gil, 2016). Even though this has enabled the appearance of considerable software improvements and magnificent tools, no research nor panel in the Wikimedia movements events has discussed the implications of this technology design process and the infrastructure governance for the engagement of new members of the community. Taking all these aspects into account, I make two reflections in this direction:

1) Wikimedia Foundation has created teams dedicated to work on User Experience and Editor Engagement. However, it is also traceable that the academic experiments with positive results have not been implemented as interface changes, even though some showed that it was possible to improve newcomers' retention – projects like Onboarding new Wikipedians[48] and The Wikipedia Adventure[49] had not been promoted and had even been interrupted. By looking at the projects developed during the past years, it is possible

[46] https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_Signpost/2014-08-13/News_and_notes
[47] https://meta.wikimedia.org/wiki/VisualEditor/Newsletter/Wikis_with_VE
[48] https://www.mediawiki.org/wiki/Onboarding_new_Wikipedians
[49] https://en.wikipedia.org/wiki/Wikipedia:The_Wikipedia_Adventure

to see they are mainly dedicated to replace actual functionalities or provide new and more advanced tools for current editors[50, 51].

Considering that the main strategic needs of the Wikimedia Foundation are attracting more readers and improving the content's quality, I wonder: to what extent can the Wikimedia Foundation prioritise software changes and take decisions in favour of new editors, such as simplifying the bureaucracy and editing process (even though this could imply deleting old editors' proposals and tools)? Is it possible to benefit from the entrance of new editors, and at the same time, satisfy the needs of the current editors and develop tools that help the latter produce new and more high quality content?

2) Communities accept or reject the implementation of new tools or interfaces by consensus of their members. In order to tackle a change, which would hypothetically improve the engagement of new editors, it would be necessary to raise the awareness of the engagement problem in the Wikimedia Foundation and in the communities. Since the Wikimedia Foundation is already aware of the magnitude of the engagement problem and of some of the possible solutions, this implies that it exclusively depends on the communities' awareness, especially of those most active editors. I wonder: is the community aware (and if yes, up to what point?) that a bigger and more diverse community in constant renewal is needed? Is it possible to raise the current community's awareness on the engagement problem so that the current editors accept these hypothetical changes that would favour the retention of new editors?

3.7 Summary of Conclusions

Wikipedia is self-defined as a "free encyclopaedia" but it is truly a new and unique genre in the Internet that combines multiple characteristics. One of them and perhaps the most well-known, is the fact that Wikipedia is constantly created and edited by a multitude of engaged volunteer editors, with the clear goal of gathering the sum of human knowledge. Wikipedia is made possible by an infrastructure of software, hardware, and most importantly, by a set of rules and roles. Nonetheless, engagement is the most important factor required for the project to continue, which in this case is dependent on both design, content and social aspects.

By applying the model of digital engagement to Wikipedia, I provided a systematic overview of the different and abundant aspects studied by academic research. The present chapter is the first attempt to put together in a consistent way some - certainly not all - of such research which allows understanding any acknowledge aspect which influences Wikipedia Editor Engagement. It is aimed at the research community, although I believe it may be useful to any curious Wikipedian who wants to know more about the object he is helping to co-create.

Regarding the review of the aspects of engagement in Wikipedia, I want to highlight two

[50] https://www.mediawiki.org/wiki/Category:New_Editor_Engagement
[51] https://www.mediawiki.org/wiki/Category:WMF_Projects

conclusions. First, any editor who wants to achieve a fluent dialogue with Wikipedia has to learn the social, technical and rules literacy, which according to current studies is substantial. Second, Wikipedia does not provide complex algorithms aimed at giving suggestions to editors or aiding the process of editing, which made me conclude that the editors' continuity is mostly self-directed. Motivation plays a fundamental role in engagement, explaining why editors invest big amounts of time and effort at creating Wikipedia.

According to the literature, editors motivate themselves to contribute to Wikipedia because of several reasons, going from ideological affinity with the project, to social aspects such as recognition within the community and the sense of belonging. Therefore, there is reason to think that editors develop some sort of identity within the community and this may be a source of motivation. Likewise, even though the representation of personal and other identities is not encouraged, it will be interesting to see how identities influence motivation and engagement - which is the aim of Part 3 of this thesis.

In this overview, I found no studies covering aspects of the Wikipedia editor engagement such as the editors' attentional states. This is understandable, considering that the most valuable manifestation of engagement is a higher interaction - e.g. both small and long edits are important. A remarkable gap[52] in research lies in not tackling the influence of the variety of devices employed to create content. In fact, most of the studies analyse data without taking into account the devices (such as smartphones and Mobile technologies) users constantly employ in editing content.

On the contrary, there are several studies on new editors and measurements. Without newcomers, Wikipedia is at risk of dying out, because old editors may not be replaced. Studies covering the characterisation of Wikipedia engagement found out that editor participation is highly unequal and, starting from 2008, it has decreased in number of active editors. On top, there are difficulties in retaining newcomers. According to several experimental studies, newcomers could be engaged in participating for longer periods of time if they socialized with other peers. Other studies argue that specific interactions such as bot responses and editors reverts are the causes that lead to newcomers' abandoning the editing process.

In Wikipedia, decisions concerning design and strategies are determined by a co-governance between the communities and the Wikimedia Foundation. The problem of engagement ultimately depends on how the actors involved in the technological design process favour introducing the right changes to address it. In this sense, the Wikimedia Foundation is aware of the situation and aims to improve editor engagement. Instead, community awareness and editors' interpretation of the engagement problem are more difficult to infer. The editors' common background and technological culture inherent to the hacker ethics tend to assume that all users must self-empower to overcome difficulties. This is diametrically opposite to the design culture, which assumes that design should serve users' needs and any user behaviour depends on object's features such as usability. Nonetheless, I want to stress the importance of awareness as the only possible way to initiate changes and improve engagement, because in the end, any design implementation

[52] https://en.wikipedia.org/wiki/Wikipedia:Editing_on_mobile_devices

in a Wikipedia language edition depends on how it is accepted by the community (in their distinctive group decision-making process of achieving consensus).

All in all, Wikipedia is one of the most challenging objects to study engagement, especially in a moment of impasse when solutions can be valuable. Wikipedia is a valuable online community with an educative function in society, and most importantly, made by motivated and self-empowered editors. This chapter verifies the usefulness of the model proposed in Part 1, and contextualises the empirical research presented in Part 3. In the following chapters, I will study how identities can be explicative of Wikipedia editor engagement.

"Wherefore by their fruits ye shall know them" (Gospel Mathew 7:15-20)

PART 3: IDENTITIES IN WIKIPEDIA

Chapter 4. Theoretical Antecedents

4.1 Introduction

As I have shown in Part 2, motivation has been widely studied in Wikipedia. Within these studies, a special focus has been given to the variety of ways editors are encouraged to contribute and continually engage in Wikipedia, such as for instance emphasizing community values and socialization as characteristics which sustain editors (Nov, 2007). Editors' motivations are not static. I believe they could be explored as part of a more dynamic concept such as identity, and an identity-based motivation. However, little attention has been dedicated to how identities – and which type of identities – could apply to Wikipedia. The first ones to view identity as a source of motivation are Yang and Lai (2010), who considered that the self-concept is the most important motivational aspect, closely related to reputation, accomplishment and gaining autonomy.

First of all, identity appears as a solution in order to help editors to be trusted by other editors. Most of Wikipedia content is created by registered editors in the community who are recognized as trustworthy by their peers. In fact, in online communities, leaders need a consistent identity to be recognized by others (Preece & Shneiderman, 2009). Once settled in, Wikipedians are valued according to their activity, their writing skills, the languages they speak or acknowledgements they have received from other peers, such as barnstars and praising comments. Creating an identity is a way of being positively perceived by others, with trust and reputation.

However, there are some contradictions when it comes to the notion of identity in Wikipedia. For some editors, the idea of building a free knowledge encyclopaedia as a neutral common resource poses some contradictions with the idea of building a personal identity within it. Although identity proved its usefulness in building trust, there are quite a few Wikipedia editors who do not find appropriate to construct one, as they consider developing a personal authority might conflict with a reliable encyclopaedia and a common good for society (Forte & Bruckman, 2008). This unsolved discrepancy becomes more visible if we take into account that articles are not authored (as signed) by the editors who contributed to them. They happened to be created by the sum of individual contributions but they are only listed as part of Wikipedia. Halfaker et al. (2009) analysed text changes in articles and found out that editors revert or restore a previous version of an article when some words introduced by them are removed. This shows that whether identity and content are related or not, editors cannot avoid feeling some sense of identification and ownership. Thus, building an identity in the site can sometimes be seen as possibly being in contradiction with creating a common resource.

There is instead more agreement on the fact that personal and other aspects of identity should not be developed in the project, since this would damage its reliability and neutrality. This is why Wikipedia policies set clear that it cannot be considered a Social Network or a Social Media, and, as already mentioned in Section 3.2, editors are not encouraged to disclose personal information in their personal user pages, to prevent them deviate from creating quality content. But even though the creation of content should

obey the goal of achieving the sum of human knowledge, it is likely that other identities may play a role in content choices. I first had this intuition in 2012 while being involved in the Catalan Wikipedia. I ran a survey to the editor community, and found out that one of the reasons editors contributed to Wikipedia was to recreate the cultural heritage of the country, in other words, '*fer país*' (see Appendix 1). Some studies (Lieberman & Lin, 2009; Rizoiu, Xie, Caetano, & Cebrian, 2016) suggest that aspects such as gender, religion or education can be inferred from the content. Hence it can be said that identities can shape the content of the site according to personal and group values.

All in all, despite the above-mentioned controversies, I do see some reasons supporting the idea that identity can be at the basis of both motivation and content choices. Differently than with other motivation studies, some identities can be easily equated to topics, values and particular activities linked to them. In this sense, Oyserman's (2008) identity-based motivation framework based in Social Psychology can provide background to explore and reflect on Wikipedia as a context where editors' identities matter. In this third part, I pursue the **Thesis Objective 3** of investigating the influence of identity-based motivation on Wikipedia editor engagement.

In particular, I divide the experiments in two case studies. The first aims at investigating how the community identity editors may develop while being part of Wikipedia; the second focuses on the cultural identity editors acquired from being exposed to a particular context. Quite contrary to many studies which focused on the hurdles or difficulties of achieving engagement, the present work illustrates the importance of focusing on successful aspects which foster participation, in order to be able to propose mechanisms to improve participation. The rest of the chapter is organized as follows: in Section 4.2, I first review the main postulates of Oyserman's identity-based motivation framework and its appropriateness in the to study of Wikipedia; finally, in Section 4.3 I introduce the next chapters where I develop my approach and present the results.

4.2 Identity-Based Motivation

Identity is that part of us by which we are known to others (Altheide, 2000). The development and introduction of an identity in public requires, on the one side, a process of announcement made by an individual who reaffirms his identity, and on the other side, an acknowledgement by the others (Stone, Roach, & Eicher, 1993). In regard to Wikipedia editors, I distinguish between two types of identities: one which is announced and constructed in relation to the project and the community, and another which consists in social identities which dwell outside the project, as memberships to social groups.

As I will discuss in more detail in the first case study, the Wikipedia community identity is a group identity tied to project values and purposes, such as creating a knowledge resource under a free license. However, like any identity, its construction is dynamic and it evolves through time according to the actions taken. Instead, the other type of identities can be related to the meanings and values shared across other groups of people (e.g. such as being member of a church, a local association or a country). This is a notable difference between the two types of identity, and although they may not be at the same level, both could be activated in Wikipedia to trigger participation. In this sense, identity-based

motivation theory is a social psychology framework of human motivation which explains that identities can be the drive behind people acting and making specific choices (Oyserman, 2009; Oyserman & Destin, 2010).

Identity-based motivation is rooted on theories about self-concept and identity, and links them to motivation theories from a situated social cognitive perspective. In fact, situated cognition explains that cognition and action are not independent from the context but are instead dynamically created by it. Therefore, goals are situationally created and can be cued outside conscious awareness and without an organized evaluation. During the development of an action, goals can change in purpose and scope. On this basis, identity-based motivation proposes that the formation of these goals and actions obeys to the relevance of personal and social identities in a context. The model has been employed as a foundation for the study of achievement in school, consumer choices, and health behaviours among others.

For instance, in an academic context, identity-based motivation was used to demonstrate that students' identities mattered for the outcome. In one study dedicated to analysed the scenario of a math class, part of the students was of Afro-American and Latino origins. As part of the experiment, students were reminded of their ethnic group before starting the task. Results showed that for those students who identified with their ethnic group but not with a wider group of society performance declined (Oyserman, Kemmelmeier, Fryberg, Brosh, & Hart-Johnson, 2003). In other studies, similar experiments were run with students of Asian origins: when their identity was made salient before the task, their performance rose (Shih, Pittinsky, & Ambady, 1999). Taken together, the identity-based motivation explains that when an identity is activated in a context, it shapes the participants' choices, and in a wider sense, it triggers cognitive procedures associated with that identity mindset. In the previous studies, these cognitive procedures can either facilitate or undermine academic performance.

Identity-based motivation in Wikipedia. I believe identity-based motivation can be used to explain how editors engage in contributing to Wikipedia. For this, it is necessary to understand how Wikipedia characteristics set a context in which identities can become salient and trigger some actions and procedures.

Firstly, the main postulate of the identity-based motivation model is that "people are motivated to act in identity-congruent ways" (Oyserman & Destin, 2010, p. 1011). Therefore, an editor who internalizes the values of Wikipedia will feel congruent when contributing the encyclopaedia. Since this Wikipedia community identity is in the making, the more the editor internalizes the values, the more congruent the actions become. Nonetheless, this first postulate does not imply that an action cannot be congruent to two identities. Each identity set some *action-readiness* which involves taking identity-congruent actions and avoiding undesired identities (Oyserman, 2009). Then, an editor could conciliate his activity goals derived from the encyclopaedia with those derived from other identities.

In fact, Oyserman adds that "identity is a dynamic function of the pragmatic options for action in a particular situation" […] "and these options are imbued with identity-based meaning" (Oyserman, 2009, p. 255). Hence, interactions in Wikipedia could be primarily

motivated by the fact of being a member of the community, sharing its goal and its place in society, but also and most importantly, they can alo be motivated by the meaning derived from the particular content they interact with. In other words, they can be motivated by the possibility of contributing with certain contents in alignment with personal beliefs, values and interests allows editors to fulfil several aims associated to each identity. And since "identities can be subtly cued without conscious awareness" (Oyserman, 2009, p. 250), an editor might choose to perform certain tasks oriented by a Wikipedian community identity (e.g., correcting typography errors, or introducing specific data) and orient them to content related to some specific identities.

Secondly, an action driven by identity-based motivation "may not necessarily be serving individuals' goal attainment" (Oyserman, 2009, p. 255). Each identity involves readiness to interpret the world according to a particular mindset, which (Oyserman & Destin, 2010) calls *procedure-readiness*. This also remains true in the scenario of Wikipedia, where the collective effort of constructing an encyclopaedia revolves around the idea of gathering the sum of all human knowledge[53]. I may consider that the vagueness of this goal can have considerable content implications, acting as an open call for a wide range of content, which may align with all kinds of social identities, whether political, religious or related to other characteristics. Then, meanwhile an editor can contribute aiming at this goal, other identities can become salient when choosing the specific topics to write about. For instance, if a social identity involves the goal of expansion and proselytism, contributions may result in content that is not in line with the immediate objectives of the encyclopaedia and their communities – the content which would be required most to create according to other editors.

In order to prevent undesired content, Wikipedia has suitable norms and guidelines (see Section 3.3). At an article creation level, a 'Notability guideline' avoids new unnecessary or inappropriate articles by requiring a specified minimum of verifiable sources. As far as article content is concerned, the policy of 'Neutral Point of View' requires that any text must "represent fairly all the significant views published by reliable sources on a topic." Even though these norms establish some limitations in order to correct the content, their appliance always depends on other editors' intervention, and in case of dispute, solutions are taken on a consensus basis. Therefore, in some scenarios, editors' identities may play an important role in accepting new points of view in articles. For instance, the overrepresentation of certain topics in a language edition (Kittur, Chi, & Suh, 2009a) could be explained by shared identities between editors. The more common and shared an identity and its values are within the editing community, the easier it is for the content related to such identity to remain in the encyclopaedia, as editors may be unwilling to delete such content, as such a deletion is incongruent with their identity.

For Wikipedians, the coexistence of a community identity with other identities that become salient at certain moments may imply a constant negotiation between the two. The problem is that the goal of contributing with relevant and neutral content to the encyclopaedia may sometimes collide with the impulse of creating content they feel mostly aligned with. Depending on the situation and judgement of the editors, this may

[53] https://slashdot.org/story/04/07/28/1351230/wikipedia-founder-jimmy-wales-responds

result in strong bias. One solution is editors acknowledging their preferences (such as political or ideological) in the User Page. Ristau (2011) considers that the userboxes and other spaces in the user page can be the catalyst in this cognitive negotiation. As an example, Neff et al. (2013) studied the impact of community identification on political interaction in Wikipedia and observed that editors who stated their political affiliations in their User Page also intensely presented themselves as Wikipedians. Furthermore, results also showed that editors who disclosed their political affinities tend to edit more content related to the political party they support, which suggests that the conciliation between political identity and being a Wikipedian affects and permeates all the possible places of interaction - content and user pages.

Thirdly, in any context people are set into "readiness to both act and make sense of the world in terms of norms, values and behaviours relevant to the identity" (Oyserman & Destin, 2010, p. 1003). Wikipedia is not society, but it plays an important role in it. As an online encyclopaedia, Wikipedia requires providing a wide range of topics, and responding with certain immediacy to the instantanous information needs of the different societies. In fact, it is specially accessed when readers need to understand specific concepts to follow breaking news (Keegan et al., 2013). Therefore, editors with a community identity developed within Wikipedia may be sensitive to these needs and be prone to act in order to provide the necessary information in an article. Hence, when Wikipedians contribute to the encyclopaedia their actions depend on the dynamic construction of each identity, an unconscious negotiation takes place to determine which is more relevant, to whether fulfil readers expected informational needs, and contribute with content they feel most aligned with according to their own identities.

Taken all together, identity-based motivation sheds light on both the cultural and social nature of identity, providing a deeper understanding of identity-based processes. Wikipedia's wide objective of attaining the 'sum of human knowledge' allows a wide range of collective identities to become salient and manifest their outcomes in the content.

Hence, I propose using identity-based motivation as a theoretical framework for **two case studies: the first**, on the Wikipedia community identity, an in-group identity developed exclusively in this context and linked to the project's objectives, which might encourage new actions and procedures in the encyclopaedia; **the second**, on the cultural identity, referred to as "one's sense of belonging to a particular culture or ethnic group" (Lustig & Koester, 2010, p. 141). Cultural identity is a collective identity based on the context, whose values can be shared at a certain extent within the members of any Wikipedia language community.

4.3 Case Studies Roadmap

I propose two case studies in order to measure the influence of identity-based motivation on Wikipedia editor engagement, through editor participation and the specific content contributions.

- In Chapter 5, I introduce the methodology; the metrics I use in order to measure engagement, the data in which I measure it, and the statistical methods I employ.

- In Chapter 6, I present a **case study of the influence of community identity on Wikipedia editor engagement**. I initially discuss its definition, review the literature on Wikipedia editors. I propose several procedures from a community identity which can be linked to Wikipedia editors' characteristics. Finally, I present the results which examine the influence of this identity on participation.

- In Chapter 7 and 8, I present a **case study of cultural identity, its representation in Wikipedia and its influence on editor engagement**. In Chapter 7, I initially discuss the cultural identity definition and adequacy to Wikipedia content and editors. Then, I propose mapping cultural identities to Wikipedia articles. Finally, in Chapter 8, I analyse the influence of cultural identity on both the content created and the editors' process of participation in Wikipedia language editions.

Either participation or some of the characteristics from Wikipedia community identity have been studied in previous research. Instead, to my knowledge, no research has proposed studying cultural identity neither in the Wikipedia content, nor in the Wikipedia editor engagement. For this reason, in the following chapters, I will stress more importance in this latter.

Chapter 5. Methodology

In this chapter, I describe the methodology employed to study engagement. For this purpose, I first present the Wikipedia content and its characteristics (5.1). Then, I explain the data acquisition (5.2). Then I detail the engagement metrics (5.3 and 5.4) and classify them in a schema (5.5). They will enable me to compare different levels of engagement in order to understand the influence of the identities. Finally, I describe the statistical methods (5.6) used to carry on the experiments.

5.1 Wikipedia Content

Given that Wikipedia is a 15-year-old "living laboratory", on the one hand, studies can make use of the abundant editor behavioural data to validate their hypothesis with quantitative methodologies on a longitudinal base. On the other hand, Wikipedia content is structured as a network of articles, which can be analysed through their characteristics. Due to popularity, an article and its main characteristics have been documented in several occasions (Hecht, 2013; Slattery, 2009). Each article can be clearly divided between the part which is intended to be read and navigated, and the one which mediates the collaborative activity to build it (e.g. each article provides a discussion page to debate about the appropriateness of its content). I briefly focus in the visible part in order to present the elements composing it (Figure 7), which will be later used to study content representing cultural identity. In fact, it is thanks to these features which make Wikipedia easy to use and recognise by readers.

Each article has a title with a length of at maximum 255 characters. Titles are unique, although in some occasions certain words or group of words can respond to multiple objects, and therefore, need a disambiguation page. On the contrary, a redirect page sends to an article whose title is slightly different. When accessing an article through a redirect, a redirect tag appears right below the article title.

Wikipedia is very well-known for its hypertextual structure, which allows deepening into topics by navigating through close-related articles. These links directed to the same Wikipedia language edition (Intralanguage links) are spread over the entire article text. The incoming links to an article are popularly known as *Inlinks*, whereas those inscribed in the article text and directed to other articles are known as *Outlinks*. Less known are the links between language editions (Interlanguage links). These are located on the bottom left part of the page as a list of languages in which there is an equivalent article made by that language community. Even though the title may not be exact, these articles should address the same subject.

Intralanguage and interlanguage links conform two graph structures which are crucial for this research. Depending on the article lengths, the number of outlinks increases, in order to refer to those concepts which are developed within the same text but in other articles. Therefore, it is possible to convey aspects such as the article prominence by counting the number of incoming links (Hecht & Gergle, 2009) or compare the semantic relatedness

(meaning proximity) between two articles by taking into account the outcoming links. Intralanguage links graph structure has been used to analyse the concept universality, selecting those articles which exists and are available across several language editions (Hecht & Gergle, 2010b; Warncke-Wang, Uduwage, Dong, & Riedl, 2012).

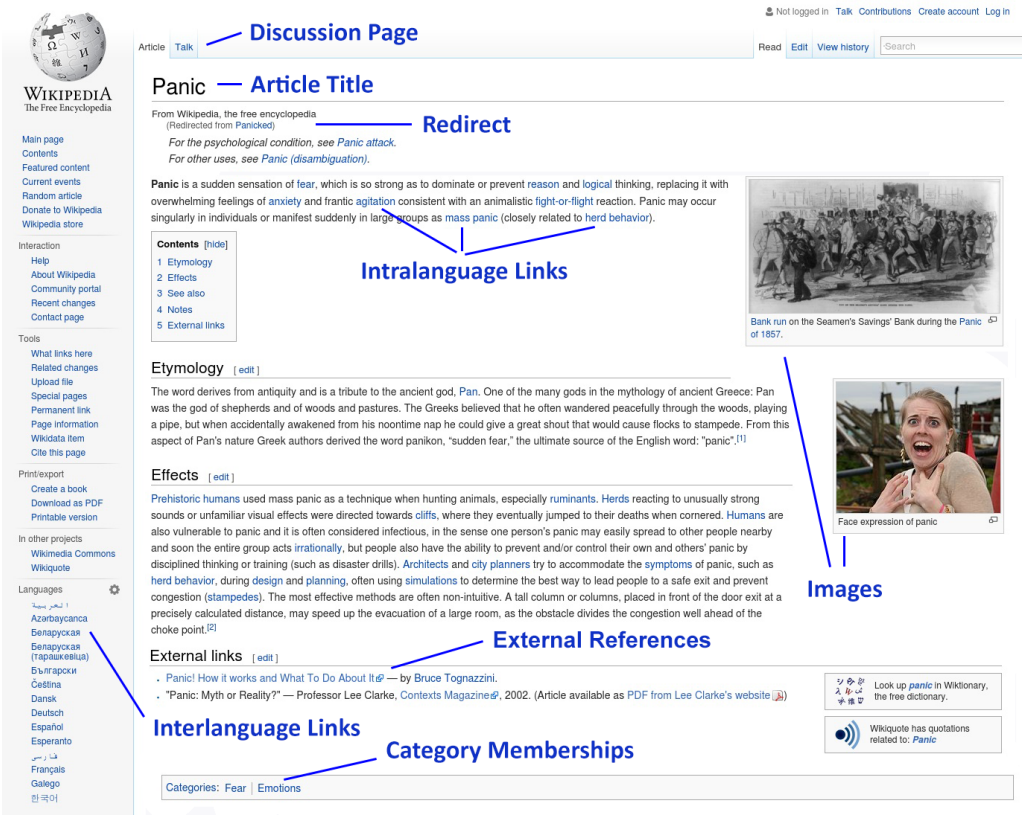


Figure 7. Article structure and its main features highlighted (Article Title, Discussion Page, Redirects, Intralanguage Links, Interlanguage Links, External References, Images and Category Memberships)

Images, external references and category memberships are other article visible and navigable features. The first illustrate the article content and are usually located in the right side, although sometimes they take a more central position. Images usually have a caption, and can be downloaded through a click on it. External references are always located in a list at the end of the article, as a requirement for any content in Wikipedia. Likewise, category memberships are located below them and classify the article in a navigable page with others into a more general topic. Sometimes, category names hold titles that may be even coincident with an article title. Category memberships conform a nested structure, in which more specific categories are located within each category, with several levels according to how each language community have designed it. Even though it could be seen as a taxonomy, these memberships sometimes contain circular references. Categories are created with no central design guidance, neither its assignation to articles.

In order to know the entire Wikipedia topical coverage, Kittur, Chi & Suh (2009a) created a method to assign an article to one or more general or macro-categories (e.g. Culture and

the arts, People and self, Geography and places, Natural and physical sciences, among others). Since this method will be later used, below I give an overview.

The method uses the category nested structure, in order to value the simplest path to find a general category and assign it to each article category membership. When the category membership is at the same distance of two macro categories, the score is equally divided. For example, if an article has three category memberships, their maximum score for the assigned macro-categories is 0.33. If one of them is assigned only to "Geography and places", while the other two is both assigned equally to "People", "Religion and beliefs systems" and "Geography and places" (since they have same number of jumps), then the final score for this article would be: "Geography and Places" 0.55, "People" 0.225 and "Religion and beliefs systems" also 0.225.

5.2 Data Acquisition

To assess the influence of cultural identity on Wikipedia content, 40 Wikipedia language editions were used. To deepen into the study of engagement both from articles and editors' perspective, 15 Wikipedia language editions were selected: Arabic, Basque, Catalan, English, German, Hebrew, Hungarian, Icelandic, Italian, Japanese, Macedonian, Romanian, Russian, Spanish and Turkish. The criteria for language selection will be specified later. These same 15 language editions were for the study of community identity. English and Catalan language editions will be preferently used for particular examples in both case studies.

Data collection and analysis are based on the Wikipedia infrastructure Mediawiki databases, provided by the project Wikimedia Labs[54] from Wikimedia Foundation. As far as the **editing history data** is concerned, entire history until January 2016 has been retrieved. As far as the **reading history data** is concerned, sample of Page views from May 2015 to January 2016 has been retrieved gathering the data from all devices (PC and Mobile). The two case studies contained in this thesis focus on Wikipedia editors' interactions and on the content produced. In order to study them, different time aggregations (see Section 2.4) are proposed. Data is retrieved from the database and it is processed considering a product or a process perspective.

5.3 Measuring the Product

Wikipedia content is the product of the entire community engagement. A product perspective aggregates data from the user or from the object. This is useful to detect and compare engagement based on the level of participation or 'higher interaction' (see Chapter 2). These are common aggregations:

[54] https://wikitech.wikimedia.org

Edit count. An edit is the mark left after inserting or deleting text from a Wikipedia page, whether it is an article or any other space. Pages can be tracked down by their revisions, which are the states between edits. An edit refers to the smallest portion of participation. It can vary substantially based on the type of work (e.g. it can be a comma or a full text), among other reasons.

Editors compare their edit count with other peers, since this is an indicator of their total participation in the project[55]. Similarly, edit count can also be aggregated for each article to compare the amount of participation involved in its creation.

Edit buckets. Aggregations like edit count provide an absolute indicator of editor participation. Following (Kittur et al., 2007; Panciera et al., 2009; Reinoso & Ortega, 2009), I employed specific buckets for the purpose of understanding the community composition. These are: 1-100, 101-1000, 1001-5000, 5001-10000, 10001+ edits. I chose these intervals and classified the editors according to the buckets.

Article Features count. Articles features can be obtained and summed through the measurement of their characteristics. For each article, I calculated the number of redirects, the article length (Bytes), the discussion page length (Bytes), the external references, the images and the different types of links. Except for the page length, these features are a mixed indicator of the characteristics of content and engagement.

Page view count. A page view (PV) is a request to the server where a website is located in order to load a single page[56]. Page view count is useful as an indicator of the appeal of a page.

5.4 Measuring the Process

A process perspective focuses on time and its measurement. Different metrics with different time-frames were considered to determine an engagement based on a 'longer duration' and 'frequent return' (see Chapter 2), which in online communities' literature is known as user retention.

Editing session and Inter-session times. The session is a short-term frame of analysis used to measure several editor actions. The most common metric, which defines the session, is the dwell time a user (editor) spends on a site. In the case of Wikipedia this information is not possible to obtain, since the only available data is regarding the time in which the editor submits the edit – whether it is a new content contribution or an existing text modification. Furthermore, editors tend to navigate through multiple pages or websites within the same session, and it might be difficult to delimit the real session duration from beginning to end.
In order to tackle this problem, Geiger and Halfaker (2013) measured the times between edits and reached the conclusion that by using a threshold to consider that the session is

[55] https://meta.wikimedia.org/wiki/Research:Metrics#Volume_of_contribution
[56] https://meta.wikimedia.org/wiki/Research:Content_consumption_metrics

over and there is no more activity, it might be possible to estimate the session duration time. This research uses the 1 hour cutoff between intra-session and inter-session edit activities calculated by Geiger and Halfaker (2013). In addition, the time between sessions were also measured (inter-session time or absence time).

Lifespan and survival/active periods. In a long-term frame of analysis, the lifespan and the survival periods provide information regarding the type of engagement an editor has with Wikipedia. The lifespan was considered the time between the first and the last edit. According to Wikimedia Foundation standards, a surviving new editor is a first-time editor who continues to edit after a period a survival period (60 days)[57]. This is used as a proxy to evaluate the newcomer survival in the project. In line with this and as an additional metric, an active or survival period has been defined as a period of sixty days in which the editor has made at least an edit. Therefore, the number of active periods were calculated for each editor in their lifespan. Six active periods are the equivalent as an active year. These are useful metrics in order to consider long-time engagement.

5.5 Engagement Metrics Schema

The following Table 2 presents all the metrics classified by focus, type of measurement, and engagement facet they respond to. The last column presents the thesis chapter in which they are used. As explained in Case Studies Roadmap (see Section 4.3), Chapter 6 is solely dedicated to see the influence of community identity on editor participation and other manifestations of engagement, while Chapter 8 explores the influence of cultural identity on both editor behaviour and content.

Table 2. Classification of metrics according to focus of measurement, time approach, facet of engagement and chapter where they are employed

Metric	Focus	Time approach	Facet of engagement	Chapter
Edit count	Editor, group, community	Product	Higher interaction	6, 8
Lifespan	Editor, group, community	Process	Longer Duration	6
Periods of time (60 days)	Editor, group, community	Process	Longer Duration	6
Session Duration	Editor, group, community	Process	Longer Duration, Frequent Return	6
Edit count	Article, group of articles, WP	Product	Higher interaction	8
Pageviews	Article, group of articles, WP	Product	Higher interaction	8
Features count (Bytes, etc.)	Article, group of articles, WP	Product	-	8

[57] https://meta.wikimedia.org/wiki/Research:Surviving_new_editor

5.6 Statistical Methods and Tests

In this thesis, some analyses require the use of statistical methods and tests in order to obtain mechanisms capable of either making assertions or taking quantitative decisions.

For most cases, a set of non-parametric tests were selected since they do not require the data to be in a normal distribution (Barry H Cohen, 2004). In fact, according to previous research (Reinoso & Ortega, 2009; Voss, 2005), several characteristics of Wikipedia including editor participation respond to a power law distribution.

The statistical tests were carried out with the data processing software SPSS. For the following tests, the significance was measured using a standard alpha cutoff of 0.05 for the p-value. I briefly comment the tests used in this research, their characteristics and purpose.

- **Gini Coefficient index** is employed to find statistical dispersion. It is usually based on a ratio between a "line of equality" and the Lorenz curve. In order to calculate it, I applied the formula proposed by (Deaton, 1997).

$$G = \frac{N+1}{N-1} - \frac{2}{N(N-1)\mu}(\Sigma_{i=1}^{n} P_i X_i)$$

  For instance, in a similar way to Ortega et al. (2008), Gini was applied to measure the degree of inequality in editor participation in Wikipedia. In the present case, μ is the mean value of edits of the entire population, P_i is the edits rank for an editor and X_i is his number of edits. Therefore, the most participative editor has the highest rank (first), and the least, N (last). This equation provides a coefficient between 0 and 1, where 1 is the maximum inequality.

- **Spearman and Pearson correlations** are employed to measure the relationship between variables. Pearson correlation provides a coefficient ρ (rho), which is either +1 or -1, depending on whether the relationship is positive or negative. The coefficient is calculated with cov as the covariance and σ as the standard deviation. The Spearman correlation coefficient is the non-parametric version of the Pearson correlation and is calculated as the Pearson correlation coefficient using the rank of the variables.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

  For instance, in this research I study the relationship between different article characteristics in order to find out whether they are providing redundant information. Since article characteristics are not a normal distribution, applying a Spearman correlation is recommended. In other cases, when data does present a normal distribution, I apply the Pearson correlation.

- **Mann-Whitney U test** is employed to compare differences between two independent groups for a not normal dependent variable. This test is often referred as the ranks version of t-test, because it uses ranks calculations in order to avoid the problems of absolute values in a non-normal distribution.

  The test requires the samples to be independent and on an ordinal scale. U is the test result, n_1 and n_2 are the sample size, and R_i is the rank value of the sample. For large samples the statistic Z is calculated as normally distributed.

$$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum_{i=n_1+1}^{n_2} R_i \qquad z = \frac{U - m_U}{\sigma_U}$$

  The test provides a mean rank for each of the independent groups, along with the U statistic which allows obtaining the significance p-value and discard the null hypothesis.

  For instance, in this thesis I compare different editor types according to their behaviour. I employ the test to see if editor differences are significant to certain events or activities.

- **Kruskal-Wallis test** is employed to analyse differences among groups of editors and articles. This test is often referred to as the ranks version of ANOVA. Since it is a non-parametric test, it is used when samples do not follow a normal distribution, hence it is an extension of the Mann-Whitney U test.

  The test statistic is obtained by:

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}$$

  Where n_i is the number of observations in a group i and r_{ij} is the rank (among all observations) of observation j from group i, $N$is the total number of observations across all groups, $\bar{r}_i$ is the mean rank of all observations in group i and $\bar{r}$ is the mean of all the r_{ij}. In order to obtain the test significance (p-value), H is contrasted to Chi-square distribution in order to discard the null hypothesis. The mean rank is useful to compare the differences between groups.

  For instance, in this research I compare the characteristics from different article topics, in order to see if their differences are significant.

  - **Dunn's Pairwise comparison** is employed to discover which groups are different which other groups. The Dunn's (1964) procedure with a Bonferroni adjustment and the Mann-Whitney U tests are usual choices to compare each pair of groups and assess the level of significance between them.

- **Sign test** is employed to determine whether there is a median difference between paired observations. This test is often referred as the paired-samples t-test for non-normal distribution.

  The paired samples can be either a measurement in two scenarios, or the measurement in the same scenario after a period of time. The sign-test determines if one of the two measurements tends to be greater than the other, by measuring the differences and ties, and uses a binomial test in order to evaluate its level of significance.

  For instance, in this research I compare an editor characteristic seven days after having registered to Wikipedia to the value obtained in the current moment. The test tells whether the characteristic has increased or decreased and its significance.

- **Simple linear regression model** is employed to model the relationship between two or more variables. The simple linear regression attempts to explain the relationship between x and Y using a straight line. The regression model is called simple because there is only one independent variable (Y). B_1 is the regression coefficient called slope, which can be seen as the change in the mean value of Y for the dependent variable x. B_0 is the regression coefficient called intercept.

$$E(Y) = \beta_0 + B_1 x$$

  The regression line is estimated by calculating the coefficients using least squares (Hahn & Shapiro, 1994), and the t-test is used in order to test the coefficients against the null hypothesis and find the significance level for the model.

  The linear regression makes several assumptions (linear relationship, multivariate normality, little multicollinearity, no-autocorrelation, homoscedasticity).

  For instance, in this thesis I model the relationship between article characteristics in order to predict the number of links they should have. The regression model provides a simple and effective method to estimate results and evaluate an article characteristic as a valuable indicator.

Chapter 6. Community Identity and Engagement

In this chapter I present the case study of community identity and Wikipedia editor engagement. In order to review the characteristics that compose a community identity in Wikipedia, I look into the theoretical antecedents of a community Identity in online communities as well as into the studies dedicated to Wikipedians (6.1). Then, I present the approach to operationalize the community identity into features linked to different editor characteristics and activities (6.2). Finally, I present the results which determine the relationship between each community identity feature and participation, and discuss their implications (6.3).

6.1 What is the Community Identity?

An online community is defined as a group with shared objectives or topics, whose members use the Internet as a primary means of communication (Porter, 2006; Preece, 2000; Yuqing Ren et al., 2007). There is a wide variety of online communities, with different social and economic goals, supporting their members and sometimes the people outside the community. In this sense, the main purpose of Wikipedia as an online community is to provide free encyclopaedic knowledge. However, the way an editor interacts with the very specific rules, guidelines, roles, and functioning (previously explained in Section 3.3) is very indicative of the degree of internalisation of the community values and of his relationship with the project.

Ren et al. (2007) studied group theories from Social Psychology and their application to online communities in order to see what makes online communities successful in terms of engagement. In this sense, they concluded that engaged members either develop and grow their motivation by internalizing the community values (which they call 'common identity') or by strengthening their bonds with other members. Later, Ren et al. (2012) studied both common identity and bond attachment in MovieLense, an online community where members contribute with movie reviews. Their results showed that common identity and member bonding both increased, and were associated to editors with a higher engagement (both the visiting retention and user participation). However, common identity was more influential than member bonding.

MovieLens common identity was associated to community features users did not necessarily interact with, during their first visits to the site, such as certain pages dedicated to the group news, its activities, or the way of self-defining as a group member. The common identity is an in-group identity that evolves throughout time as a consequence of experiences and decisions. According to the identity-based motivation framework presented earlier, when an identity is relevant to a context, it implies an action and procedural readiness. In other words, not only would MovieLens users feel more prone to act, but they would do so in accordance with certain activities related to the group values. The common identity represents a complete mind-set of the community values.

However, Oyserman's identity-based motivation would not be able explain the effect of other contextual factors like the bonds between members of a same community. It is

possible that they provide a context where actions are more congruent. For Ren et al. (2012), member bonding was complementary to common identity, and its importance on the user could vary along time. While in all online communities some member bonding might manifest, the common identity features vary according to the site characteristics. The same common identity developed in MovieLens would not necessarily appear as salient in another online community such as Wikipedia. Each identity cues an in-group mind-set and readiness to make sense of the environment in very specific ways.

For example, Danescu-Niculescu-Mizil et al. (2013) studied the characteristics of the users of two online communities: RateBeer and BeerAdvocate. In specific, they analysed the use of language in order to detect changes which could indicate the emergence of a collective identity. After studying different user's lifecycle in the communities, they detected that users who become most receptive to their norms, they also become in-synch with the language terms and topics. Later, they cease to respond to the new changes in language and eventually leave the site. The study suggests that by identifying the linguistic changes it is possible to estimate the lifespan of users and their potential to become very active members of the community.

The interesting point of the study is that the linguistic terms employed by users change in accordance to the community values. By internalising the different popular terms, the user may develop a community identity which is congruent with that specific community and motivate further contributions. However, RateBeer and BeerAdvocate showed that their dynamic nature in terms of language could be a drawback to maintain its users engaged. Therefore, the process of building a community identity is dynamic: it may also depend on the available actions and topics on the community – some may provide changes and challenges at greater pace and allow their users to identify with tasks of greater specialisation. Yet, it is expected that any community identity is based on the online community main goal. To study the influence of a community identity on engagement, it is necessary to detect how this main purposeor goal, the values, the user characteristics and the activities are linked together.

Wikipedia community identity values, editor characteristics and activities. In order to understand the influence of a community identity in Wikipedia, I need to identify the activities and editor characteristics which may appear after having internalized the Wikipedia values. Some of the characteristics of experienced editors and community members have been long-time studied by means of qualitative approaches, like ethnography and interviews (Bryant, Forte, & Bruckman, 2005; Forte & Bruckman, 2008) or by means of quantitative approaches (Arazy, Ortega, Nov, Yeo, & Balila, 2015; Hale, 2014; Welser et al., 2011). The results of these studies are useful to evaluate which editors' characteristics and activities are more likely to be part of the Wikipedia community identity.

*Self-presentation through User Page and participation***.** Even though it is possible to contribute to Wikipedia anonymously, it is highly recommended to register. Other peers need to recognize editor's skills and degree of expertise in specific fields in order to collaborate, or even evaluate and track contributions (Bryant et al., 2005). An editor who has registering a name is more trusted than anonymous editor, and it is recommended to use it throughout posts or revisions ("Sign your posts on all talk pages"; "Log in before

making drastic changes to existing articles"). Registering a name may be enough in order to belong to the community. In Wikipedia, user pages should be primarily used to present information relevant to one's work in the encyclopaedia[58], and only limited biographical information. This is in contrast to other online communities and social media, where User Pages are used for disclosure and can be related to member bonding (Yuqing Ren et al., 2007). In Wikipedia user pages, editors show the topics they are interested in, the tools they use, and the articles they have made. Sometimes editors claim their efforts on particular articles and list them in their User Page. In short, from this perspective, editors create self-appointed public characters in order to communicate their value and gain credibility and reputation.

However, according to another perspective, very active Wikipedians evaluate other editors trustworthiness based on the endurance of their edits (Krupa, Vercouter, Hübner, & Herzig, 2009). Equating identity with work brings Wikipedia close to a meritocracy. Even though newcomers are not aware of the different types of work or roles editors play in the community, experienced editors view their participation on the project as their membership in the community (Bryant et al., 2005). This is sometimes depicted as a continuum between a periphery and a centre or core, in which editors tend to increase their participation and complexity of tasks. Although an edit might imply very different types of work, there is a sort of admiration towards the editors' total number of edits ('editcountitis'), which has been documented even in Wikipedia help pages[59]. Some studies consider that only the editors who achieved a large number of edits are worthy of being called Wikipedians (Panciera et al., 2009). Hence, participation in terms of number of edits is seen as a public aspect of an editor. Therefore, both the User Page and participation could serve the function of self-presentation in the community.

Assuming rights and taking a functional role in the community. Regarding the periphery-centre analogy, Bryant et al. (2009) also noticed that editors from the community core tend to move from local focus on specific articles to a more holistic view of the project. A general aspect of online communities is that a small fraction of users participate in the governance and create the rules and policies, repair vandalized content and tutor newcomers (Preece & Shneiderman, 2009). In Wikipedia, some specific editor 'flags' are created in order to grant rights and allow trustable editors to take the above-mentioned tasks.

The process by which these editors receive the flag is either through a request or proposed by another editor; and participation is one of the determinant characteristics to obtain one (Burke & Kraut, 2008). Even though different flags exist, some encompass more rights than others. For instance, the administrator flag allows the most important governance actions in Wikipedia, such as protect and delete pages, as well as delete user accounts. Editors with a flag have a functional role in the community: the rest of editors expect them to fulfil certain actions they have been entrusted with. Functional roles also present a continuum editors tend to climb to achieve a more central position; for most of the communities, oversight and checkuser are the highest level followed by administrators (sysops and bureaucrats), while registered editor is the lowest rank (Arazy et al., 2015).

---

[58] https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_is_not_a_social_networking_site

[59] https://en.wikipedia.org/wiki/Wikipedia:Editcountitis

Obtaining a functional role is the confirmation of having internalized the project's values, and is tantamount to having acquired a specific position in the community core.

Participating in community and social oriented activities. Some studies have focused on identifying other types of roles, defined by the recurrence of their activities (Welser et al., 2011; D. Yang, Halfaker, Kraut, & Hovy, 2016). In order to assess the complexity of wiki-work, these studies took into account the different namespaces existing in Wikipedia, which are dedicated to articles, discussions, user conversations, community infrastructures and governance. Some of the roles they found were social networker, vandal fighter or fact updater. Editors who take up these more flexible roles have already assumed a position in the community, and therefore, see their contributions in consonance with other editors' work. Some of the activities carried out by the editors who play these roles are oriented to the whole community, such as creating templates, uploading files or creating rules. However, social activities and personal communication could be part of the community mind-set or not, depending on whether the focus is on newcomers or on bonding with other editors. Hence, it is not easy to distinguish whether a community identity should limit those activities which do not clearly embrace the project's values or goals, but which are, nonetheless, equally beneficial to the community members.

Contributing to multiple language editions as an attachment to the project as a whole. In an online community or more generally in a group, having internalized the group values tends to transition into an affect towards the entire group and goals their members are pursuing (Yuqing Ren et al., 2007). In Wikipedia, a considerable part of editors contributes to multiple languages editions to make a bigger contribution to the global project. Even though editors choose their language edition according to their competences, other sociological factors matter; some editors prefer the English language edition, considering their own language has a low status, while others think in terms of impact on a large audience of readers. Contributing to multiple language editions should be seen as the desire to help the project grow as a whole. A 2011 survey found that half of respondents contribute to other language editions, and an overwhelming majority (72%) read Wikipedia in more than a single language (Glott, Schmidt, & Ghosh, 2010). However, Hale (2014) studied the edits produced during two months and saw the percentage of multilingual contributors is only a 15% of active editors. And when non-English editors decide for a second language edition, in most cases it is English.

All in all, a community identity in Wikipedia might take shape in the different sorts of activities and editor characteristics I described with the available current research. *I expect participation to appear both as a central value in community identity, and as a consequence of it. The internalization of the different community identity values can make each further interaction more identity-congruent, which in turn would raise the participation and create a continual motivational reinforcement.*

6.2 Operationalizing the Community Identity

In this section, I describe the operationalization of the community identity in Wikipedia, in order to link group values to particular features based on editor characteristics and activities. Hence, I first evaluate the different lengths of user pages, and then I present the

different editor types and functional roles. Finally, I define language affiliation and the main community-oriented activities.

6.2.1 User Pages

Editors' self-representation in a User Page is a step towards the community. Hence, each editor's User Page length has been measured and classified according to different lengths chosen arbitrarily. In number of Bytes: no User Page, one line (<= 600 Bytes), page (600 > and <= 1200 Bytes), profile (1200 <= 4000 Bytes) and complete profile (>4000 Bytes). Figure 8 presents five different real User Page examples from Catalan Wikipedia. Besides the User Page with one line, they all provide minimal description of themselves. Some of them (b, c, d) contain some user labels in order to specify the languages they speak along with some personal affiliations. The longest profile (Kippelboy) contains a wiki-biography explaining his career in the project, both in the site and as chapter organizer. The User page is a community identity feature, which editors create to find new collaborations or to build a self-image in order to continue contributing.

6.2.2 Editor Types

The three main editor types are: anonymous, registered editors and bots. Nonetheless, only the registered ones can be considered the real community members. In fact, there are no available studies explaining the anonymous editors' profiles; I suspect that anonymous editors are mostly spontaneous editors who do not want to register as they prefer making little contributions, rather than full registered editors who do not want to login. Anonymous editors are tracked by their contributions, as they are signed in with their current IP (which is reportedly used to obtain data such as location information). Bots are algorithms operated by registered editors, and their contributions are signed with the bot registered name, like any other registered editor.

In order to build and preserve the encyclopaedia from vandalism attacks and in order to manage the community, Wikimedia Foundation designed some flags to be granted to registered editors[60]. These flags or rights are attributed through community consensus. Editors receiving them need to ensure trust, and on the basis of their achievements and activities they can transition to more power in their career in the future (Arazy et al., 2015). These roles are often called functional, since they allow making special actions and they serve specific purposes. Current literature locates functional roles in the core of the community, subdividing them into levels from 0 to 5 (Arazy et al., 2015). These are defined by their purposes and degree of responsibility in the community. As far as the community identity is concerned, taking a functional role implies switching from an individual mind-set to a collective point of view.

In a similar way to Arazy et al. (2015), the different functional roles were classified into groups - levels (from 0 anonymous to 4 security force), configuring the same progression

[60] https://en.wikipedia.org/wiki/Wikipedia:User_access_levels

towards the core of the community (see Table 3). The middle and high level group were simplified and merged as these two flags served a similar purpose. These changes respond to the need to find equivalence to the different variations found across language editions, and be able to compare results.

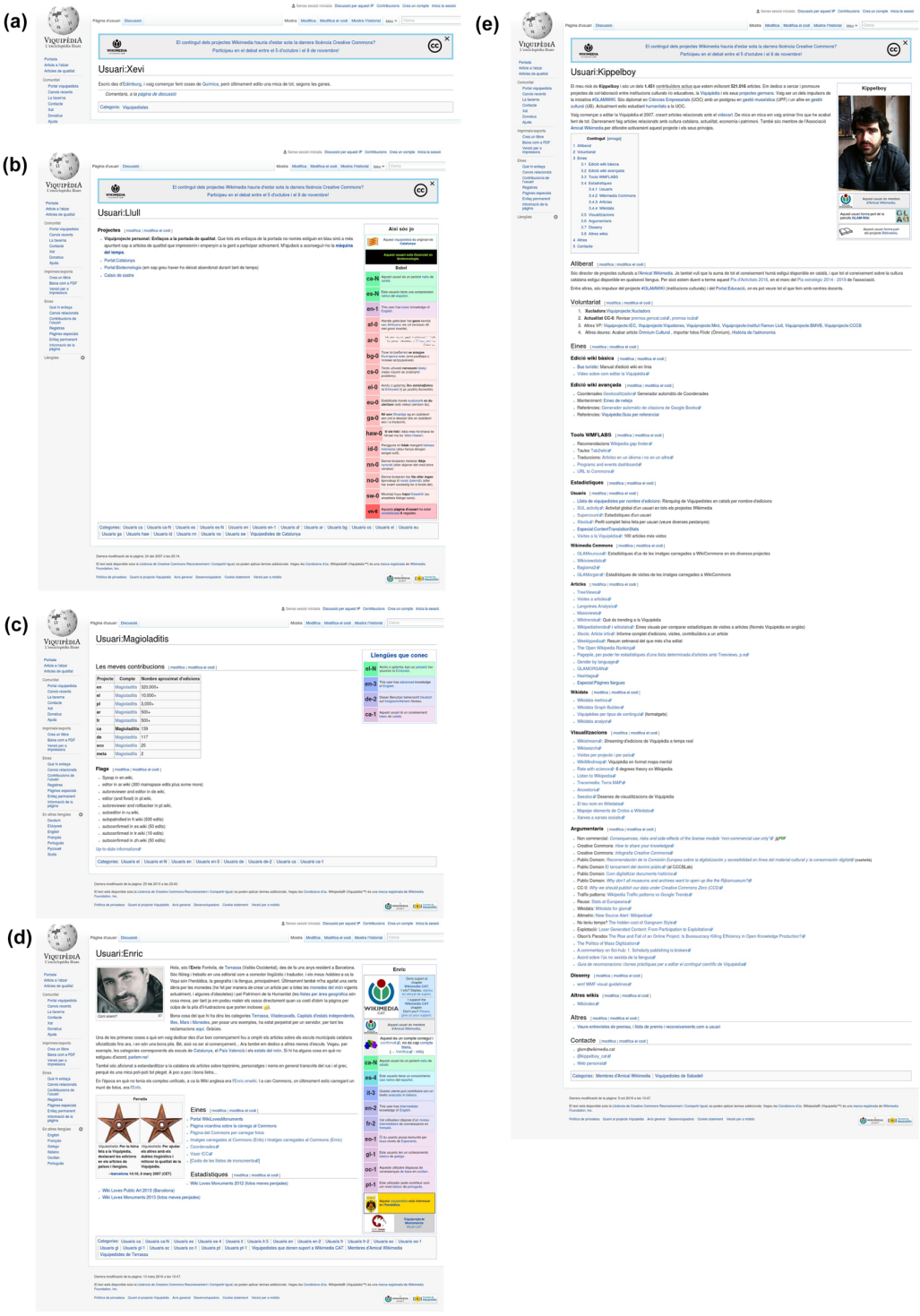


Figure 8. Examples of User Page. *(a) Xevi (one line), (b) Llull (page), (c) Magioladitis (short profile), (d) Enric (profile), (e) Kippelboy (complete profile)*

To put an example, in the Hungarian Wikipedia there are 690 editors with a flag named trusted, while there are no autopatrolled editors. Instead, in the Russian Wikipedia there are 1581 autoeditors and no autopatrolled or trusted editors. All of these are classified together under the group label 'Production Force', since their goal is to create new content in an agile way. Leaving aside Security Force, where, as already mentioned, there are very few editors, the group of administrators is the one encompassing more rights (they usually have the flag 'sysop', and very seldom 'bureaucrat'). Administrators have a wide variety of actions such as blocking editors, protecting and renaming pages without restriction.

The fifth level from Arazy et al. (2015) which includes the flag steward has not been considered as this is a flag created and managed by the Wikimedia foundation and only grants technical privileges. Other technical flags such as accountcreator, filemover or template editor have not been included for the same reasons: they respond to technical purposes, they encompass few editors and not every language edition has them.
Bots have been excluded from the analysis because these are non-human editors, whose behaviour is automatized by other editors, and therefore it is not useful to evaluate them in terms of identity-congruent actions. In order to do this, two heuristics suggested by WMF were followed[61]. a) In principle and most generally, bots are registered with a 'bot flag' (in a similar way to functional roles). Yet, there are exceptions: in some other cases, bots do not have this flag and still operate in various languages. In addition, b) when user names contain or end with the word 'bot' in its different upper-lowercase possibilities, they can be included as a probable bot – whether they have the bot flag or not. Therefore, bots were included using the two heuristics. However, since there could be user names containing the string of characters ‘bot’ for other reasons, the list of collected bots was verified against the list of editors with a functional role. This way it was possible to avoid including by mistake an administrator as a bot[62].

Table 3. Editor types classified with level, description and access flags.

Level	Editor Type	Description	Access Flags
Level 0	Anonymous	Non-community members	-
Level 1	Registered	Registered editors	user
Level 2	Production Force	Editors who are granted some trust and therefore their contributions are not monitored	autopatrolled, editor, autoeditor, autoreview, trusted
	Quality Patrol	Editors who are involved in patrolling new content and reverting problematic changes such as vandalism.	reviewer, rollbacker, patroller, abusefilter
Level 3	Administrators	Editors who can perform block actions on both content and community	sysop and bureaucrat
Level 4	Security Force	Editors who work to keep the community healthy from malicious editors.	check-user, oversight and steward

[61] https://www.mediawiki.org/wiki/Analytics/Metric_definitions#Active_editor
[62] This was introduced after erronously identifying as a bot the Catalan Wikipedia administrator named “Paucabot”, who by the way is as active as many bots.

6.2.3 Language Affiliation and Multilingualism

As far as the community identity values are concerned, it was argued that contributing to multiple language editions can be a sign of internalizing Wikipedia values and developing affect towards the project as a whole. Before 2013, it was necessary to register in each language edition. At the present time editors are able to contribute to any other language with the same name under a unified name account[63]. In order to analyse multilingual editors, it is necessary to distinguish their relationship towards each language edition.

In line with previous research (Hale, 2014; Kim et al., 2016), I identified an editor as primary to a language edition when the majority of her contributions and interactions were made in that language edition. Complementarily, the same editor is non-primary to the rest of other language editions to which he or she contributes to. Among the editors who are primary to a language, I identified the non-multilingual editors. By applying these exclusive definitions, a list of non-primary editors, primary multilingual and primary non-multilingual editors were identified for each Wikipedia language edition.

6.2.4 Activities and Namespaces

Any interaction in Wikipedia, either a text insertion or a deletion, is stored as an edit. Contributions to articles are edits to the article's namespace number 0, while content discussions, policy writing or personal messages are edits to other namespaces[64]. Most of the editors' work is devoted to articles. However, depending on the tasks they negotiate and self-attribute, editors also embark in other types of activities or roles (Welser et al., 2011; D. Yang et al., 2016).

Three main types of activities were identified with their aims, places and namespaces (Table 4). Besides the production activity, which concerns edits in articles, the other activities are intended to reach either the community or specific members of it. More particularly, Community Communication included all the edits meant to be read only by other editors and with a community communication function. These may range from policy creation (object governance) to article text discussion (discourse). Data Spaces is another joint classification encompassing those contributions whose content can be reused by editors to create articles - from files, images to categories.

Hence, Community Communication and Data Spaces are two activities completely community-oriented, aimed at increasing its resources and ensuring its well-functioning. Editors who have incorporated such activities in their behaviour internalised the project's values and decided to collaboratively work with the other editors. Personal Communication is another activity that could aim at both community members bonding and leisure. Even though it does not reflect a community mind-set as the others do, it is used as a control feature.

---

[63] https://en.wikipedia.org/wiki/Wikipedia:Unified_login
[64] https://en.wikipedia.org/wiki/Wikipedia:Namespace

Table 4. Wikipedia main activities classified by community function, aim, namespace and namespace number

Activity	Place	Aim	Namespace	Namespace Number
Production	Edits in Articles	Providing final text for reading	Page article	0
Data Spaces Contributions	Edits in Data Spaces	Creating material for the articles and organizing the articles content	Files, categories, portals and templates	6, 7, 14, 100, 10
Community Communication	Edits in Article Discussions	Discussing text	Page talk	1
Community Communication	Edits in Wikipedia, Guidelines, Help	Creating and discussing about the Wikipedia governance	About Wikipedia, guidelines and help	4, 5, 8, 9, 11, 12, 13, 15
Personal Communication	Edits in User Pages	Expressing personal preferences. Giving barnstars, Notifying other users about work, etc.	User, user talks	2, 3

6.3 Community Identity and Wikipedia Editor Engagement

In this section, I present the results that estimate how the construction of a community identity influences Wikipedia editor engagement. I undertake a broad characterization of the communities of fifteen Wikipedia language editions in order to analyse the relationship between editor participation and community identity development.

6.3.1 Research Questions

The level of participation is the first defining trait of the Wikipedia community identity. Editors consider their own participation as equivalent to community membership (Forte & Bruckman, 2008), and evaluate other editors' reputation according to this same principle. At the same time, participation can be the consequence of internalising the values and acquiring the characteristics and activities common to the community identity.

The User Page is a space where editors can present their public character. To date, only one study assessed the percentage of registered users having a User Page (54% in Chinese Wikipedia) (X. M. Zhang & Zhu, 2011). Although editors employ the User Page to present their skills and convey credibility, this is often deduced from the participation itself. In this light, developing a User Page can be is tantamount to a step towards the community. This leads to the following question:

RQ1. *Which is the relationship between self-representation through a personal User Page and participation level?*

In Wikipedia, like in most online communities, participation is unequally spread among few engaged editors (Ortega et al., 2008). Some of these very active editors are administrators (Kittur, Pendleton, & Kraut, 2009b). Even though a certain level of participation is a requirement, the guide to requesting adminship advocates that factors such as trust and confidence are determining. Other functional roles are elected on similar criteria. The second question is:

RQ2. *Which is the relationship between having a functional role and participation level?*

Generally in online communities, members who assume the community identity values end up developing an attachment to the project as a whole (Ren et al., 2012). One way to apply this principle to Wikipedia could be seen in the development of multilingual participation. Hale (2014) measured edits in a time-frame of two months and found that the percentage of multilingual editors oscillated around 15%. However, not much is known about what types of editors would be more prone to extend their activity to multiple languages, neither about the possible consequences on their present participation in their primary language. This leads us to the third research question:

RQ3. *Which is the relationship between multilingualism and participation level?*

Some studies found that certain types of editor roles or profiles can be distinguished based on their activities in Wikipedia spaces (Welser et al., 2011; Yang et al., 2016). For

instance, certain editors are mainly oriented to editing in social network spaces, which implies they develop an attachment to other members. Others give priority to some work in favour of the entire community. In Wikipedia, the community identity may imply adopting some of these community-oriented activities. This leads to:

RQ4. *Which is the relationship between engaging in community-oriented activities and participation level?*

Previous research showed that the editors who spent the highest amount of hours in editing sessions were also the ones who performed the most edits (Geiger & Halfaker, 2013). Obviously, these editors reached high levels of participation with a long-term dedication and editing regularity on a daily basis. As already explained in Chapter 3, Section 3.5, editor retention refers to the continued activity of editors, and is a key aspect for any online community project (Yuqing Ren et al., 2007). While a strong participation implies retention, the opposite is not necessarily true. The last question is:

RQ5. *Which is the relationship between editor retention and participation level?*

6.3.2 User Pages (RQ1)

Editor participation level is considered an indicator of credibility in front of the other peers in the community. Sometimes, editors create a User Page for this very purpose. I propose examining how editors with different levels of participation differ in the development of their User Page, in order to assess the importance of this characteristic for a community identity.

Results. Figure 9 represents the percentage of editors by User Page length according to the edit bucket they are assigned to. By looking at the different edit buckets, several observations can be made. The vast majority of editors who are in the first 100 edits do not have a User Page at all (they range from 77.71% to 90.50%), and this group is even bigger if we consider those who have a User Page with only one line. In this edit bucket, the editors who have invested more in developing their User Page are less than the 3%. This implies that building a User Page is not an initial activity performed with the aim of introducing oneself to the community. The majority of editors between 101 and 1000 edits do have a User Page, but consisting in a single line. After 1000 edits and before 5000, editors who write more than a line in their User Page become a majority. From 5001 to 10,000 edits and beyond 10,001 edits, the trend tends to invert: editors having a complete profile are the most numerous, followed by groups with a corresponding shorter User Page.

To answer the first research question **(RQ1)**, there exists a relationship between participating more and building a larger User Page. This pattern is shared and it is visible in the 15 Wikipedia language editions analysed and depicted in Figure 9. Nonetheless, exceptions appear in small communities like the Macedonian and the Icelandic, where participating more does not always equate to building more complete profiles (see fourth edit bucket). Editors from languages like Hungarian or Hebrew have the largest share of longer user pages, while editors from Turkish seem to have developed them the least. However, cultural variations in the creation of a User Page are minimal.

Discussion. The results have shown that participative editors tend to develop longer user pages. User Page is therefore a community identity characteristic whose development may be positive for the editor. In fact, this relationship manifests in all 15 Wikipedia language editions with very slight variations.

Even though there seems to be a relationship between participation and developing a longer user page, not all participative users develop this personal space. Hence it is not as essential as it would seem. For instance, a majority of editors who exceeded 1000 edits but not yet 5000, (which is still a considerable amount of work), feel no special need or desire to present themselves through a User Page containing more than a line. The majority of new editors start editing without paying special attention to developing a User Page. On the other hand, only between 33% and 50% of the extremely active editors (more than 10,001 edits) build complete profiles.

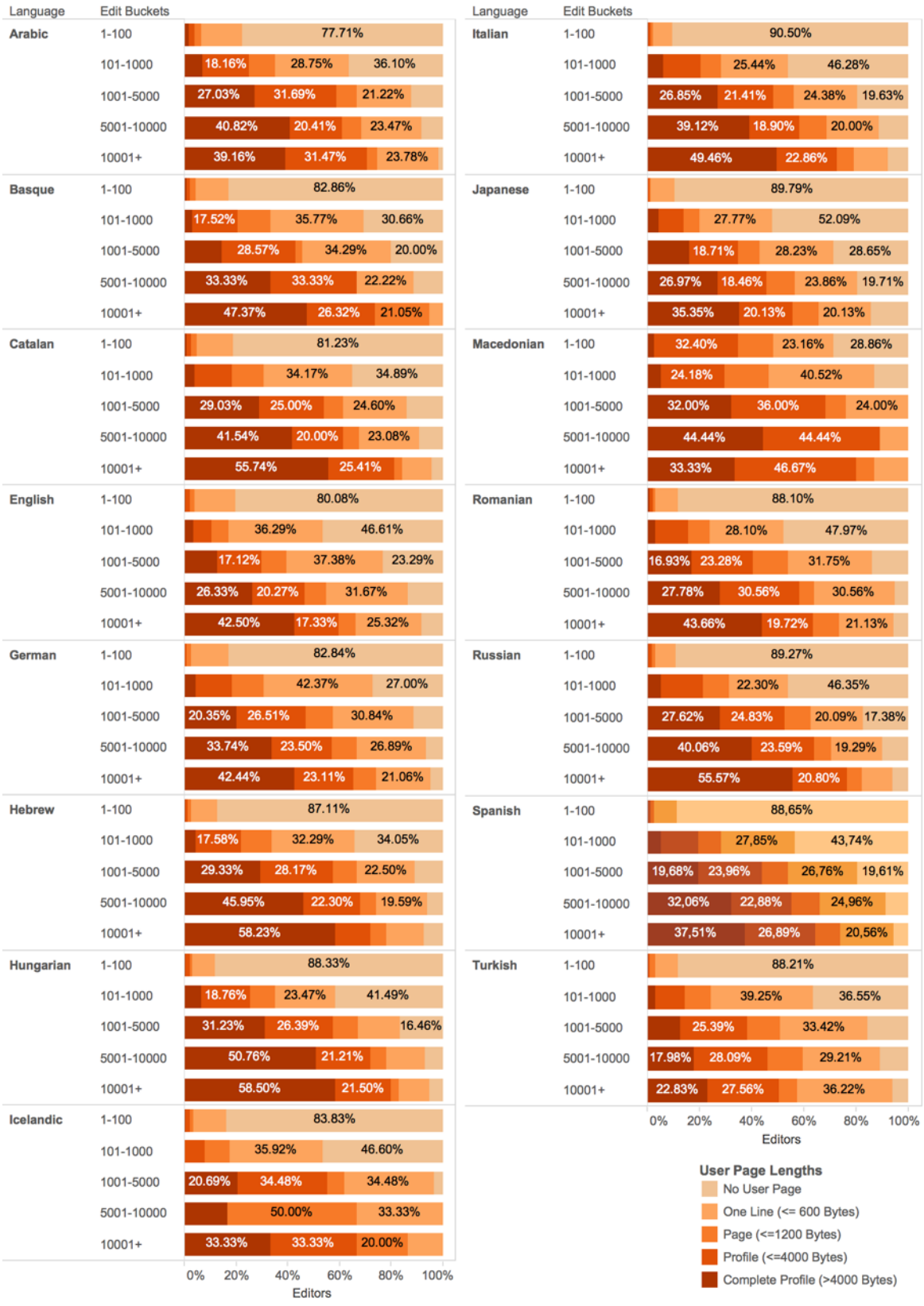


Figure 9. Proportion of editors by User Page length within each edit bucket.

Generally, these results confirm the idea that editors present themselves through their work and not so much through their User Pages. According to the Wikipedia 'User Page guideline'[65], it should display information relevant to work on the encyclopaedia. Even though editors may differ in their participation, they could all show their topics of interest, skills and achievements. This applies especially to new editors, who could be demanded personal information in view of a smoother integration into the community dynamics. A simple mechanism to motivate editors to share relevant personal aspects could be useful to their peers and help establishing collaborations and start building a community identity.

6.3.3 Editor Types and Participation (RQ2)

Obtaining a functional role shapes the editor mind-set towards serving specific functions in the community. Functional roles are limited by definition, and one of the requirements for its obtainment is a certain level of participation. In order to shed light on the relationship between the two, I propose various calculations regarding community participation (edits made by bots were not taken into account).

a) Community core: participation and functional roles

Results. Table 5 shows the percentage of edits made by the top 100-500-1000 editors ranked by the number of edits in each Wikipedia, the percentage of edits made by functional roles and the Gini coefficient of the edits made by the registered editors (Reg.) and by the editors with a functional role (F. Roles).

Table 5. Participation inequality in Wikipedia language editions and coincidence with functional roles. *The table shows the percentage of edits made by the top 100, the top 500, and the top 1000 editors with more edits. It also shows the percentage of edits made by editors with a functional role (F. Roles) and by admins (Admin.). The last two columns show the Gini coefficient of the edits made by the registered editors (Reg.) and by the editors with a functional role (F. Roles).*

Language	% Edits by					Gini Coefficient	
	Top 100	Top 500	Top 1000	F. Roles	Admin.	Reg.	F. Roles
Arabic	53.63	80.80	86.29	76.05	14.52	0.96	0.85
Basque	46.65	48.76	49.22	29.14	24.26	0.97	0.34
Catalan	67.58	87.37	90.95	73.36	17.80	0.97	0.66
English	8.84	19.53	26.84	47.57	20.62	0.91	0.69
German	15.55	36.11	48.61	87.46	29.22	0.97	0.85
Hebrew	48.54	79.15	87.28	66.7	15.38	0.97	0.74
Hungarian	51.47	81.77	88.43	77.47	9.16	0.97	0.74
Icelandic	44.23	48.10	48.87	50.74	35.26	0.96	0.48
Italian	30.81	58.58	70.61	58.81	12.96	0.97	0.63
Japanese	20.95	40.92	52.12	2.58	2.25	0.95	0.53
Macedonian	44.79	47.75	48.57	73.1	35.61	0.97	0.82
Romanian	31.84	40.90	43.13	53.61	16.18	0.96	0.75
Russian	23.55	49.48	62.37	69.11	5.82	0.97	0.91
Spanish	26.72	51.61	63.33	35.26	6.06	0.96	0.59
Turkish	54.01	76.28	81.99	67.21	7.85	0.96	0.85
AVG	37.94	56.47	63.24	57.87	16.86	0.96	0.69

[65] https://en.wikipedia.org/wiki/Wikipedia:User_pages

The results show a considerable participation inequality in all language editions; language variations due to their scale. In small communities like Hungarian, Catalan or Turkish, the first hundred editors in number of edits made around half of the Wikipedia editors' edits, a percentage which was only achieved by the first thousand in German Wikipedia. Regarding the functional roles, the percentage they created often accounts for a much larger share than the most participative editors selected in the top 100 ranks and smaller than the one made by the top 1000. There is a close relationship between participation and having a functional role: functional roles participation largely coincides - although not completely - with the most participative editors but (**RQ2**).

In fact, the most participative editors and those with functional roles are both considered the 'core' of the community - in opposition to the periphery, which is populated by editors with very few edits. It is usually assumed that administrators are the most relevant functional roles in this core, as they hold more rights to influence the community governance. But results show that in many communities, administrators' participation represents a small contribution in the entire group of editors with a functional role. The Gini coefficient of the edits made by the editors with a functional role show there is less inequality than in the entire community, but its high value supports the idea that the functional roles are not a homogeneous group, and that some of the roles differ in the level of participation.

b) Community composition

Results. Figure 10 represents each editor in the Catalan Community with a bubble; the colour of the bubble is related to the editor type (its functional role or being a registered editor) while the bubble size represents the total number of edits. In order to represent the core-periphery continuum, functional roles are located in agreement with their level of rights (Security Force, Administrators, Quality Force and Production Force). On the top left subfigure, editors are represented according to editor types, using the same colours but without considering the size. On the left bottom subfigure, editors are represented using the number of edits as the size and coloured according to the edit bucket they belong to. These graphs illustrate the inequality of the community. At the same time, they support the community core-periphery continuum (those at the centre hold a functional role and tend to be bigger). In this language edition, the functional role Production Force takes a vital importance: some editors have higher participation levels than administrators. As pointed in Table 3, Production Force consists of the trusted editors whose work does not need to be observed by others. Quality Patrol, a group of editors dedicated to content surveillance and fighting vandalism, is reasonably smaller than the administrators group, and smaller than the Productive Force. Security Force are top administrators who supervise other editors in order to prevent malicious activities – so they do not need to be very numerous.

The Catalan Wikipedia community is diverse in terms of roles, but an equal distribution does not occur in other communities either. For instance, as seen in Table 5, the percentage of edits made by the Japanese administrators and overall by the editors with functional roles is very low. This is due to the few number of editors with an administrator role and to the lack of allocation of the functional roles among editors. German Wikipedia relies on a large number of editors in the Production Force, while the Basque Wikipedia is based on few administrators.

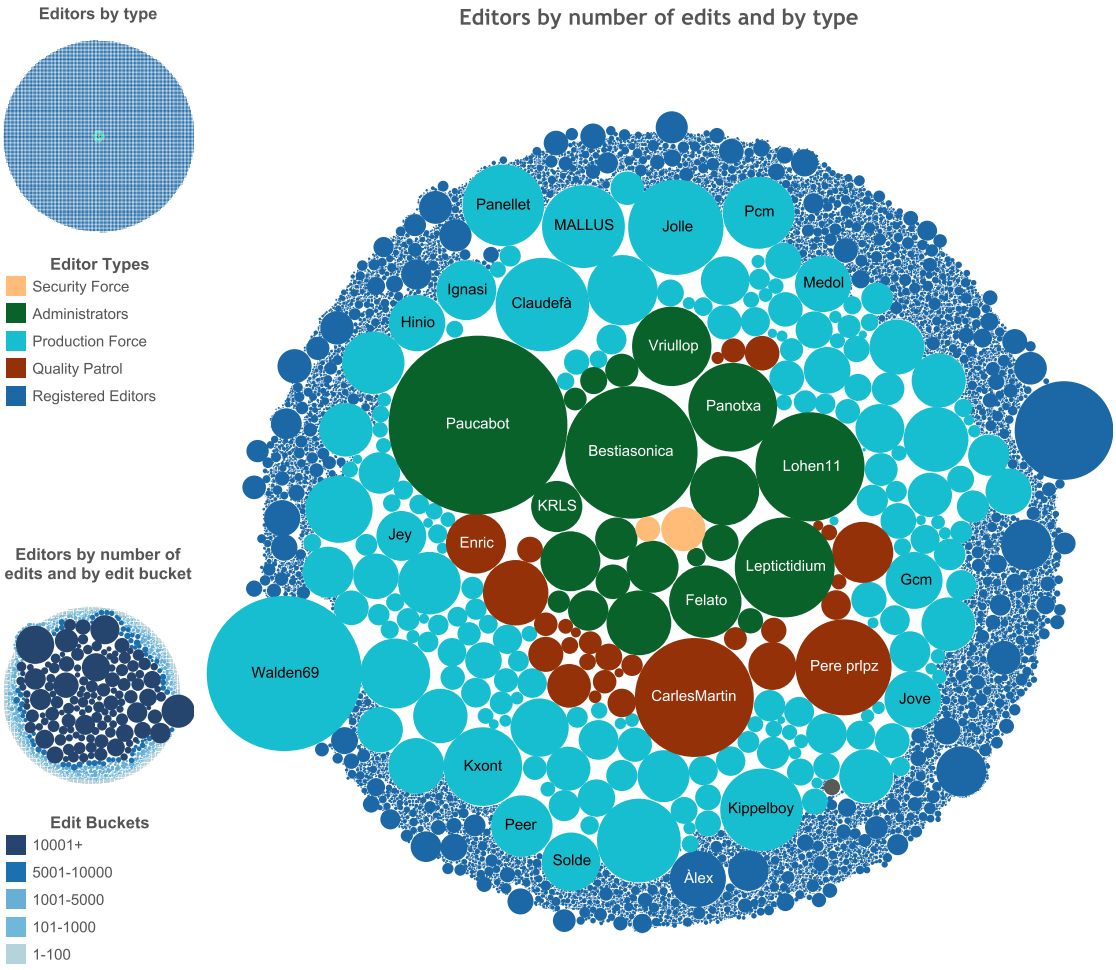


Figure 10. Representation of the Catalan Wikipedia community by number of edits. *Centre right: Editors by number of edits (area size) and editor type (colour). Top left: Editors by editor type (colour). Bottom left: Editors by number of edits (area size) and edit bucket (colour).*

To have an additional perspective on the relationship between functional roles and participation, in Figure 11 I sorted the number of editors in the edit buckets, and the percentage of edits each edit bucket accounts for (see Section 5.3 for a definition of edit buckets). Next to this, in Figure 11 I also depicted the share of edits made by each editor type. In fact, editors in this fifth bucket account for almost two thirds of the total of edits. These are the core of the community, they are not less numerous than the fourth bucket (5001-1000 edits), which make around ten times less. The third bucket, 1001-5000 makes also a larger percentage than the fourth bucket, which may imply there exists a plateau difficult to surpass, and when editors do surpass it, they end up contributing beyond 10,000 edits. Regarding the functional roles contribution, the figure shows very different distributions throughout the different language editions; Japanese language edition is mainly made by registered editors, while, in Basque, Icelandic and Macedonian an important share of the edits is made by administrators.

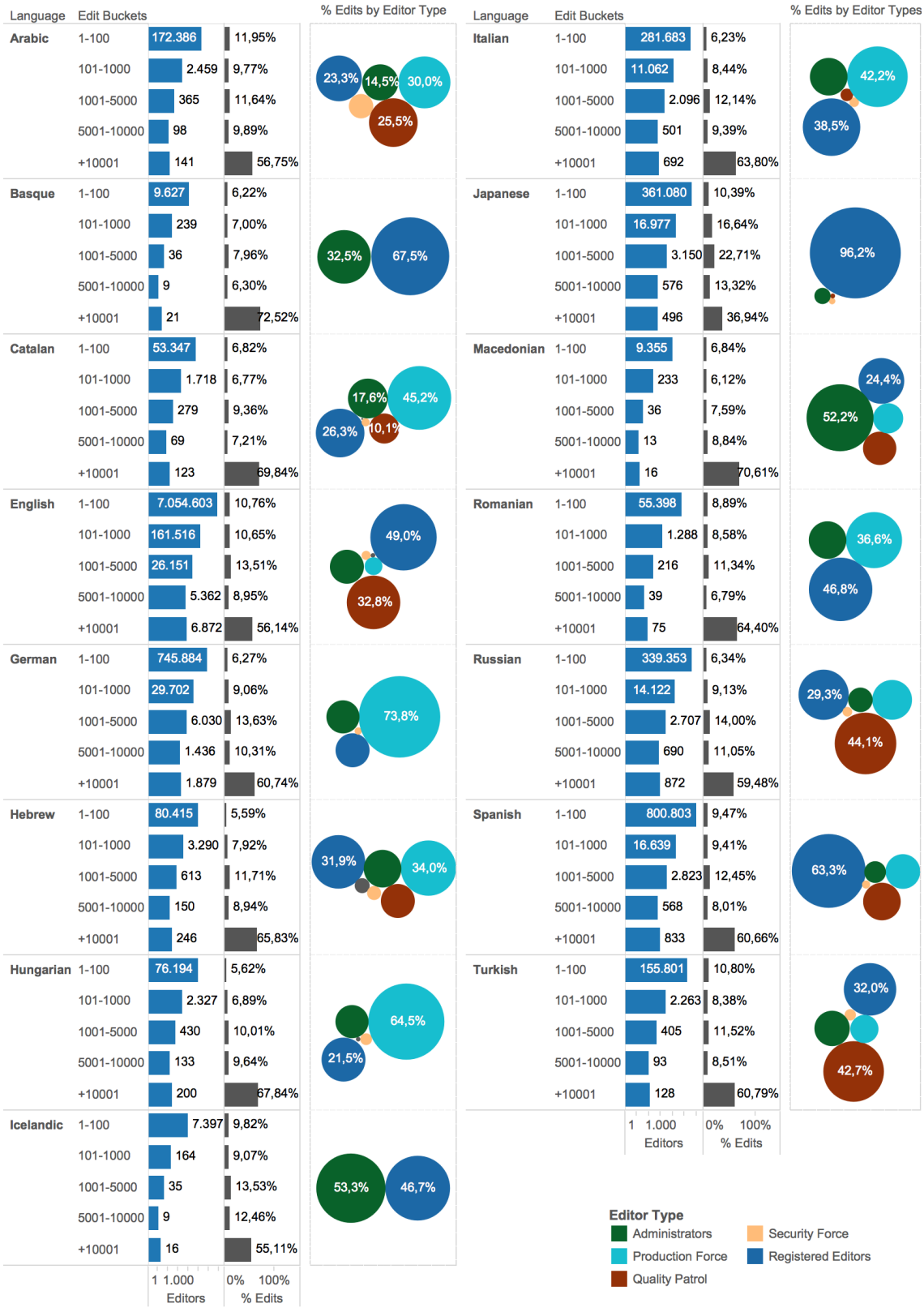


Figure 11. Number of editors by edit buckets and proportional contribution by edit buckets and editor types. *The first two columns show the number of editors by edit bucket, and percentage of edits in each Wikipedia language edition by editors in edit buckets. The last column represents the percentage of edits made by editor type.*

c) Editor types and session characteristics

To further analyse participation, I use the session as a compound work indicator additional to the plain edit. This way, it is possible to explore in greater detail the differences between editor types, and their relationship with participation. With this aim in mind, I identified each session for each editor throughout the entire editing history of the Catalan Wikipedia (2001-2015) by taking into account the one hour cut-off suggested by Geiger & Halfaker (2013). I obtained the number of Bytes and edits performed, as well as the time they lasted and the time elapsed between sessions (inter-session time).

Results. Figure 12 represents the proportion of editor sessions according to the hour of the day they started and by editor type. Generally, editors follow a 24-hour rhythm: they mostly start at 6-9 am, then stop for lunch at 12-13, to conclude at 20-22. However, functional roles (especially quality patrol and administrators) tend to start of their activity and reach its peak earlier, and be more stable during the day. Instead, plain registered editors and anonymous ones carry on their activity mostly in the afternoon.

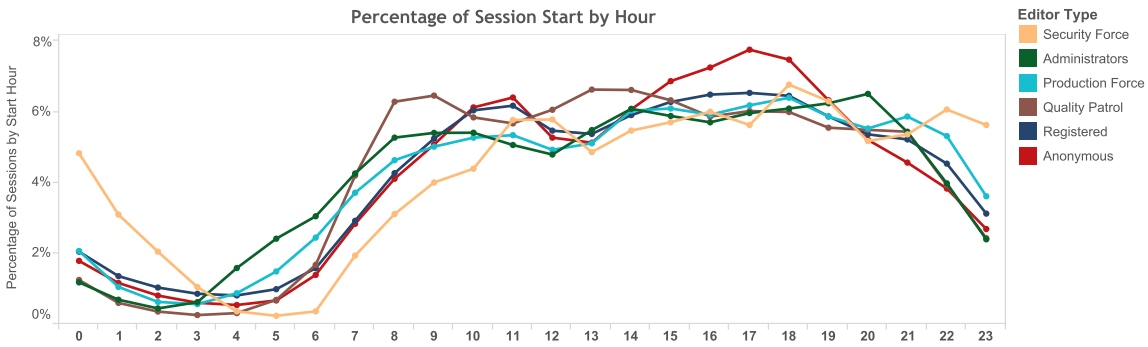


Figure 12. Number of sessions by editor type and hour of the day.

Figure 13 represents the proportion of editor sessions according to the month and quarter of the year by editor type. This graph also supports the idea that administrators and quality patrol are the most stable in their engagement, closely followed by production force. In the third quarter of the year (July-August-September) there is an important decrease in the number of sessions by anonymous and registered editors, which can be easily attributed to the appearance of seasonal activities (e.g. holidays, etc.). Nonetheless, administrators and quality patrol maintain their participation at a more stable level throughout the year, as compared to the rest of functional roles. In fact, August is, unsurprinsingly, the most peculiar month of the year in terms of participation – and its effects over anonymous editors are very intense.

In order to thoroughly examine the process of engagement, Figure 14 uses a box-plot to depict the time elapsed between sessions by different editor types. The data takes 1.5 times the Interquartile Range from left to right whiskers (the box is defined by lower and the upper quartile with the median in the middle). For the different editor types, the administrators present the lowest inter-session time in number of hours (median 17 hours), followed by security force (median 28 hours), quality patrol (median 32 hours), production force (median 46 hours), and registered editors (median 134 hours). The data for each editor type also shows the transition core-to-periphery. Namely, the centre of the

community is constituted by few security force editors, administrators and quality patrol. These editors generally carry on their editing activity on a daily basis.

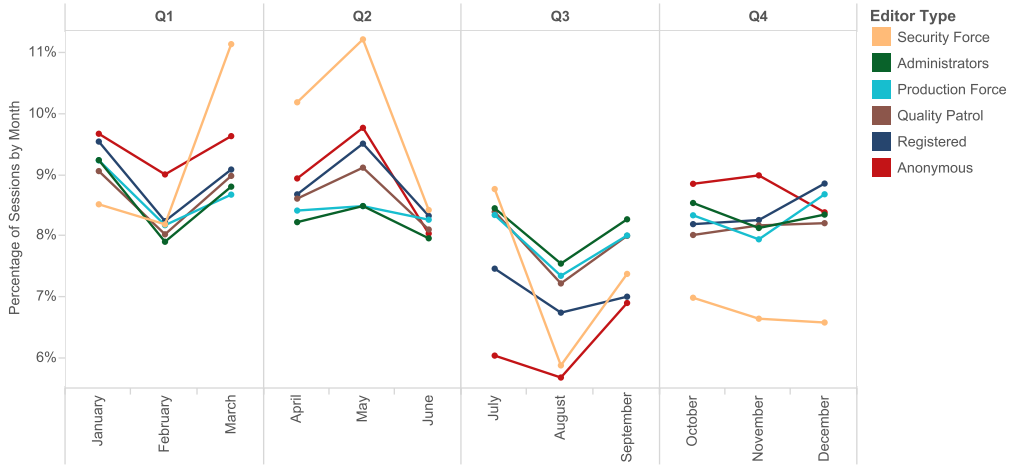


Figure 13. Number of sessions by editor type and quarter of the year.

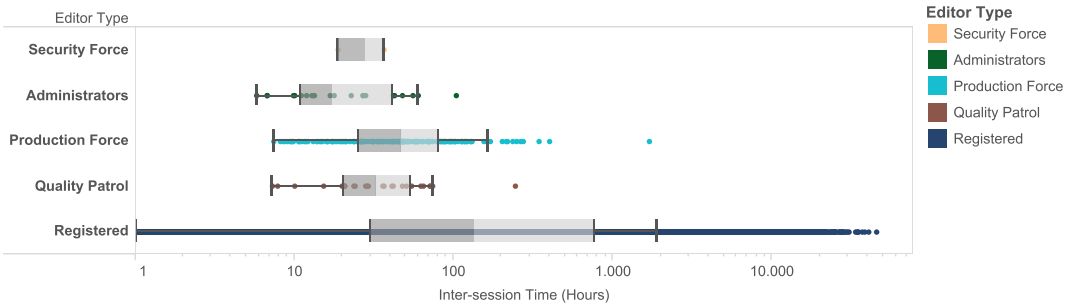


Figure 14. Inter-session time by editor type. Each dot is an editor.

In Figure 15, I compare the session characteristics of the different editor types and editors in the edit buckets. When comparing edit buckets, it is clearly visible that editors who have achieved a higher level of participation have longer sessions, perform more edits, and contribute more Bytes to the pages they interact with. As I previously pointed out, the distance between the edit buckets 1000-5000 and 5000-10,000 is not as big as the one between as 5000-10,000 and the last bucket, whose editors have achieved a higher participation which also corresponds to their editing session characteristics. As far as editor type is concerned, their session characteristic reveals the pattern transition of administrators > production force > quality patrol for the number of Bytes they contribute in the session, and the session duration. This is the core-periphery transition already mentioned, although in this case the security force editors were not the most engaged – it is important to note that they are only two editors. Generally, there seem to be a relationship between the level of rights and the session characteristics.

In order to determine whether differences between groups of editors are statistically significant, a Kruskal-Wallis test was conducted to the different editor types and within the edit buckets. In addition, to verify which group is different from which other group, I also conducted a Dunn's test (1964) procedure with a Bonferroni adjustment. Results are consistent with those from Figure 15 for both editor types and edit buckets with a p-value

< 0.001 for most cases (test results are located in Table 28 at Appendix 3). Regarding edit buckets 1001-5000 and 5001-10,000, it turns out their differences are not significant in term of number of edits and Bytes performed within the same session. The rest of results indicate that a high participation might also depend on different session characteristics. In other words, maybe some editors achieve a high participation because they create specific habits which affect the session characteristics. Regarding the editor types, the transition core to periphery becomes again visible. Nonetheless, editors from the group production force and quality patrol are not significantly different in number of edits, neither are production force and administrators in number of Bytes.

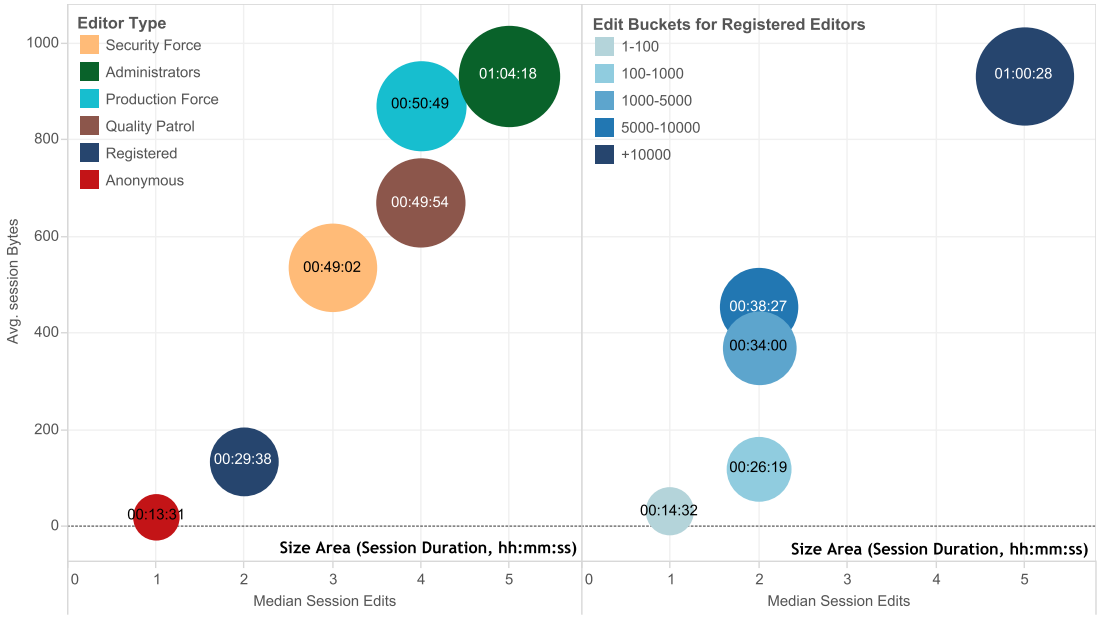


Figure 15. Session characteristics (median number of edits, median number of Bytes and median session duration) by editor type in Catalan Wikipedia.

Discussion. All in all, there is a strong relationship between achieving a high participation and holding a functional role in the community. The functional role is a community identity characteristic which determines many aspects of the editor behaviour - albeit there can be diversity in the behaviour of the editors with functional roles. Productive force and quality patrol editors develop complementary tasks in the community, in order to grow and protect the content. Administrators seem to encompass both productive and patrolling tasks, showing higher levels of engagement if we consider their participation in number of edits, session duration or Bytes introduced. Each language community decides how to distribute the functional roles among the community editors. Functional roles can be either requested by the editor himself or offered by the current members. The distribution of roles among the community members is subject to clear size issuess (small communities do not complement roles that clearly and tend to have only administrators), but also to cultural factors. In fact, the Wikimedia Foundation analytics team found out that the administrator flag is no longer granted to new editors in the English Wikipedia[66]. Given the high coincidence between participation and functional roles, this is an issue each community should plan and work on.

[66] https://strategy.wikimedia.org/wiki/Editor_Trends_Study

6.3.4 Multilingualism (RQ3)

Editors who have built a community identity may develop an attachment to the project as a whole, and adapt their behaviours to favour it in a more complete way than initially planned. While this could relate to the topics they write on, it could also imply a shift towards acting in multiple language editions. To analyse the relationship between participation and multilingualism, I have identified each editor as primary multilingual, primary non-multilingual and non-primary in relation to each language they edit (see the criteria at the basis of this division in Section 6.2.3). Further on, I calculated the percentage of primary multilingual editors for each edit bucket and each language edition. In addition, I calculated for each editor (considering all the edit buckets and editor types) the ratio of number of edits in the primary language as compared to his total number of edits in all languages. Such ratio shows the focus on the primary language and the percentage of multilingualism penetration as a condition, but also as a quality of participation.

Results. Figure 16 shows that the more participative editors are in their primary language, the higher the probabilities they become multilingual (left column: % Primary Language Multilingual Editors). Still, instead of distributing their participation into multiple languages, they tend to gradually focus on their primary language (% Primary Language Edits becomes larger). These patterns seem to be shared across the analysed Wikipedia language editions, with the exception of Icelandic, which has few editors within the bucket of 5001-1000. Big and small languages differ in the percentage of primary multilingual editors both in total (e.g. they are a 14.50% in the Catalan, 12.06% in the Icelandic vs a 2.79% in the Japanese and a 2.92% in the English), and also within each of the edit buckets. In general, small languages show a higher penetration of multilingualism, confirming the results obtained by Hale (2014), who considered a sample of two months of editing activity. Additionally, in Figure 17 the same analysis is repeated for the editor types. As expected, the known pattern core-to-periphery appeared: security force, closely followed by administrators are the most multilingual type of editors. Quality patrol group of editors shows a higher percentage of multilingual editors than production force editors. One reason could be that quality patrol is a more homogeneous group than production force in terms of participation. Regarding the distribution of edits in primary language with respect to the total edits in all languages, quality patrol and administrators show a higher focus on the primary language than the production force.

In order to verify differences at editor level, a Kruskal-Wallis test has been first applied to the edit buckets and editor types. Later, the Dunn procedure with Bonferroni adjustment was conducted to determine significant differences among each group. Test results are mostly significant with a p-value<0.000 (Table 29 for edit buckets and Table 30 for editor types Appendix 3). Regarding the edit buckets, it is possible to affirm that the more participative editors are, the more they focus on their primary language. However, differences between the highest buckets (1001-5000, 5001-10,000 and 10,001+) are sometimes not significant for medium-small language editions, which implies that for editors with a high number of edits, this trend is not that clear. When it comes to editor types, there are significant differences only between registered editors and functional roles. This means the latter generally have the same editing behaviour throughout the distribution or spread of their edits among the different language editions.

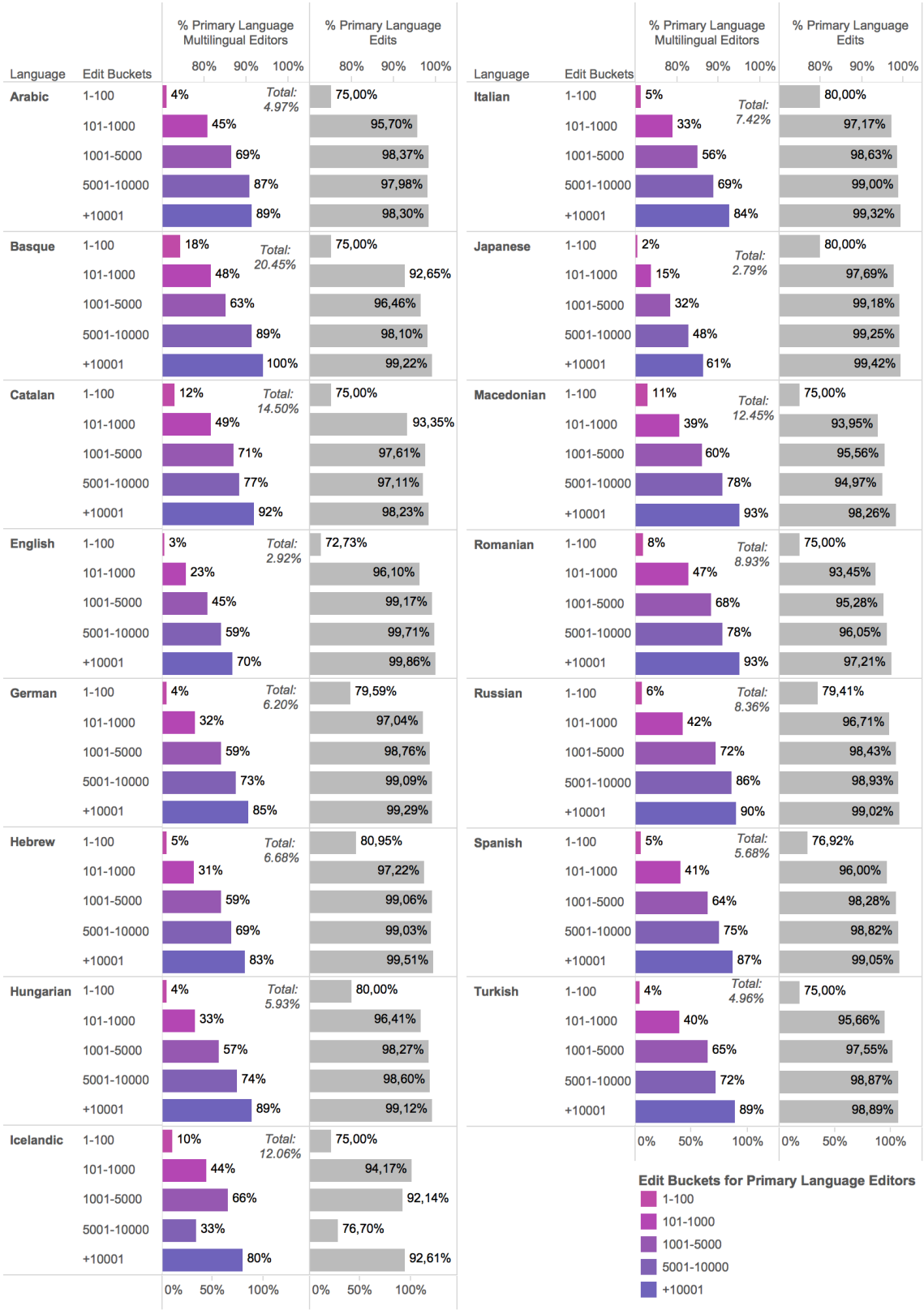


Figure 16. Percentage of primary multilingual editors by edit bucket (% Primary Language Multilingual Editors), and percentage of edits in primary language in relation to total edits (% Primary Language Edits). *Kruskal-Wallis test was applied to the % of Primary Language Edits. The values of chi-square are statistically significant for all the results and p-values are always lower than 0.001.*

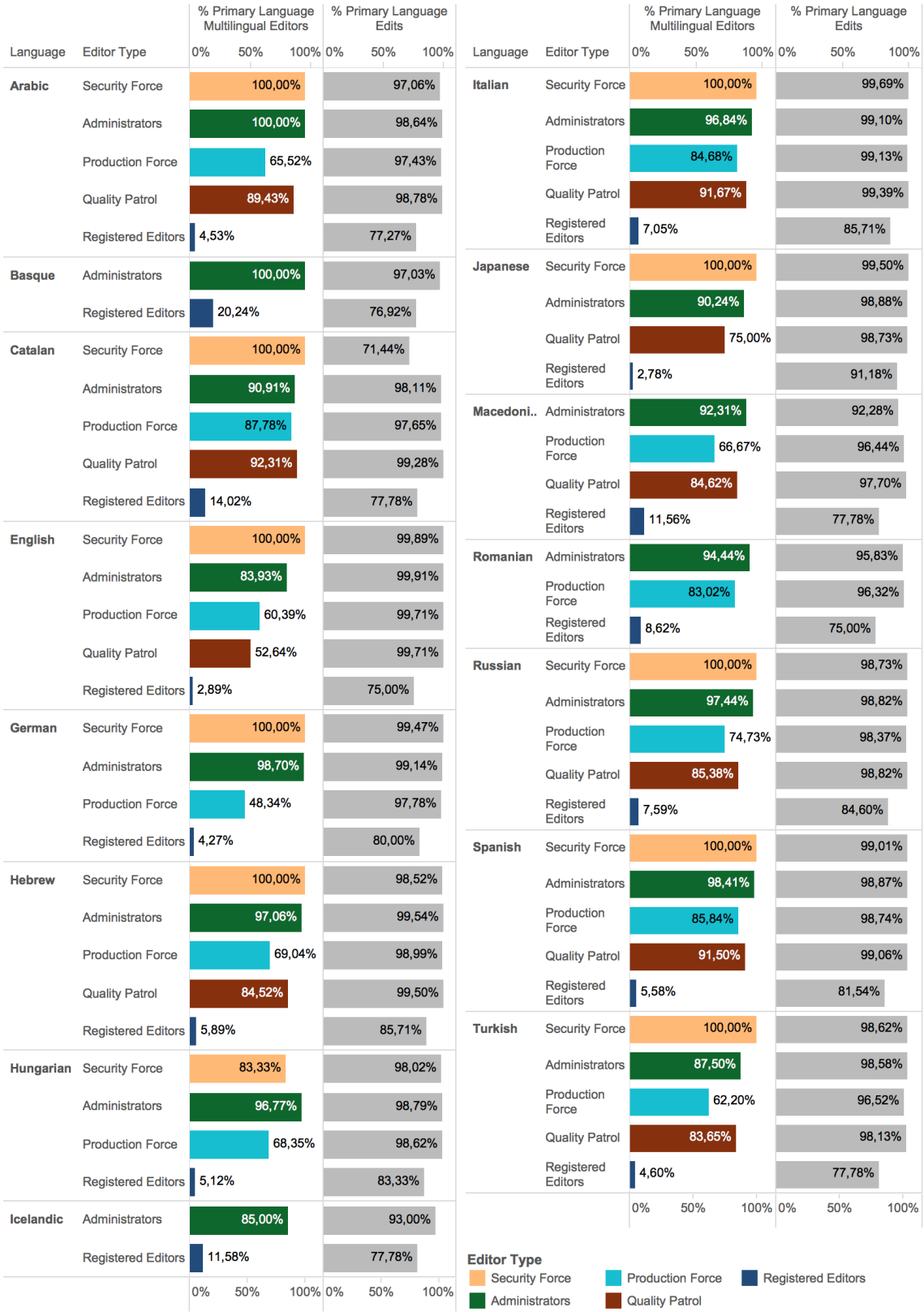


Figure 17. Percentage of primary multilingual editors by editor type, and percentage of edits in primary language in relation to total edits - primary language and non-primary languages (% Primary Language Edits). *A Kruskal-Wallis test was applied to the % Primary Language Edits. The values of chi-square are statistically significant. The p-values are lower than 0,01 for Basque and 0,001 for the other languages.*

Multilingualism appears as condition of those editors who have reached higher participation levels **(RQ3)**. In this sense, an additional analysis in Table 6 shows the percentage of non-primary editors in each Wikipedia language edition classified by editor type. The small proportion of editors who achieve a functional role in a non-primary language confirms the difficulty of combining activities in more than one language community. Production force appears to be the one with a higher percentage of non-primary language editors, and this is also the functional role that is more easily granted a flag. In Table 6 it is also possible to see that small languages have a higher percentage of non-primary editors (e.g. in the Icelandic Wikipedia, the non-primary editors are the majority!). This is in line with the results on the higher percentage of primary multilingual editors, already shown in Figure 16.

Table 6. Percentage of non-primary editors to each Wikipedia language edition by editor type.

Language	Security Force	Admin.	Production Force	Quality Patrol	Registered Editors
Arabic	16.67	3.45	35.45	4.65	9.30
Basque	-	0	-	-	56.05
Catalan	0	0	6.74	13.33	35.52
English	0	1.25	0	6.77	5.62
German	0	0	8.67	-	11.31
Hebrew	0	0	4.01	1.18	12.56
Hungarian	0	0	5.26	-	16.68
Icelandic	-	4.76	-	-	55.97
Italian	0	0	3.07	7.69	15.19
Japanese	0	0	-	11.11	7.03
Macedonian	-	0	32.76	0	49.84
Romanian	-	0	28.95	-	24.26
Russian	0	0	13.12	2.79	10.69
Spanish	0	1.56	1.96	4.67	8.55
Turkish	0	0	29.15	5.36	10.73

Discussion. The results show that multilingualism is a further step in building a community identity, considering that the editors should embrace the Wikipedia project as a whole. Editors with a high participation in their primary language tend to become multilingual. This is in agreement with the core-periphery continuum with the functional roles: central roles are more likely to become multilingual editors. Yet, the analyses show that multilingualism has limits. It does not settle as a complete behavioural change, since editors do not equally distribute their participation throughout the various languages they edit.

Furthermore, editors do not usually obtain a functional role in a language edition where they are not primary editors. Table 6 shows that in each Wikipedia language edition there is a small percentage of editors with a functional role who are non-primary editors. This could be explained by the language barrier, or by the difficulty of achieving credibility and coordination with editors from more than one community. Editors who obtain a functional role in a non-primary language could be editors native in that language but who nonetheless prefer to edit in English. The proportion of non-primary editors is bigger amongst plain registered editors, especially for smaller languages like Basque, Icelandic and Macedonian. This could be explained by the effect of editors whose primary language edition is not their native language, and by a simple matter of scale. Most multilingual

editors edit sequentially the same content across languages - as a way of spreading their contribution to multiple languages (Hale, 2014). With only a few edits by each editor, in small language communities the number of these non-primary editors seems larger proportionally, compared to how it would seem if the language community were bigger.

6.3.5 Community Oriented Activities (RQ4)

One way to check whether an editor embraces the Wikipedia project as a whole is looking at the level of multilingualism. Nonetheless, since editors do not distribute their participation throughout the different language editions they contribute to, perhaps they engage in other activities in the same language which are clearly oriented to the benefit of the entire community. In order to assess the proportion of participation in these activities, the percentage of edits dedicated to Data Spaces, Community Communication and Personal Communication have been measured for each editor. I chose these three activities as they represent very different contributions with different aims and repercussions. I propose looking at the proportion of edits dedicated to these activities to see whether it increases or decreases along with the editors' overall participation in Wikipedia.
It must be remarked that these are side-activities, different from contributing to articles. For instance, for editors with more than hundred edits, Data Spaces accounts for a median (to all editors in a language and to all languages) of 1%, Community Communication a median of 3%, and Personal Communication a median of 5%. To untangle the possible relationships between participation and community-oriented activities, I have relied again on the Kruskal-Wallis test to see if the edit buckets show significant differences among them, and the Dunn procedure with the Bonferroni correction to precisely determine the differences between edit buckets. Detailed results are located in Table 32 and Table 33 in Appendix 3.

Results. Figure 18 presents the mean rank values for the three activities and for each edit bucket. There is a visible increase for the three activities in each edit bucket and in all languages. Since the test is applied to the same editors, it is visible that the proportion of edits in Data Spaces grows more than in the Community Communication, which in turn is higher than the one in Personal Communication. A growth in participation to Data spaces and Community Communication is visible in all fifteen languages. Instead, Personal Communication stops growing at the third bucket. This suggests that Community Communication and Data Spaces are more strongly related to participation than Personal Communication. In fact, the pairwise comparison showed significant results (p-value < 0.000) mainly between the first two buckets and the other buckets. Only in bigger languages such as English or Japanese, editors who exceeded the 1001-5000 edits also grew their proportional participation in these activities. The general trend also shows that differences between buckets in terms of Personal Communication are less significant than in the other activities, in line with the mean rank values. It is interesting to notice that even though in absolute numbers they dedicate more edits to Personal Communication, what is most indicative of highly participative editors is their participation in Community Communication and Data Spaces. All in all, it is possible to state that there is a possible relationship between adopting community oriented activities and growing in overall participation **(RQ4)**.

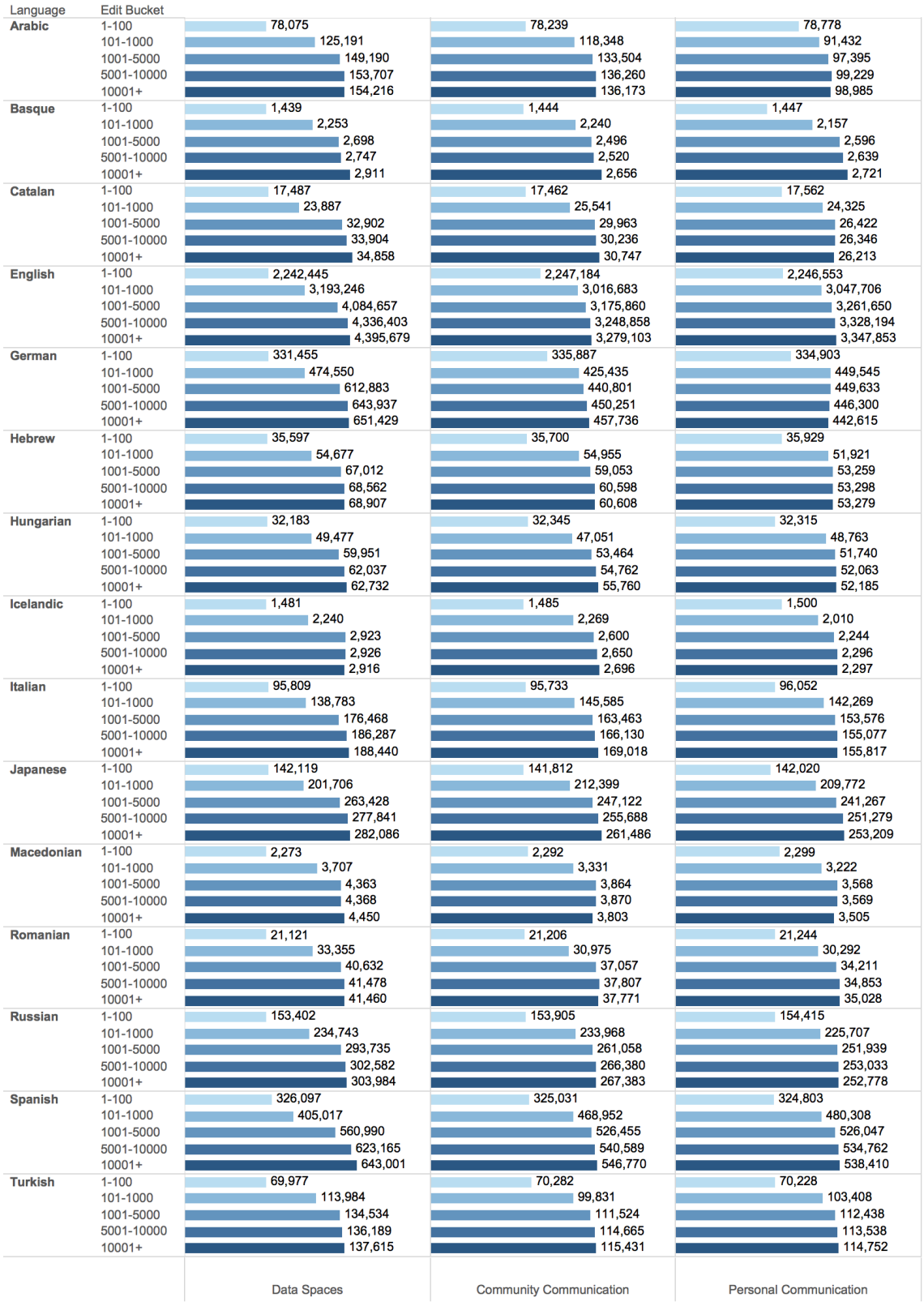


Figure 18. Proportion of edits in different activities (Data Spaces, Community Communication, Personal Communication) by edit buckets. *Mean ranks for the proportion of editor edits in CIRA by Edit Bucket. Results of a Kruskal-Wallis test are statistically significant for all languages with p-values always lower than 0.001*

Discussion. These results are consistent with the idea that an editor developing community identity dedicates greater efforts to coordination, creating tools for others, uploading images, among others. Personal Communication is an activity that receives more participation, but it is nonetheless the one which grows less. In a similar way to the results from the MovieLens online community (Ren et al., 2012) presented in Section 6.1, participating in wider-scope activities for the community is associated to a higher overall participation. Perhaps if new editors were introduced to the editing activity by means of training provided by more experienced editors, perhaps they could initially achieve higher participation in the activities of Community Communication and Personal Communication. This could help them engage in building their community identity from the very beginning, and possibly increase their overall contributions to Wikipedia.

6.3.6 Retention and Survival (RQ5)

a) Long-term retention: active periods

In the previous sections, I established the different features associated to a Wikipedia community identity. I assumed that an editor who cultivates such features would feel more identity-congruent with every new interaction, and consequently, increase the participation in a positive loop. In this section, I turn the attention to the relationship between participation and retention. It is expected that editors who participate more will have a stronger retention. However, there is not much research on not very active editors who still return to the project – a different manifestation of engagement than the usual participation. In order to analyse this relationship, five languages have been selected (Catalan, German, Hebrew, Japanese and Russian) and measured for each editor the number of periods of 60 days in which s/he made at least an edit. This period is the 'survival period' the Wikimedia Foundation establishes as a standard to consider an editor continues in the project after the first edit. Periods of sixty days in groups of six periods have been aggregated and counted as an active year. Later on, editors have been divided by number of active years. In each year, the proportion of editors belonging to each edit bucket has been measured and represented.

Results. Figure 19 shows the relationship between participating more and having a higher retention **(RQ5)**. Editors with more than 10,000 edits tend to spend more than four years in the project, being the majority after the ninth year. In the first and second year, most editors have achieved around 1000 edits. Yet, interestingly enough, after the fourth active year there is still a numerous group which have not yet surpassed the first hundred edits: roughly and in average, they made a maximum of two edits every month. In fact, the number of these editors is very similar or even larger than the core of the community.

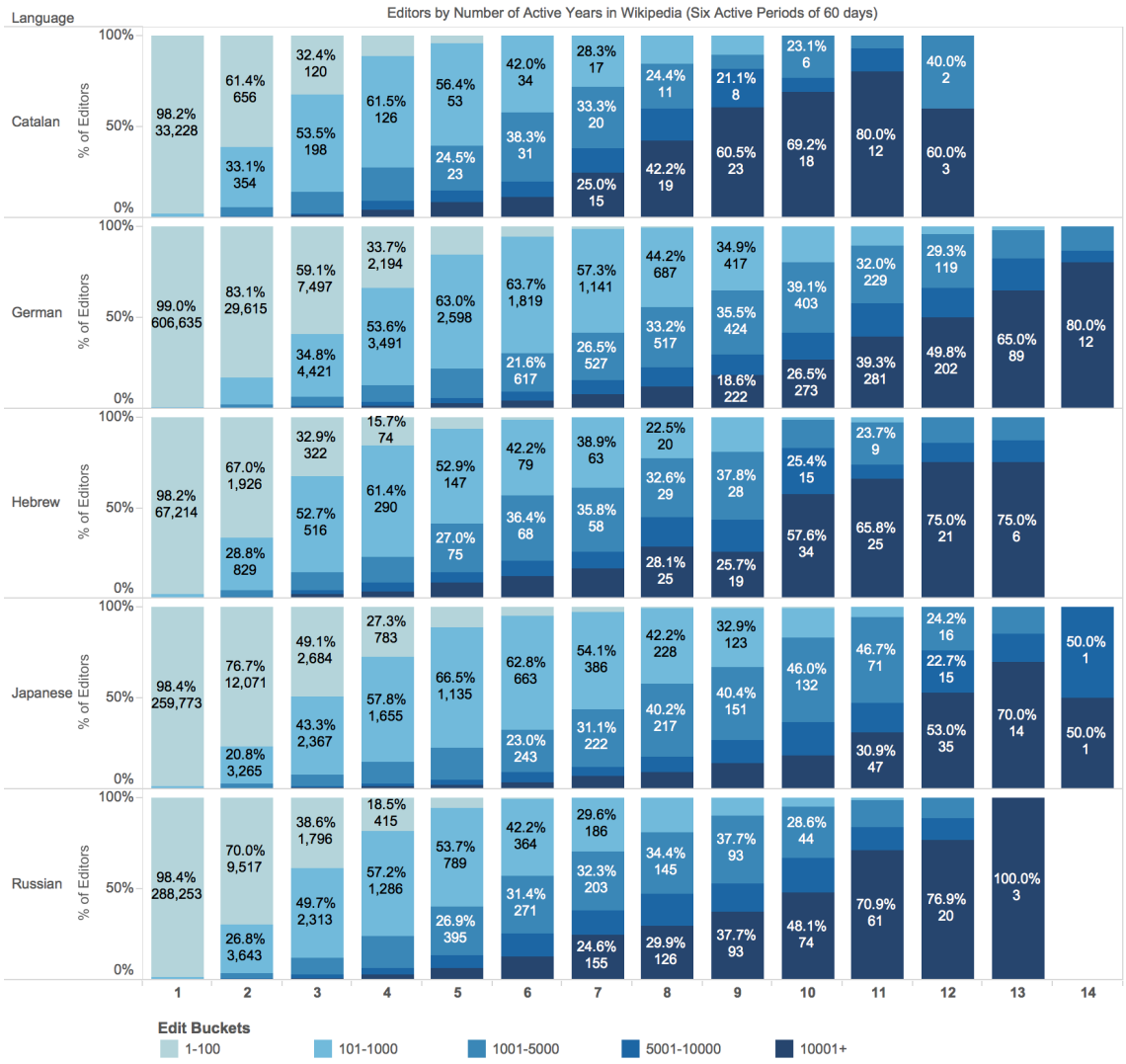


Figure 19. Editors (total number and percentage) by number of active years in Wikipedia (an active year is made of six periods of 60 days with at least an edit in each) and by edit bucket.

Figure 20 presents the same analysis on editor types. It is readily apparent that editors who obtain functional roles last at least three entire years in the project. Also, the core-periphery continuum pattern is visible: administrators are engaged on the longest term, followed by quality patrol and production force. This suggests that, as in the case of the number of edits (or session characteristics), the time spent in the project is definitely a user characteristic associated to those editors who obtain functional roles with higher levels (administrators and security force). Furthermore, the graph clearly shows that different language editions employ different strategies in distributing functional roles. For instance, languages like German give the production force flag to editors after the first year, while Japanese does not have production force and only grants few administrator flags and after at least five years.

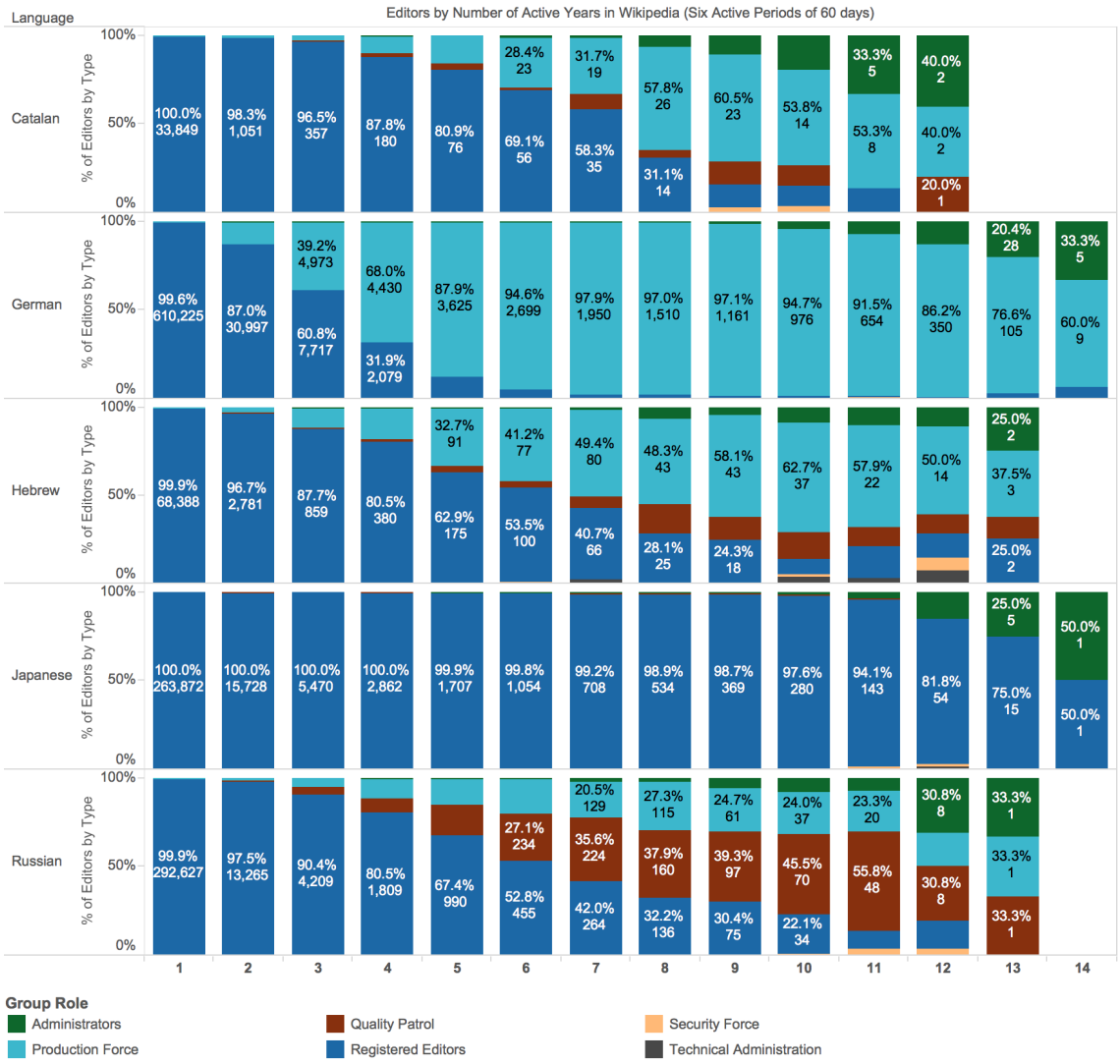


Figure 20. Editors (total number and percentage) by number of active years in Wikipedia (an active year is made of six periods of 60 days with at least an edit in each) and by editor type.

b) Editor survival and User Page length

As said before, some analysis conducted by the Wikimedia Foundation counted the editors who continue editing after an initial period called 'editing survival period' of sixty days. The proportion of survivors has become lower over the years. Nonetheless, I suspect that the editors who survived this period also developed specific characteristics of the community identity. Following this idea, I analysed the proportion of editors who survive at least a year, in relation to the length of their User Page.

Results. Figure 21 shows a general decrease (the total in blue): the proportion of survivors is around the 50% in 2004, but it decreases until less than 10% in 2014. Among those editors who had developed a longer User Page there is a bigger proportion of survivors; longer User Pages are associated to a higher percentage of survivors. Those who developed complete profiles multiply by 5-6 times the final proportion of survivors.

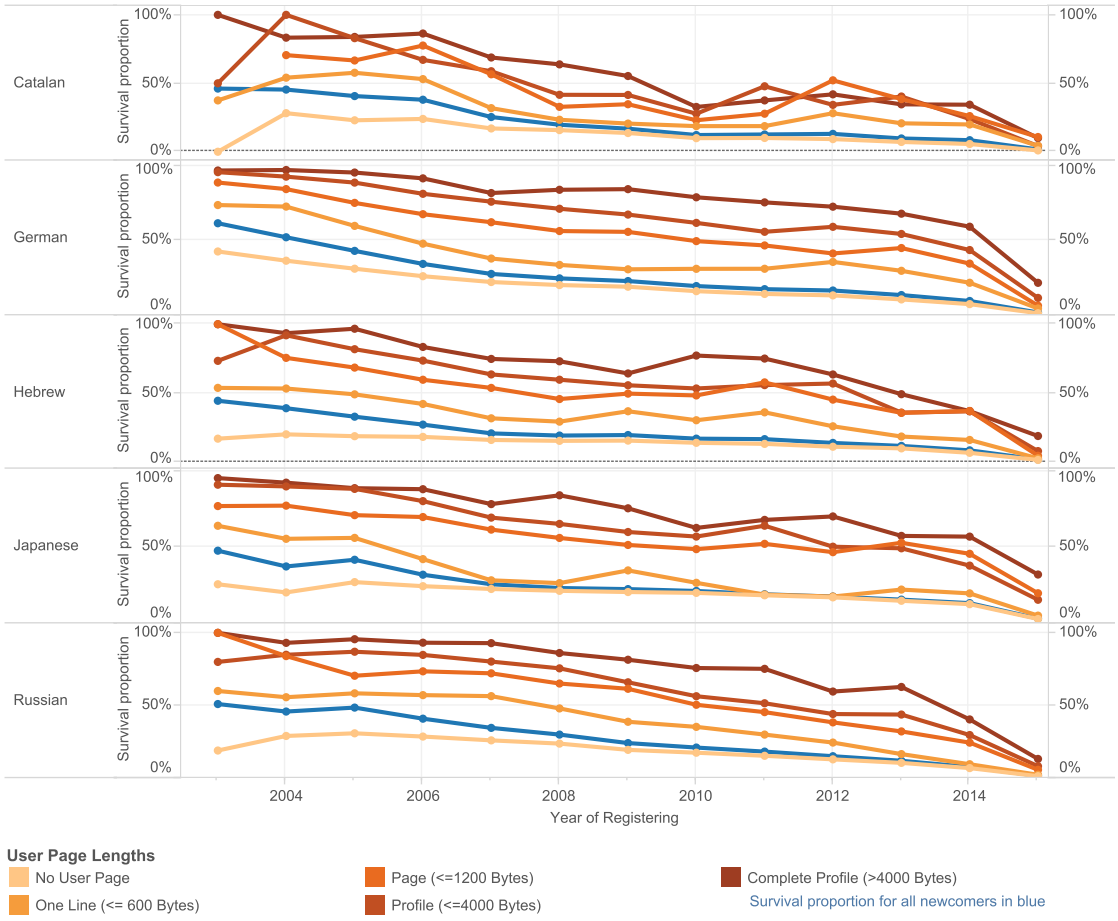


Figure 21. Percentage of editors by their User Page length and who survived a minimal period of six months after registering (percentage of total survivors in blue). *Percentage of survivors are aggregated in the year they registrated.*

Discussion. The results confirm the relationship between participation and retention. While the participation seems to imply retention, there is also a group of editors who enter and perform few edits in the site for a long period of time (over 2-3 years), having accumulated less than a hundred edits. Such behaviour is exceptional considering the usual activities, the structure of the community and its proportions. This group of editors demonstrated to be motivated by the project, although they might not increase their participation (previous research showed that very participative editors can be identified from their very first days in Wikipedia (Panciera et al., 2009)).

Leaving these editors aside, newcomers are the most desirable group of editors to retain. Currently, a very small percentage of new editors survive the first six months. However, results showed that among the editors who had created longer user pages there was a higher proportion of survivors. In fact, creating a User Page is possibly the easiest characteristic to develop from the community identity. Nonetheless, the analysis in Section 6.3.2 showed that those who create a User Page during the first 1-100 edits are below the 10%. The results do not establish a causality between creating a User Page and surviving but, still, suggest that among those who created larger user pages there are higher percentages of survivors.

6.3.7 Summary of Results

This chapter showed that participative editors are the ones who most developed the community identity features. By mining the data from 15 Wikipedia language editions, I characterised entire communities and demonstrated in which features participative editors differed the most. Positive results for each of the proposed features suggest that when editors internalise community values and the project mind-set, consequently become more motivated to continue participating. Participation is in itself the first characteristic of the community identity, and hence, this cycle may repeat as a positive loop.

Editors primarily build their community identity through their work, and not so much through their User Page, which is usual in other online communities or social networks. Editors with a higher participation feature longer User Pages **(RQ1)**, but their use is very unequal. Regarding the entire community, only a five per cent of the editors in their first 100 edits and fifty per cent of the group of editors of 101 to 1000 edits develop a User Page. On the other end, after 5000 edits, about two thirds of the editors develop their User Page with more than just a line. This suggests that participative editors use these spaces as an accessory communicative tool, and new editors do not find useful to create and expand one. In any case, Wikipedia communities could benefit from introducing mechanisms to help new editors introduce themselves to the community.

Editors who acquired a functional role in the community are among the most participative editors. Language editions distribute their functional roles and grant them according to particular strategies, in which factors like their community size and cultural background may also matter. All in all, there is clear a relationship between functional role and participation, both in the total number of edits and in the particular session characteristics **(RQ2)**. Results from session analysis demonstrated that, within this time-frame, administrators are the most regular and participative, followed by quality patrol and production force. In fact, the examination of session characteristics re-confirmed the core-to-periphery pattern: those at the core showed significantly higher values for each of the session characteristics.

Editors tend to become contributors to multiple language editions when they increase their participation in their primary language **(RQ3)**. Hence, becoming a multilingual editor is a way of embracing the Wikipedia project as a whole. This responds to the attachment developed by group members towards a project as a whole, after they had built a group identity (Ren et al., 2012). Results confirm this statement in Wikipedia, although multilingualism is reflected more in spontaneous and casual activity, rather than shaping the entire editor participation. Editors' contributions to non-primary languages increase, but they represent a smaller proportion, taking into account that editors increase more the number of contributions to their primary language.

Editors increase their proportion of participation to community-oriented activities, such as contributing to Data Spaces (e.g. files and images) or Community Communication (e.g. article talk pages, Wikipedia policies, among others) along with their participation **(RQ4).** Acquiring these activities confirms the internalisation of community values, and therefore, a way of developing a community identity. Editors with different levels of participation tend to differ more in Data Spaces than Community Communication. I

compared these results with the proportion of participation in Personal Communication spaces, and I observed that for the most participative editors, the participation in Personal Communication spaces increased in a lower proportion as compared to the other activities. In line with Ren et al. (2012) on the movie-based online community, internalising community identity values has a stronger influence on engagement than developing bond attachments with other editors through Personal Communication spaces.

In the last section, I additionally examined the proportion of different editor types and editors with different levels of participation considering the number of active years spent in Wikipedia **(RQ5)**. One the one hand, results showed that those editors with most retention rarely remain without a functional role. Also, the core-to-periphery continuum is again visible when it comes to the time dedicated to the project: the administrators are the ones who had spent more years on the project. This could suggest time is an indicative factor for role transitioning towards higher levels. There is a relationship between participative editors and having the most retention. On the other hand, even though editors who spend more time in the project have a higher number of edits, still a considerable number of editors had spent 3-4 years in the project with an amount of 100 edits.

All in all, by analysing the community identity features I provided a generalizable view on participative editors across languages. Wikipedia is a very unique object, for its scope and massive use by all sorts of readers, but their editors can be compared to other online communities. In addition to these findings, I suggested several explanations based on the current design. Since the community identity features are positively linked to participative editors, the Wikipedia design could be easily and intentionally changed towards helping new editors develop them.

In particular, I stress the importance of helping new editors to develop these features during their first days in the project. This period of time is decisive in order to integrate them into the community and help them acquire the mechanisms of contribution and socialization. Like any other identity, community identity is a dynamic construction that evolves in time according to the actions taken, and the relationships between those who share it. In Section 9.3 I propose some design recommendations in line with these conclusions.

Chapter 7. Cultural Identities in Wikipedia

In this chapter I present the case study of cultural identities and Wikipedia editor engagement. I initially explain why cultural identity may become salient in any Wikipedia language edition and influence editor participation, and I review the cultural identity main definitions in order to understand its characteristics (7.1). Then, I present the approach to map cultural identity meanings to articles in 40 Wikipedia language editions (7.2). Finally, I analyse the extent of cultural identity representations, its topics and cross-language availability (7.3).

7.1 What is the Cultural Identity?

The identity-based motivation model explains that when an identity becomes salient in a context, it triggers an action and procedural-readiness to act congruently with that identity. Chapter 6 showed that Wikipedia editors who had internalized the values of the community identity gave proof of a higher participation. However, given the broad Wikipedia goal of obtaining 'the sum of human knowledge', I believe that other identities may become salient in Wikipedia. Since editors are allowed to choose the topics they contribute to, they can easily act identity-congruently. Any social identity could become salient, as long as the contributions follow the specific content rules and guidelines. Cultural identity is a collective identity whose meanings may be shared, at some extent, by editors from any language edition; this is why I believe it is worth studying it. But, what exactly is cultural identity?

"Cultural identity refers to one's sense of belonging to a particular culture or ethnic group" (Lustig & Koester, 2010, p. 142). Differently from the social identity, which is often considered empty in terms of meaning, cultural identity involves also learning, embracing, and embodying "the traditions, heritage, language, religion, ancestry, aesthetics, thinking patterns, and social structure of a culture" (Lustig & Koester, 2010, p. 142). Any individual can become member of one or more cultures. Therefore, cultural identity is a broad and useful concept to analyse content whose meanings are shared by a group of people (e.g. Wikipedia editors). However, it is necessary to understand how cultural identities are constituted and created, as well as the relevance of context.

Cultural theorist Hall (1990, p. 223) defines cultural identity as "the common historical experiences and the shared cultural codes". Culture is about shared meanings, such as language, territory places, artistic creations, traditions, among others. He emphasizes the idea that meanings originate around a place. This is a very prevalent idea in social sciences. The anthropologists G. Hofstede, Hofstede, & Minkov (2010) affirmed that "culture is a collective phenomenon because it is shared with people who live or lived within the same social environment".

According to Hall, one of the most important aspects of cultural identity is its dynamic nature. It is a matter of becoming as well as of being. Its creation is not fixed, and it is in constant relationship with history, culture and power in territories. Likewise, individuals'

cultural identities can undergo changes because of their integration into different places, their mixing with other communities where different cultures are practiced. People's cultural identities are the sum of experiences with other people in precise places.

Hall affirms that cultural identities are represented, and that happens when their "shared meanings or shared conceptual maps" use language system as a vehicle (Hall, 1997, p. 18). In fact, they can coexist in language: for instance, British and North American cultural identities may share meanings despite being in different territories. Some languages may also coexist in the same territory, giving place to different cultural identities with shared meanings about their surrounding environments. This makes the creation and representation of a cultural identity a variable geometry. Only in some cases where territory sovereignty coincides with the territory of cultural practice, cultural identity shared meanings coincide with those of a national identity (as in the case of Icelandic cultural identity). This reaffirms the idea that originally, cultural identities are tied to territory, but they constantly evolve in their representation.

7.1.1 Cultural Identity Meanings in Wikipedia

In Wikipedia, articles can imbue meanings related to any editor identity, including the cultural identity codes associated to the territories editors live in. Editors' cultural identity is the set of meanings they identify with and that can be possibly developed into articles; local associations, places or traditions are candidates to be part of the encyclopaedia. *In line with identity-based motivation, I hypothesize that editors will act congruently with their cultural identities and repeatedly engage in representing them in the encyclopaedia*. The saliency of this identity might encourage the participation of all kinds of editors, and its effect might be considerable in the content. Because whenever it is possible to choose between different options, the option linked to one's own cultural identity possibly becomes more salient and motivating as compared to others.

Wikipedia's initial mission statement and the current specific content policies do not encourage the representation of language communities' idiosyncrasies. However, each language content has proven to be diverse. Any community initiative involving the editors' most immediate environment has been a success. For instance, global projects such as 'Wiki Loves Monuments' and 'Wiki Loves Earth' (where editors had to contribute with pictures of monuments and of landscapes respectively) have been a success in many Wikipedia language editions. This may anticipate that editors, in an appropriate context, make identity-congruent contributions to their Wikipedia language editions, and cultural meanings are more easily contributed.

For each Wikipedia language edition, I propose obtaining all the articles representing the shared codes of their associated cultural identities. Therefore, looking at its characteristics it will be possible to determine if there has been a higher engagement, as an indicator of the influence of this motivation type. In addition, having such a corpus of articles may be useful to study articles' composition; moreover, the corpus represents a very valuable tool to understand the editors' culture. In Wikipedia, previous works have already employed the term cultural contextualization to highlight the content differences in different Wikipedia languages editions, and such differences have been attributed to the context

(Hecht, 2013). So far, contextualization effects in Wikipedia can be classified in two groups:

1. **Community effects**
   These effects are based on the idea that since each language edition constitutes a community, its editors only contribute with limited points of view to the content of Wikipedia articles. Wikipedia is organized around language communities, but generally there is no direct correspondence between languages and nations. Hence, the terminology employed to refer to the editor's point of view varies: 'linguistic point of view', 'national point of view', or 'cultural bias'. For instance, some studies have remarked that these differences in the point of view are more prominent when it comes to controversial topics, where history and politics is seen from opposite positions (Apic, Betts, & Russell, 2011; Massa & Scrinzi, 2011).

   Other authors like Callahan & Herring (2011) explored how the biographical articles of well-known people are more complete in the language editions associated to the territories where the person is from. Similarly, Rogers et al. (2012) compared an article dedicated to the historical event 'The Srebrenica Massacre' throughout different Wikipedia language editions including English and Balkan languages. The study shows how the *same* article in different language editions adopt a different point of view to illustrate facts; such points of view are sometimes unified, other times in total disagreement when it comes to the terminology employed and its political connotations.

2. **Geographical effects**
   These effects are based on the idea that geographical context affects editors' interests in two ways. First of all, editors decide to focus on their territories, and secondly, their interests differ from the interests of editors located in distant locations. For instance, in a study on editing interests, Karimi et al. (2015) gathered all the editors' interests and analysed their relationships in order to determine how close their affinities were. Results showed that editors from close places tended to edit more similar articles than editors from distant geographical locations. In a different way, Hecht and Gergle (2010a) computed the location of each anonymous edit in geolocated articles and discovered that many of the contributions were made from close distances.

   Another effect detected by Hecht and Gergle (2009) called 'Self-focus bias' explains that articles located in the countries local to each language edition are linked by many more articles (i.e. they have more inlinks) than the articles located in the other countries. The geographical factor is also used to explain the fact that, while each Wikipedia language edition presents a diversity in content and has unique articles, those language editions whose editors' territories are geographically closer, tend to share more articles (Warncke-Wang et al., 2012).

This study proposes going one step further in the study of cultural contextualization and provides a more *complete* and *explicative* analysis of the phenomenon. When saying 'complete' it is meant that by obtaining all the articles that represent the editors' cultural identities, a valuable corpus is obtained to study more in depth the Wikipedia cultural contextualization in each language edition than it has been done in previous studies. This

enables to quantify the extent of this content, the diversity in terms of subjects, and to compare the relationships of this content with the rest of Wikipedia's content. To my knowledge, this is the first study to extract and quantify the extent of content representing the editors' cultural identities in each Wikipedia language edition.

By 'explicative' I mean that although previous research had measured the effects and identified some of the causes of cultural contextualization, most of these studies lack an explanatory framework to describe the way content editing unfolds. **The present study proposes to investigate the influence of an identity-based motivation that favours identity-congruent participation and the creation of articles imbued of cultural identity meanings.**

7.1.2 The Influence of Cultural Identity on Participation

Building upon previous research on identity-based motivation (see Chapter 4), I argue that an identity-based motivation could foster participation if the scenario has certain characteristics (e.g. freedom to choose certain actions and meanings). Hence, by analysing the data and studying the articles related to cultural identities, their creation and the editors' editing history, I aim to explore the influence of this motivation type. Nonetheless, there are two considerations to be made before starting this analysis:

a) Identity-based motivation is not comprehensive of all the motivational factors that drive participation.

Identity-based motivation framework is rooted in socially situated cognition and states that identities are dynamically constructed in context and therefore, they can trigger specific choices and behaviours. Yet, it does not explain how people get into the context. Although it could be a choice triggered by another identity, it is difficult to advocate that such action responds to an identity. Perhaps because of this, Oyserman (2009, p. 251) affirms that "broader identities (e.g., female) are more likely to be cued than narrower ones (e.g., professor). Gender and race-ethnicity are both broad and also often psychologically salient". Hence it is unlikely that identity-based motivation alone drives participation into Wikipedia; other motivation types like those explained in Section 3.4.3 might also influence the decision of contributing to Wikipedia (e.g. ideology or fun).

This suggests that every individual might be influenced to a different degree by cultural identity and by other types of motivation. Perhaps for someone very aligned with the project's ideology, the influence of cultural identity might only help in choosing between two different topics, while for someone else it could even be the very trigger to register in Wikipedia, login in and edit an article about his hometown article. Both situations are possibly influenced by identity-based motivation and cultural identity, but the composition of motivation may vary for every editor. The editors who feel a very deep connection with their cultural identity history, language and traditions, may mainly focus on representing this content in the encyclopaedia. Instead, editors who are more motivated by some aspects of socialisation, the identity-based motivation may only emerge in very specific moments.

b) Identity-based motivation cannot entirely explain the participation in content representing cultural identities.

In Chapter 2 I discussed the aspects of digital engagement and I explained that user interaction also depended on the object's aspects (i.e. usability, the new actions proposed, among others). Therefore, I hypothesize that participation in content representing cultural identities may complementarily be influenced by other motivation types or aspects. Therefore, some of these factors may also matter*:* a) the cumulative effect of some more prominent articles (already developed and easy to find); b) the community dynamics or duties in case of holding a functional role; c) the information demand by readers; or d) the easiness to write about topics already known – at least to some extent – such as the cultural identity representations. In any case, these hypothetical factors certainly do not invalidate the influence exerted by the identity-based motivation. Differently put, an editor can have a superficial knowledge on cultural identity related topics, but he needs motivation to find the proper references, among other requirements.

These other factors are important to bear in mind; it would be faulty to assume that one can prove the existence of identity-based motivation and cultural identity by measuring participation alone. At the same time, it is not possible to isolate these different factors, as then I would not be measuring Wikipedia editor engagement as it really is. Instead, I propose different analyses to measure the influence of cultural identity and identity-based motivation on editor participation, while taking into account the point of view of the content created to represent the cultural identities and the characteristics of the editors engaged in editing such content.

The very first step of the study is proposing a method to identify and retrieve the Wikipedia content that represents the cultural identities of the editors from each language edition (7.2). Later in this chapter, I will present the research questions and the analysis of cultural identities' representation in Wikipedia (7.3).

7.2 Mapping Cultural Identities to Wikipedia Articles

In this section, I describe the method to map cultural identities to Wikipedia articles in each language edition, in order to construct a dataset. First, I select the list of languages I retrieved the articles from. Second, I explain the criteria by which I include an article into the dataset. Third and finally, I propose a simple mechanism to manually assess the method's success.

7.2.1 List of Languages

Wikipedia language editions are not a fair representation of the language diversity in the world. Among the reasons that explain why some languages do not have a Wikipedia language edition or have it underdeveloped, Ensslin (2011) and Van Dijk (2009) mention, among others, the reduced number of speakers, the digital divide speakers may suffer, and the low online reputation of their language. Some languages like Arabic and Chinese[67]– the latter being a notorious case of censorship[68] – have a reduced number of articles per number of speakers; in Catalan and Swedish language editions the opposite phenomenon occurs. The enthusiasm of small language communities could be an indicator of language health and of speakers' motivation to create a comprehensive encyclopaedia and raise the status of their language. However, this aspect is unclear and might require a more qualitative approach.

As seen before, in general most editors who contribute to a language edition exclusively focus on that language edition. Most of these editors are presumably natives or hold a degree of familiarity with that language and hence share part of the heritage and cultural identity meanings. While this is true in almost all cases, it is also known that languages like English deserve special attention since content is also created by speakers from many other languages (Hale, 2014; Van Dijk, 2009). For the study of cultural identities, it was considered that having a long and varied list of languages and sociological contexts extends its validity. The selection of languages includes the 30 largest Wikipedia language editions in number of articles (as of July 2015), in addition to 10 more language editions which fulfil distinct sociolinguistic factors: language editions from all the continents, different linguistic roots, different speaking community sizes, and also different editing community sizes. The 10 added languages are Basque, Estonian, Greek, Macedonian, Hebrew, Swahili, Afrikaans, Icelandic, Nepali and Guarani.

7.2.2 Dataset Construction: Cultural Identity Related Articles (CIRA)

a) Article Selection and Retrieval

Once languages were selected, it was necessary to map the content of each Wikipedia language edition to cultural identity meanings. The aim was to elaborate a method to collect a comprehensive set of Cultural Identity Related Articles (from now on CIRA) for every Wikipedia language edition. The CIRA from every language are expected to be a set of articles encompassing a wide variety of topics to represent the shared meanings related to the corresponding territories and cultures. Furthermore, more than one cultural identity can co-exist and be shared by the speakers of the same language. With this method, a single CIRA Dataset for every language was created, including all the cultural identities of their speakers. After having taken all this into consideration, before being able to elaborate the method, I still needed a first **ground-truth** with some reliable, certain and central meanings for each language related cultural identities.

[67] https://meta.wikimedia.org/wiki/List_of_Wikipedias
[68] https://en.wikipedia.org/wiki/Censorship_of_Wikipedia#China

In this sense, I identified for each language: the language name, geographical entities (top political territories such as country names), where it is spoken, and its demonyms. A database linking language name, country codes and demonyms was establisehd. To do so, it was necessary to use the ISO code 639 already employed by Wikimedia Foundation to classify Wikipedia language editions (e.g., 'es' for the Spanish language Wikipedia: es.wikipedia.org) and ISO codes 3166 and 3166-2 to identify each country and its subdivisions at a regional level. These codes are widely used on the Internet in geolocation services. This way it was possible to pair each of the selected language editins with its native words to specify the territories where it is spoken (*de iure* or *de facto*), its inhabitants' demonym and language name (e.g., eswiki españa mexico … español castellano) (see Appendix 2 for the complete list). This word list has been generated by crossing ISO databases, and for cases such as a language spoken in a region that does not appear in the database, or a second name for a language, it has been manually revised and extended using information from the specific articles in the correspondent Wikipedia language editions. Once the ground-truth was obtained, a computational implementation of the method was developed applying and integrating the three strategies. The first two gathered the articles considered totally reliable, while the third filtered the undesired ones.

The first criterion and strategy (i) implied examining article location tags such as the coordinates and the ISO code. It was necessary to obtain articles clearly located within such territories. Articles satisfying the first criterion were directly retrieved from the databases of each Wikipedia language edition, which are updated in real time (I was provided access to these articled by the Wikimedia Foundation[69]). Nonetheless, the implementation of coordinates is unequal throughout the different language editions and may contain errors. Therefore, articles with only a couple of coordinates were verified using a *reverse geocoder* tool in Python[70], which provided a ISO code to be verified in the database. Later, it was possible to add articles that were not tagged with coordinates and did not have a territory ISO code, but that could be matched to the corresponding articles in other language editions, where they were properly geolocated (e.g., an article about a city in Nepal which was not geolocated in the Nepali Wikipedia, but it was in the English Wikipedia).

Figure 22 shows the territorial coverage of a sub-selection of 15 languages, where some of the analysis will be carried out. As it can be seen, differences between languages are noteworthy; the second graph in the figure also shows the percentage taken by Geolocated articles in each language edition, and their distribution into countries. Even though the top languages in percentage of geolocated articles are still the big Wikipedias in terms of total number of articles (or in terms of geographical extension) such as the Russian or the English Wikipedias, it is remarkable that less spoken languages, as for instance Catalan and Romanian, had created a large amount of geolocated articles. In Appendix 2 Section 2.2 I examine in depth the characteristics of Geolocated articles.

The second criterion and strategy (ii) implied examining the articles that included in their title keywords related to the language or to the corresponding territories (e.g., "England National football team", "English law", etc.). These two criteria ensured a high reliability, but unfortunately, they could not completely guarantee that all the articled which were supposed to be included in the CIRA selection were actually included.

[69] http://wikitech.wikimedia.org
[70] https://pypi.python.org/pypi/pygeocoder

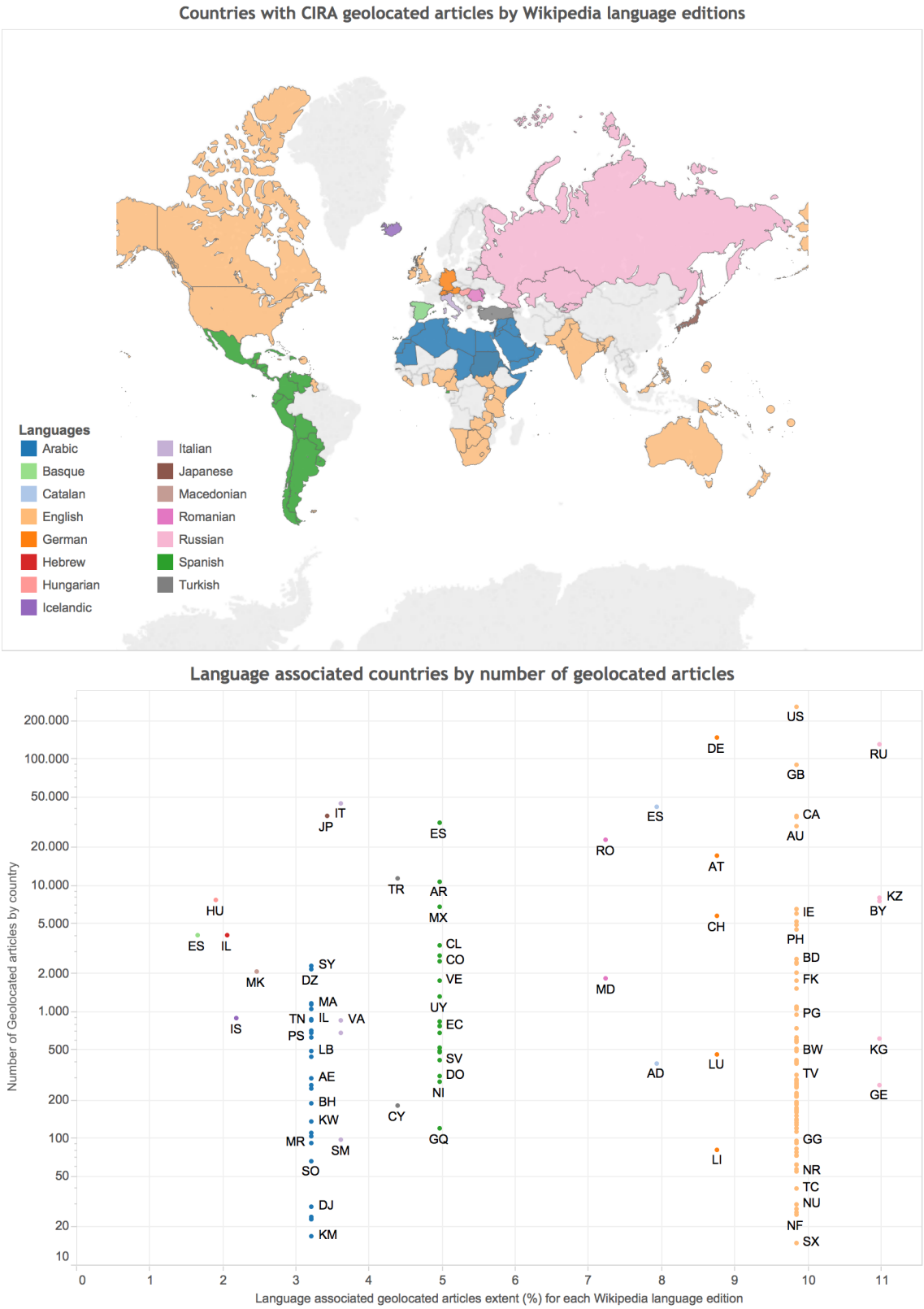


Figure 22. Geolocated articles from each language edition sorted by territories (criterion i). *Top: extension of the territory for Wikipedia language edition. Bottom: the number of geolocated articles for each country and language edition (y), the percentage they occupy in their Wikipedia language edition (x).*

The third criterion and strategy (iii) aimed to retrieve the articles more generally related to particular keywords. Wikipedia articles are classified into categories that are named according to the topics developed in the articles. These categories are organized in a hierarchical tree structure. Hence, starting from a few categories at a general level, it is possible to crawl down the classification structure and gather all the articles on a particular topic. In a similar way to article retrieval (see the second criterion), I used the keywords to retrieve all the categories that include them in their titles; for example: "Performing Arts in England" or "Disputes in English Grammar". These categories contain articles and other categories which contain in turn more specific articles (see Figure 23), until at a certain level, the process of crawling and gathering articles finishes. This depends on the way each editing community constructed the category structure, but it generally happens around the tenth level.

The main advantage of this method is that it allows to obtain articles related to some top-level keywords. However, the distance to the top matters: while the category "Films directed by Charlie Chaplin," is part of the category "Performing Arts in England", its content will be far more specific. The downside of the category crawling is that sometimes the categorization includes circular references or incorrect links (e.g., a more general category appears under a more specific one), which may produce interferences in the final collection (e.g., "World War II" category placed under "Wars involving the United States" category would determine including articles about the German army as related to the English Wikipedia related cultural identities). Possibly because of this interference issue, when I used this method in 2011 with the following keywords: territories, demonyms and language names, I only took into account the first four levels. In the current case, only English language had a limit of five levels of iteration. I let the rest of languages to complete the iterations until the down category graph went extinct. At the same time, I limited the crawling algorithm not to repeat any path.

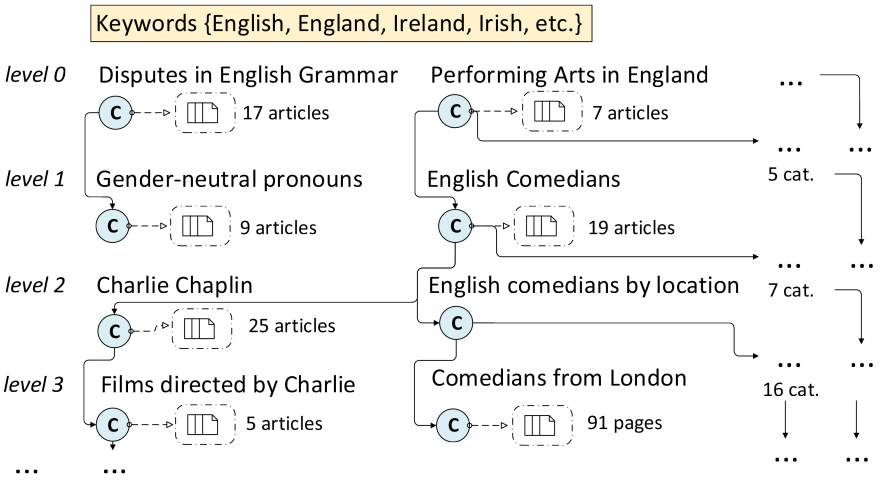


Figure 23. Crawling down the category graph with keywords (strategy and criterion iii).

b) Filtering

Since most of the articles obtained using this third criterion and method can be considered CIRA, the interference issue was tackled with a **filter**. In order to be effective, the filter had to discriminate whether the article was related to the editors' cultural identities, i.e.

whether the links contained in the text directed the reader to other CIRA articles. Fortunately, the geolocated articles and those including the keywords in their title could serve as an initial reliable set of articles and second ground-truth. As a heuristic, when articles from the bulk category crawling selection had a 15% of their text links pointing out to ground-truth articles, they could be added to this group for a further iteration. While the algorithm usually did not add more articles after the third iteration, in large Wikipedia language editions such as English it was necessary to limit the algorithm to the fifth iteration because more articles considered for the new ground-truth had an attracting effect with interference from the bulk. Using this 15% threshold I obtained a definitive CIRA slightly smaller than the bulk selection, but avoided most of the interference.

Table 7 shows the number of articles for each Wikipedia language edition and the percentage of articles obtained through each strategy. Even though the three strategies were used to obtain CIRA (whose extent is provided in the further section), for the sake of the analysis, three separated segments were kept: articles with keywords in title (CIRA Keywords); articles with a geolocation tag (CIRA Geolocated); and other (the rest of CIRA articles with none of these two characteristics).

7.2.3 Manual Assessment

Finally, to check the precision of the method and filter against interference, for each language edition 100 random articles classified as CIRA, and 100 random articles from the remaining ones were retrieved for manual assessment. An automatic translator was used to translate the text of each article. The articles were manually classified according to their content as belonging to CIRA or not. False positives were totally unrelated articles about specific topics from nearby countries, or articles related due to anecdotal relationships, such as a football player who played a competition in one of the countries associated to a language.

In a few cases, articles were considered to be part of CIRA despite not being exclusively focused on the country speaking the corresponding language, but because they were somehow relevant to the country's history or society, and this was reflected in the content of the article. For example, the article about the disputed French region of Lorraine was important to explain the history of Germany, especially during the first decades of the 20th century, when Lorraine used to be part of the German Empire; as a consequence, it is categorized in the German Wikipedia as "Historical Territory (Germany)". This language edition provides references about Lorraine in this historical period, but so does the French Wikipedia language edition. Instead of debating between original or imported concepts, the CIRA selection should be seen as a *continuum* from those more central to a culture - in Hall's words (1990, p. 223), "the common historical experiences and the shared cultural codes" - to those more peripheral but still maintaining an important semantic value to explain a society's imaginary. As it has been tested, deleting periphery is possible by reducing the filter's 15% threshold or adjusting the number of iterations to less than 5.

Table 7. Percentage of articles obtained by selection strategy for each of the 40 Wikipedia editions. *The columns show: the number of articles (WP art.); the percentage of articles identified through geolocated tags in the corresponding territories (GL %); the percentage of articles identified through keywords in their titles (KW %); the total percentage of articles identified through the category crawling (CC %).*

ISO code	Language	WP Art.	GL %	KW %	CC %
af	**Afrikaans**	35966	5.95	0.91	19.53
ar	**Arabic**	375282	3.21	2.44	35.88
eu	**Basque**	208630	1.65	0.42	16.25
ca	**Catalan**	467486	7.93	0.83	18.58
ceb	**Cebuano**	1211531	0.00	0.06	0.06
zh	**Chinese**	851670	6.25	1.17	67.92
cz	**Czech**	326187	9.04	1.15	29.31
da	**Danish**	205764	6.11	1.00	39.56
nl	**Dutch**	1828148	1.64	0.33	9.29
en	**English**	4917741	9.84	2.75	58.62
et	**Estonian**	136362	6.06	1.73	33.51
fi	**Finnish**	375347	2.31	1.03	23.69
fr	**French**	1642276	6.88	1.70	31.25
de	**German**	1834147	8.76	1.85	37.89
el	**Greek**	108090	6.44	0.60	35.97
gn	**Guarani**	3031	13.96	3.27	24.05
he	**Hebrew**	174667	2.06	1.61	34.53
hu	**Hungarian**	326146	1.91	1.45	21.67
is	**Icelandic**	39554	2.19	1.49	32.18
id	**Indonesian**	363529	1.01	0.58	32.76
it	**Italian**	1210801	3.62	0.65	20.50
ja	**Japanese**	973955	3.42	1.01	56.36
ko	**Korean**	320742	2.37	0.83	99.88
mk	**Macedonian**	82743	2.46	1.33	20.47
ms	**Malay**	275031	1.40	0.75	22.08
ne	**Nepali**	29114	11.77	2.16	40.23
no	**Norwegian**	415015	5.51	0.77	29.55
fa	**Persian**	460523	10.33	0.71	30.86
pl	**Polish**	1122218	9.42	1.08	23.91
pt	**Portuguese**	880529	1.99	1.01	24.24
ro	**Romanian**	329925	7.24	1.11	24.11
ru	**Russian**	1237127	10.98	1.14	33.68
sr	**Serbian**	321912	3.22	0.14	13.04
es	**Spanish**	1147742	4.96	1.98	30.33
sw	**Swahili**	29168	3.58	0.99	21.26
sv	**Swedish**	1970808	4.34	0.42	12.31
tr	**Turkish**	249061	4.39	2.06	44.79
uk	**Ukrainian**	581735	6.78	1.01	26.56
vi	**Vietnamese**	1137180	0.88	0.23	4.55
war	**Waray**	1259278	0.00	0.02	0.05
AVG.	***Average***	736654	5.05	1.14	29.53

7.3 The Representation of Cultural Identities in Wikipedia

In this section I present the results of the selection of articles which represents the editors' cultural identities in each Wikipedia language edition, as the outcome of engagement. I measure the extent of the selection and analyse several characteristics of the articles in order to understand the nature of the representation of cultural identity in Wikipedia.

7.3.1 Research Questions

Oyserman's model of an identity-based motivation can be useful to shed light on the process of contributing content to Wikipedia as an identity-congruent act. When choosing the content to contribute with, editors may choose articles imbued with meanings congruent with their cultural identity and therefore the presence, in each language edition, of a considerable number of articles related to local cultural identities may reflect the influence of this motivation. We expect to find a considerable portion of each Wikipedia dedicated to its own related cultural identities. This leads to the following research question:

RQ1. *What is the extent of editors' cultural identity representations in each Wikipedia language edition?*

Cultural identity representations are expected to be covered by a significant number of articles. I believe, first of all, that the creation of this content may play a role in motivating editors to contribute to Wikipedia, and second, that cultural identity is extensible to all editors. Studying the creation of these articles over time can confirm this motivation persists, as well as provide an interesting perspective on its possible evolution in the future. This leads us to the second research question:

RQ2. *How has the content representing cultural identity been created over time?*

Editors of a language edition who lived in the same context acquired, most probably, a great extent of the meanings of that cultural identity. However, I am interested in knowing the topics these cultural identity meanings can be classified into, as they serve to editors to make sense of their world. Looking at the topical coverage of the cultural identities representations may allow to inspect which topics from each language edition are common to all language editions. Likewise, I also expect each language based cultural identities to require diverse topics to represent their context according to their location and their historical background. Therefore, I wonder:

RQ3. *What is the topical coverage of editors' cultural identities representations in each Wikipedia language edition?*

In order to define a cultural identity, some meanings like geographical places or history have a strong importance (Hall, 1990). Therefore, I expect that the structure of content representing cultural identities may involve self-references. Previous research demonstrated that Wikipedia language editions have more links to the articles located in

their geographical domain (Hecht & Gergle, 2010b). This leads to the fourth research question:

RQ4. *Where do links pointing at cultural identity representations come from?*

Cultural identities are also framed in terms of difference and otherness. There exists a relativism among identities, implying that sometimes there is a lack of equivalence from a culture to another, and in order to translate meaning, it is necessary to move from one mind-set to another (Hall, 1990). In Wikipedia, different language editions show a considerable amount of unique, not shared content (Warncke-Wang et al., 2012) which ends up producing a language gap. This is partially explained by the fact that some languages split large topics into more than one article (Hecht & Gergle, 2010b). In addition, I expect the content related to cultural identities to be mainly unique and partly responsible for the language gap. In the last question, I examine the content available across language editions:

RQ5. *What is the availability of content representing editors' cultural identities across different Wikipedia language editions?*

7.3.2 Extent of CIRA in Wikipedia (RQ1)

I start by presenting four prototypical articles from the English Wikipedia***, representing the content types:*** CIRA Keywords***,*** CIRA Geolocated***,*** the rest of CIRA***,*** and the rest of Wikipedia (Figure 24). A good example of CIRA Keywords is 'English Literature' because it explains perfectly this type of article***:*** an article which contains the word 'English' in its title, and whose text is dedicated to summarize a wide topic (English writers' biographies and works). CIRA Keywords articles are often a synthesis of a topic aggregated by the demonym or the territory name. From CIRA Geolocated, I chose the 'Times Square' article, as it represents an article within the geographical territories associated to the English Wikipedia. Even though this is a very iconic place, in CIRA Geolocated there are articles with all levels of notability – from small towns, nation-wide companies and famous monuments. A good example of the rest of CIRA articles is the 'Banbury Cake' article. After the CIRA geolocated and CIRA keywords articles, the rest of CIRA articles dedicated to specific themes of local scope represent the majority in CIRA.

Figure 24. Examples of articles from English Wikipedia. *CIRA Keywords (English Literature), CIRA Geolocated (Times Square), the rest of CIRA (Banbury cake) and the rest of Wikipedia (Sun).*

a) Selection of articles

Results. The Venn diagram shown in Figure 25 presents the average proportion of CIRA in the 40 language editions, and the proportion of these articles that were identified via geolocation tags and keywords in the title. As it can be observed, about 1 over 5 articles in the CIRA set was identified via geo-coordinates, while only about one over 20 was identified via keywords in the title. The intersection between the two subgroups is rather

small. Data for the articles identified via the category hierarchy are not shown, as they represent almost the totality of CIRA (29.5% on average).

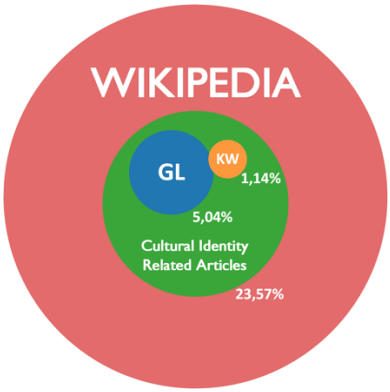


Figure 25. Average proportion of CIRA, and of CIRA detected through geolocation and keywords. Sizes are in scale according to their propotion.

As it can be observed in Table 8, almost a quarter of each Wikipedia language edition (mean 23.2%, median 24.2%, standard deviation 11.1%) belongs to Cultural Identity Related Articles **(RQ1)**. These results show that a non-negligible percentage of each Wikipedia is dedicated to concepts representing editors' cultural identities. This suggests that the influence of an identity-based motivation on editors' article creation is plausible. Table 8 reports the total number of articles and the percentage of articles classified as CIRA at the end of the process for each of the 40 considered language editions. Furthermore, the table shows the percentage of articles that were identified through Criterion 1 (i.e., through keywords in the title) and Criterion 2 (geolocated articles). I omit the percentage of articles selected with Criterion 3 (category crawling), as for most language editions, it is very close or almost equal to the final percentage of articles included in the CIRA set.

Therefore, it is difficult to compare and explain the proportion of CIRA across languages. The English Wikipedia is the biggest in number of articles, and its CIRA set is proportionally one of the largest (46.8% of the articles in the English encyclopaedia are CIRA articles). Only the Japanese Wikipedia has a larger proportion of CIRA (49.2%). For all the other languages, the proportion of CIRA is below 40%. Low proportions of CIRA observed for some languages are due to the presence of automatically translated content. For example, the Vietnamese, Cebuano and Waray-Waray Wikipedia language editions are among the top ten in number of articles but have strikingly low proportions of CIRA; this is because these editions have been mostly grown by an automatic program (bot) which massively created and translated articles from other language editions[71].

These cases are especially interesting because they indicate that CIRA may exist as long as there are editors involved in the community. To further investigate this relationship, the Pearson correlation was computed between CIRA percentage and the total number of editors for each language edition[72]. This showed a correlation of 0.405 ($p=0.013$), which implies that more editors contributed in a language edition, and more articles related to

[71] http://www.wsj.com/articles/for-this-author-10-000-wikipedia-articles-is-a-good-days-work-1405305001

[72] https://meta.wikimedia.org/wiki/List_of_Wikipedias

the corresponding cultural identities were published. This is consistent with the idea that identity-based motivation and cultural identity tend to affect all editors regardless of their activity level.

Table 8. Percentage of CIRA articles in Wikipedia language editions. *For each of the 40 editions, columns show: total number of articles (WP art); percentage of CIRA articles in relation to the entire Wikipedia (CIRA %); percentage of articles identified through geolocated tags in the corresponding territories (GL %); percentage of articles identified through the category hierarchy; percentage of Featured articles among CIRA (FA %); percentage of false positives (FP %); percentage of false negatives (FN %); resulting f1-score (F1).*

ISO code	Language	WP Art.	CIRA %	GL %	KW %	CIRA FA %	FP %	FN %	F1
af	**Afrikaans**	35966	19.20	5.95	0.91	13.75	1	1	1
ar	**Arabic**	375282	26.92	3.21	2.44	42.89	3	12	0.92
eu	**Basque**	208630	10.05	1.65	0.42	36.30	2	0	0.98
ca	**Catalan**	467486	16.17	7.93	0.83	17.91	0	0	1
ceb	**Cebuano**	1211531	0.03	0.00	0.06	0.00	2	0	0.98
zh	**Chinese**	851670	32.87	6.25	1.17	12.43	10	6	0.92
cz	**Czech**	326187	25.97	9.04	1.15	20.13	5	2	0.96
da	**Danish**	205764	31.70	6.11	1.00	30.77	6	5	0.94
nl	**Dutch**	1828148	7.77	1.64	0.33	19.53	1	2	0.98
en	**English**	4917741	46.84	9.84	2.75	75.07	4	12	0.92
et	**Estonian**	136362	31.06	6.06	1.73	50.00	2	5	0.96
fi	**Finnish**	375347	21.95	2.31	1.03	18.34	1	3	0.98
fr	**French**	1642276	29.00	6.88	1.70	32.83	9	5	0.92
de	**German**	1834147	36.77	8.76	1.85	45.53	9	6	0.92
el	**Greek**	108090	33.55	6.44	0.60	33.84	3	3	0.98
gn	**Guarani**	3031	23.59	13.96	3.27	-	0	5	0.98
he	**Hebrew**	174667	31.72	2.06	1.61	40.87	4	4	0.96
hu	**Hungarian**	326146	18.50	1.91	1.45	16.24	2	1	0.98
is	**Icelandic**	39554	30.70	2.19	1.49	20.00	1	2	0.98
id	**Indonesian**	363529	27.02	1.01	0.58	-	3	2	0.98
it	**Italian**	1210801	19.24	3.62	0.65	36.76	1	2	0.98
ja	**Japanese**	973955	49.24	3.42	1.01	38.82	0	9	0.96
ko	**Korean**	320742	32.60	2.37	0.83	23.17	12	7	0.9
mk	**Macedonian**	82743	15.88	2.46	1.33	12.88	5	1	0.96
ms	**Malay**	275031	19.47	1.40	0.75	32.43	1	1	1
ne	**Nepali**	29114	29.69	11.77	2.16	-	1	13	0.92
no	**Norwegian**	415015	26.82	5.51	0.77	24.42	2	1	0.98
fa	**Persian**	460523	11.03	10.33	0.71	6.83	2	13	0.92
pl	**Polish**	1122218	23.15	9.42	1.08	25.86	1	1	1
pt	**Portuguese**	880529	19.05	1.99	1.01	21.58	4	0	0.98
ro	**Romanian**	329925	20.74	7.24	1.11	19.02	3	2	0.98
ru	**Russian**	1237127	31.23	10.98	1.14	29.10	1	1	1
sr	**Serbian**	321912	12.05	3.22	0.14	22.75	2	2	0.98
es	**Spanish**	1147742	27.65	4.96	1.98	30.60	5	1	0.96
sw	**Swahili**	29168	18.30	3.58	0.99	31.84	2	2	0.98
sv	**Swedish**	1970808	11.42	4.34	0.42	13.64	9	2	0.94
tr	**Turkish**	249061	33.90	4.39	2.06	0.00	6	0	0.96
uk	**Ukrainian**	581735	24.84	6.78	1.01	32.20	3	2	0.98
vi	**Vietnamese**	1137180	2.47	0.88	0.23	8.31	2	0	0.98
war	**Waray**	1259278	0.04	0.00	0.02	-	2	0	0.98
AVG.	***Average***	736654	23.25	5.05	1.14	26.02	3.3	3.4	0.96

To inspect the quality of content related to the cultural identity of each Wikipedia, I looked at 'featured articles', a special category of articles that according to editors deserve a mention of quality according to their characteristics[73]. I calculated the proportion of CIRA among featured articles (CIRA FA %) for the 35 languages in our dataset in which this category exists, and I found an average of 27.8% (median 27.5%, standard deviation 13.7%). This proportion is higher than the proportion of CIRA articles, which indicates that editors also engage in creating high quality articles to represent their cultural identities.

b) Manual assessment

Results of the manual assessment of CIRA selection quality are also shown in Table 8, which reports, for each language edition, the percentage of false positives (FP) and false negatives (FN), together with the corresponding F1 score. Overall, I have found that across the 40 languages there were, on average, 3.3% of false positives, and 3.4% of false negatives. The average value of F1 is 0.96. The selections with more interference are Korean and Chinese (12% and 10% FP respectively). This is mainly due to the fact that the category hierarchy of these Wikipedias does not strictly follow a general-to-specific principle, many articles are short and underdeveloped and contain very few links, which makes the 15% threshold ineffective in filtering out anecdotal links. Some improvements were achieved by setting a different threshold in each language edition, but on the other hand I believe that always using the same value for the parameter makes the results more coherent and comparable across languages, with acceptable accuracy levels.

Discussion. The selection of articles representing the editors' cultural identities occupies in average almost a quarter of each language edition. The method proposed integrates different criteria as a groundtruth: CIRA keywords and CIRA geolocated. Later on, I included in the corpus the articles selected through these two characteristics as additional sets of articles, which synthesise content concerning the entire territory or their inhabitants (CIRA Keywords) or describe places, events or people from specific locations (CIRA Geolocated). Yet, the whole CIRA encompasses many other articles. The extent of CIRA in each Wikipedia language edition is slightly correlated with the number of editors with a coefficient of 0.405 (p-value=0.013). This is consistent with the assumption that the creation of this content corresponds to motivated identity-congruent acts, considering that every editor has a cultural identity.

Often, the English language edition in Wikipedia has been considered as a possible referential language, for several reasons: it was the first to be created; it is the widest in number of articles; and importantly it has a status of lingua franca as a global reference with editors from all countries. The most popular language edition editors chose to contribute to (after their native language editions) is English (Hale, 2014). However, far from having a reduced proportion of CIRA, as one could expect from a markedly multicultural encyclopaedia, the English Wikipedia has a 46.8% of articles related to its cultural identities, and it comes second after Japanese.

To obtain Cultural Identity Related Articles, I used CIRA Geolocated articles and CIRA Keywords as a reference to filter undesired content. Manual assessment resulted into a

[73] http://en.wikipedia.org/wiki/Wikipedia:Featured_articles

3.3% of false positives and 3.4% of false negatives. To improve accuracy, thresholds could be adjusted, although the more a Wikipedia language edition improves article characteristics (geolocation tags, outlinks and categorisation), the more reliable the ground-truth will become. Other strategies to diminish interference could be using articles solidly included as CIRA for another language as a negative ground-truth. Machine learning approaches could also be used to improve accuracy. I want to remark that the method I have proposed could also be applied to other kinds of editor identities across languages, such as religion, professional careers, hobbies, gender, etc. This would require finding proper keywords, which may not be straightforward especially in this last case, and setting additional filtering to ensure low interference.

7.3.3 CIRA Creation Over Time (RQ2)

The considerable extent of CIRA shows that editors engage in contributing with content related to their context. One could think that topics about the very near context may be finite or stop being notable, especially in comparison with the amount of universal content which deserves being included in an encyclopaedia. However, I argue that editing CIRA is influenced by an identity-based motivation, and as such it might be sustained over time. Hence, this may prevent that after a long period of time, each Wikipedia becomes diluted into other types of articles. An analysis of how CIRA has been created over time may explain the most productive period, the current influence of this motivation type, and predict future scenarios. To investigate whether the creation of cultural identity related content is consistent over time, I counted the number of articles created every year in each Wikipedia's CIRA since the creation of each Wikipedia language edition until January 2016, and compared it to the overall number of articles created every year in each of the 15 language editions under analysis.

Results. Figure 26 shows the growth of each Wikipedia language edition in terms of number of articles, depicting CIRA as the green area and the rest of the articles as the remaining grey area. Figure 26 also represents the percentage of CIRA created every year (green and red indicate respectively values above and below the overall percentage). In general, CIRA creation tends to remain as a stable part of the activity over the years, although some general patterns can be appreciated.

The most prolific period tends to be located between 2005-2010, when Wikipedia language editions experienced their most important growth. It is the same period when the highest percentages of CIRA for most languages occurred, which suggests that the most important bursts in content creation have been dominated by local cultural identities. After the years of "content boom", the proportion of CIRA tends to get stabilized for most of the languages. Generally, big Wikipedia language editions with strong communities such as the English and the German ones exhibit a more balanced growth, less affected by spikes in the creation of content, as it happens for instance in the Icelandic or the Macedonian Wikipedia.

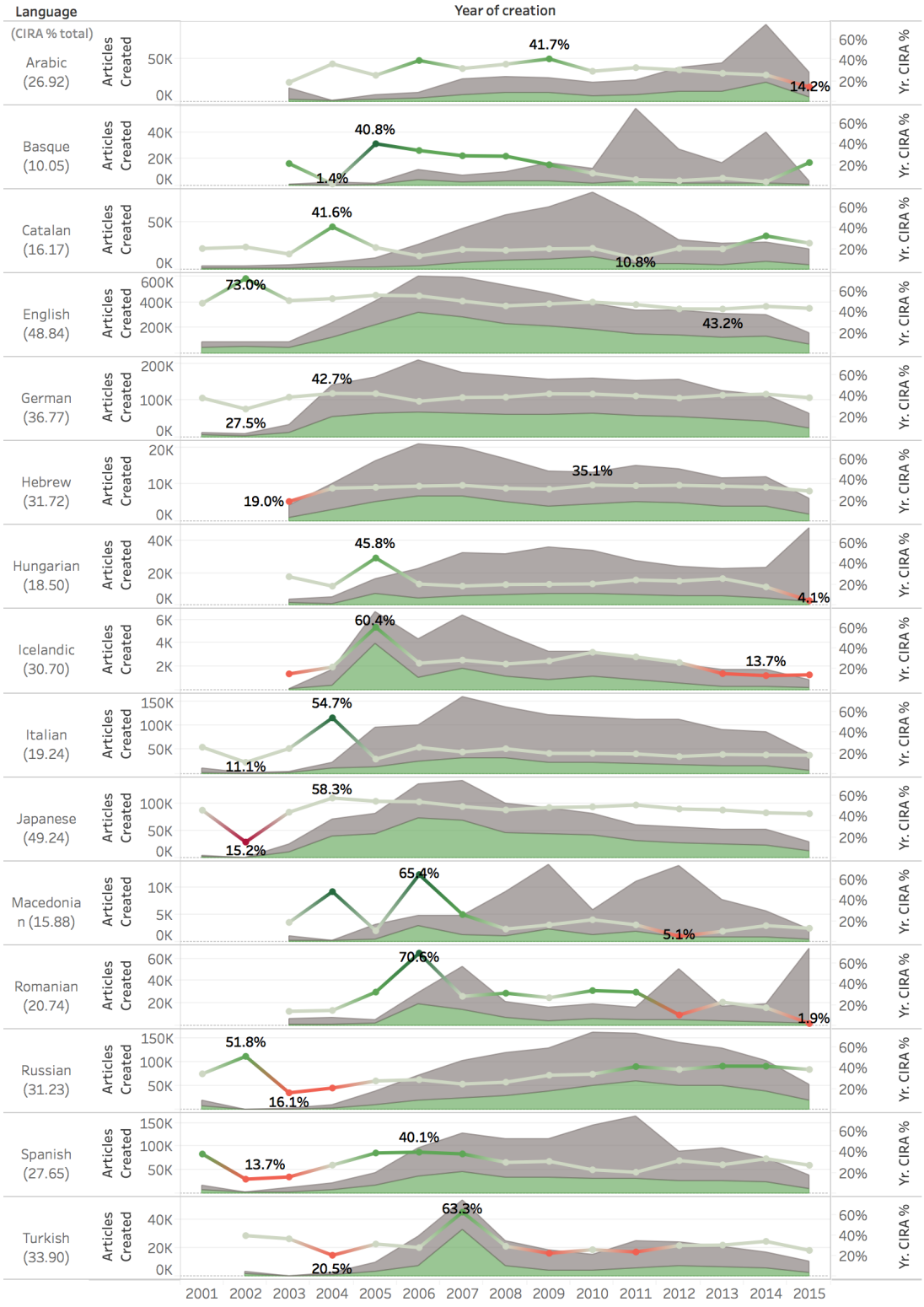


Figure 26. CIRA creation over the 15 years of Wikipedia. *For each language edition, the green area represents the absolute number of CIRA created over years, and the grey area the rest of the articles created. The line shows the percentage of CIRA over the total number of articles created during each year; it is depicted in grey when it is in line (less than 10% variation) with the final overall percentage of CIRA in the encyclopaedia, in green or red when it is higher or lower, respectively.*

Table 9 proposes a similar temporal analysis but with a greater detail into the different types of articles (CIRA, CIRA Geolocated, CIRA Keywords and entire Wikipedia), in order to understand the importance of specific years. Each line represents an article type group, and each column a period of time with the percentage in number of articles created in relation to the final number of articles in the group itself in 2015. Cells are coloured in different shades of green representing the continuum from the smallest to the highest percentage in each language. In the first place, and consistently with the previous figure, most of CIRA, CIRA Geolocated (CIRA GL) and CIRA Keywords (CIRA KW) were created between 2006 and 2010.

Generally, during those years the different language editions grew and found its maximums production peaks, as the percentage of WP also demonstrates. Observing the table, I can assert that the decrease in the creation of articles affects all the different language editions. Some language editions like Arabic or Russian had a special productive year in 2014 for both the entire Wikipedia (WP) and CIRA. An important percentage of the CIRA Geolocated articles were created in one specific year for most of the Wikipedia language editions (usually 2005, 2006 or 2007). This is not surprising, since an important part of the geolocated articles are based on the cities, towns and common places. Considering that editors organize the creation of articles in topics, during those years, this gap was filled. In order to see which segment of CIRA or article group presents a more stable creation, I have computed the standard deviation for each line and averaged it for the total of language editions. That is, CIRA GL shows more variation (8.72), followed by CIRA (5.41), WP (4.86) and CIRA KW (4.06), which confirms this point.

Discussion. The creation of articles from CIRA spreads over years, with a specific period of time during 2006-2010 when most of Wikipedia language editions created more articles in general and also in CIRA. Usually CIRA has grown parallel to Wikipedia, but in those years, it also grew more proportionally, occupying an important percentage of the entire Wikipedia's creation of articles. In those years, most of the CIRA Geolocated articles had been created. This indicates that the core of the content representing Cultural Identities may be finite (e.g. capital cities, historical places, etc.), but not the entire group of CIRA. If we look at the bigger picture, we see that the percentage of articles created in CIRA every year did not show an important decrease.

All in all, these results show that editors are motivated to continually expand the meanings from their cultural identities. Bear in mind that I have analysed the number of articles created, but contributions can be reduced down to each edit. Therefore, perhaps at a certain point the extent of CIRA decreases in percentage of articles, but in percentage of edits it may still attract the same degree of attention. Likewise, other factors which could increase further development of CIRA are the readers' demands of information. Some of these ideas will be explored in Section 8.1. In general, it can be concluded that the sustained interest over time suggests that in the future editors may find more specific topics from their context to make new articles, and probably, some of the most relevant CIRA articles already existing in the encyclopaedia will be developed in greater detail.

Table 9. Different article types creation over the years. *Each period (column) shows the number of articles created in that year divided by the final number of articles in that group. Column 'Total WP' shows the total number of articles in that Wikipedia and the extent (%) of the other article types. Darker shades of green highlight higher values of the percentage.*

Language	Article type	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Total WP
Arabic	CIRA			2.6%	0.5%	1.8%	3.9%	7.7%	9.8%	10.5%	6.1%	7.5%	11.4%	11.5%	22.5%	4.4%	26.92%
	CIRA GL.			3.1%	1.1%	4.9%	7.8%	9.6%	13.4%	14.5%	4.7%	10.3%	12.4%	9.3%	5.7%	3.2%	3.21%
	CIRA KW.			6.1%	0.7%	1.8%	4.1%	8.8%	11.2%	10.0%	6.7%	8.5%	9.9%	12.1%	13.1%	7.0%	2.44%
	WP			4.2%	0.4%	2.2%	2.9%	7.0%	7.8%	7.2%	5.9%	6.6%	10.7%	11.9%	24.1%	9.1%	375282 art.
Basque	CIRA			0.8%	0.1%	2.8%	17.3%	9.3%	13.0%	14.8%	6.6%	14.9%	5.7%	5.1%	6.1%	3.4%	10.05%
	CIRA GL.			0.5%	0.3%	7.8%	17.0%	11.6%	14.9%	18.6%	4.9%	5.2%	3.0%	6.8%	6.5%	3.0%	1.65%
	CIRA KW.			3.5%	0.6%	3.0%	8.8%	12.4%	8.6%	11.3%	10.4%	19.4%	9.8%	5.4%	4.4%	2.5%	0.42%
	WP			0.4%	1.1%	0.8%	5.4%	3.5%	4.9%	8.0%	6.0%	27.7%	13.1%	8.1%	19.3%	1.6%	208630 art.
Catalan	CIRA	0.6%	0.7%	0.6%	3.1%	2.6%	3.8%	9.7%	11.8%	13.8%	17.2%	6.6%	6.8%	6.1%	10.2%	6.4%	16.17%
	CIRA GL.	0.5%	0.6%	0.8%	3.1%	2.0%	2.3%	5.6%	10.2%	17.4%	23.7%	6.9%	7.4%	4.2%	11.1%	4.2%	7.93%
	CIRA KW.	0.7%	1.2%	1.5%	2.4%	5.5%	7.2%	15.1%	13.6%	9.7%	8.2%	5.9%	5.5%	9.5%	7.8%	6.1%	0.83%
	WP	0.6%	0.6%	0.9%	1.4%	2.4%	5.5%	9.3%	12.3%	14.1%	17.4%	12.5%	6.6%	5.8%	6.0%	4.6%	467486 art.
English	CIRA	1.7%	2.3%	1.8%	5.1%	9.5%	13.8%	12.3%	9.9%	9.1%	8.0%	6.5%	5.9%	5.5%	5.7%	2.8%	46.84%
	CIRA GL.	1.4%	7.9%	1.7%	4.9%	10.5%	14.1%	11.1%	10.7%	9.5%	7.8%	5.2%	5.1%	4.2%	4.2%	1.8%	9.84%
	CIRA KW.	2.8%	2.0%	1.9%	5.2%	8.4%	11.9%	12.4%	10.3%	8.8%	7.5%	7.4%	5.9%	5.7%	5.6%	4.1%	2.75%
	WP	1.7%	1.7%	1.7%	4.8%	8.3%	12.2%	12.0%	10.8%	9.5%	8.0%	6.8%	6.8%	6.3%	6.2%	3.2%	4917741 art.
German	CIRA	0.5%	0.3%	1.8%	8.1%	9.4%	10.0%	9.3%	8.9%	9.1%	9.3%	8.5%	8.0%	7.0%	6.4%	3.3%	36.77%
	CIRA GL.	0.5%	0.3%	2.0%	11.3%	10.9%	8.5%	8.1%	7.8%	8.1%	7.9%	8.1%	8.7%	7.4%	7.0%	3.5%	8.76%
	CIRA KW.	0.7%	0.5%	2.6%	8.2%	9.2%	10.8%	8.9%	8.2%	8.2%	9.0%	8.9%	8.8%	6.9%	6.5%	2.7%	1.85%
	WP	0.6%	0.5%	1.8%	7.7%	8.9%	11.4%	9.6%	9.0%	8.5%	8.8%	8.4%	8.5%	6.9%	6.1%	3.4%	1834147 art.
Hebrew	CIRA			1.3%	5.4%	9.0%	12.3%	12.1%	9.4%	7.3%	8.2%	9.1%	8.6%	7.0%	7.0%	3.2%	31.72%
	CIRA GL.			1.5%	12.6%	24.4%	12.8%	11.3%	7.8%	8.1%	4.7%	4.4%	4.2%	3.4%	3.5%	1.4%	2.06%
	CIRA KW.			2.7%	5.9%	7.5%	12.8%	9.6%	11.7%	7.8%	7.8%	9.9%	7.2%	8.3%	6.1%	2.9%	1.61%
	WP			2.4%	5.8%	9.4%	12.0%	11.5%	9.8%	7.8%	7.6%	8.6%	8.0%	6.7%	6.9%	3.5%	174667 art.
Hungarian	CIRA			1.3%	1.2%	10.8%	7.1%	9.3%	9.7%	11.5%	11.2%	10.6%	8.7%	9.1%	6.4%	3.1%	18.50%
	CIRA GL.			1.0%	1.9%	56.0%	6.0%	7.1%	6.6%	4.3%	4.2%	3.2%	3.4%	2.8%	1.9%	1.5%	1.91%
	CIRA KW.			1.4%	1.3%	5.7%	6.1%	9.0%	11.7%	10.8%	7.1%	14.5%	12.8%	8.4%	8.1%	3.1%	1.45%
	WP			1.0%	1.4%	5.1%	6.9%	10.0%	9.7%	11.1%	10.5%	8.5%	7.4%	6.8%	7.1%	14.7%	326146 art.
Icelandic	CIRA			0.1%	2.7%	32.4%	8.2%	14.5%	9.0%	7.2%	9.4%	7.0%	4.6%	2.1%	1.8%	1.0%	30.70%
	CIRA GL.			0.2%	9.4%	16.8%	22.7%	9.6%	15.4%	7.0%	8.6%	4.4%	3.0%	1.2%	1.4%	0.2%	2.19%
	CIRA KW.			0.3%	8.8%	9.5%	12.2%	17.3%	14.6%	9.3%	10.2%	4.2%	5.8%	2.9%	3.7%	1.4%	1.49%
	WP			0.2%	4.4%	16.7%	10.8%	16.1%	11.9%	8.3%	8.3%	6.9%	5.6%	4.4%	4.3%	2.1%	39554 art.
Italian	CIRA	1.1%	0.0%	0.3%	4.6%	5.5%	10.4%	13.2%	13.3%	9.5%	9.3%	8.8%	7.7%	6.9%	6.4%	3.1%	19.24%
	CIRA GL.	0.6%	0.1%	1.0%	18.4%	3.3%	9.0%	15.1%	15.3%	8.8%	6.4%	6.0%	4.7%	5.1%	4.2%	1.9%	3.62%
	CIRA KW.	1.8%	0.1%	0.6%	7.9%	5.2%	9.9%	10.3%	12.1%	9.6%	8.3%	8.4%	7.5%	7.9%	6.6%	3.9%	0.65%
	WP	0.8%	0.0%	0.3%	1.9%	7.9%	8.3%	13.1%	11.4%	10.0%	9.6%	9.3%	9.3%	7.6%	7.1%	3.5%	1210801 art.
Japanese	CIRA	0.4%	0.0%	2.1%	8.0%	9.0%	15.1%	14.5%	9.5%	9.2%	8.4%	6.4%	5.4%	4.9%	4.6%	2.4%	49.24%
	CIRA GL.	0.3%	0.0%	9.8%	26.9%	12.7%	12.7%	9.5%	5.6%	4.2%	4.6%	3.8%	2.8%	2.7%	3.2%	1.3%	3.42%
	CIRA KW.	0.2%	0.0%	2.9%	8.1%	6.5%	14.2%	14.4%	11.2%	10.3%	8.6%	6.8%	5.7%	5.2%	4.1%	1.7%	1.01%
	WP	0.4%	0.0%	2.5%	7.1%	8.2%	13.8%	14.5%	10.2%	9.4%	8.4%	6.3%	5.7%	5.3%	5.3%	2.8%	973955 art.
Macedonian	CIRA			1.1%	0.1%	2.1%	21.5%	9.3%	7.9%	17.0%	8.8%	13.2%	5.0%	5.6%	6.3%	2.2%	15.88%
	CIRA GL.			0.7%	0.0%	3.0%	87.5%	1.4%	1.7%	1.9%	1.3%	1.0%	0.6%	0.4%	0.2%	0.1%	2.46%
	CIRA KW.			4.0%	0.7%	4.8%	13.2%	7.0%	12.0%	7.1%	14.7%	10.2%	9.9%	6.2%	7.9%	2.3%	1.33%
	WP			1.1%	0.0%	3.7%	5.8%	5.8%	10.9%	17.0%	7.0%	13.2%	16.8%	9.3%	6.7%	2.8%	82743 art.
Romanian	CIRA			1.0%	1.2%	2.1%	27.3%	20.6%	9.1%	6.0%	8.5%	6.8%	6.2%	5.1%	4.3%	1.6%	20.74%
	CIRA GL.			0.6%	1.4%	1.2%	60.5%	8.3%	2.5%	1.9%	4.2%	6.3%	7.2%	3.0%	2.3%	0.6%	7.24%
	CIRA KW.			3.8%	2.2%	5.0%	10.5%	12.1%	9.8%	7.4%	12.7%	7.7%	9.3%	5.9%	8.6%	4.9%	1.11%
	WP			1.7%	2.2%	1.5%	9.0%	16.1%	6.4%	4.9%	5.7%	5.0%	15.5%	5.2%	5.8%	21.0%	329925 art.
Russian	CIRA	1.4%	0.0%	0.0%	0.4%	2.4%	4.8%	5.9%	7.2%	10.0%	12.7%	15.4%	12.7%	12.6%	10.0%	4.7%	31.23%
	CIRA GL.	1.4%	0.0%	0.0%	0.4%	1.5%	3.4%	4.2%	6.0%	8.4%	13.4%	17.7%	14.0%	14.7%	9.2%	5.8%	10.98%
	CIRA KW.	2.6%	0.0%	0.1%	0.8%	3.1%	5.3%	7.8%	8.9%	10.1%	12.4%	9.9%	10.7%	13.2%	10.6%	4.5%	1.14%
	WP	1.4%	0.0%	0.1%	0.7%	3.0%	5.7%	8.4%	9.6%	10.4%	13.1%	13.0%	11.5%	10.6%	8.2%	4.3%	1237127 art.
Spanish	CIRA	1.7%	0.1%	0.4%	1.6%	4.8%	10.9%	14.1%	10.2%	10.5%	9.6%	9.7%	8.2%	7.8%	7.4%	3.1%	27.65%
	CIRA GL.	1.1%	0.1%	0.6%	3.3%	8.4%	15.5%	20.7%	9.1%	8.0%	7.6%	7.5%	6.9%	5.3%	4.5%	1.2%	4.96%
	CIRA KW.	3.3%	0.2%	1.0%	2.1%	5.0%	10.3%	9.1%	8.4%	9.2%	9.6%	11.4%	9.3%	9.4%	7.9%	3.8%	1.98%
	WP	1.3%	0.2%	0.9%	1.9%	3.8%	8.3%	11.1%	10.1%	10.0%	12.6%	14.2%	7.7%	8.3%	6.4%	3.3%	1147742 art.
Turkish	CIRA		1.1%	0.0%	0.4%	3.3%	8.6%	38.5%	8.0%	4.5%	4.4%	6.6%	8.1%	7.0%	6.5%	2.9%	33.90%
	CIRA GL.		1.6%	0.0%	1.6%	5.2%	11.6%	55.3%	6.0%	2.2%	3.0%	5.0%	2.8%	2.5%	2.0%	1.2%	4.39%
	CIRA KW.		1.0%	0.0%	0.5%	2.8%	7.6%	11.8%	9.7%	8.6%	6.0%	6.7%	8.8%	11.3%	17.8%	7.3%	2.06%
	WP		1.0%	0.0%	0.8%	3.8%	11.2%	21.4%	9.8%	7.2%	6.1%	9.9%	9.4%	8.1%	6.8%	4.2%	249061 art.

7.3.4 Topical Coverage of CIRA (RQ3)

Once I have measured the extent of CIRA and its evolution during the years, I proceed in analysing the characteristics of this content. I want to see if the Cultural Identities representations comprise all sorts of meanings. For this purpose, I analyse the topical coverage of CIRA to see the different shared meanings necessary to understand the editors' territories, and local contexts from each language edition. In order to do so, I use the method created by Kittur, Chi, & Suh (2009a), which consists of assigning each article's categories to one or more top level categories representing general topics, choosing the closest in the category hierarchy. Hence, it is possible to obtain a distribution of topics for a group of articles. In line with Farina, Tasso, & Laniado (2011), I expanded the top level categories to a total of 18 main categories to cover all the very different encyclopaedic themes, and only analysed the 15 language editions having an equivalent category for each of them.

Results. The result of this process is shown in Figure 27 for the 15 Wikipedias. On average, I find Geography as the biggest category in CIRA (22%), followed by People (19.4%), Culture (14.7%), Society (9.8%), Social Sciences (6.2%), and others. This confirms that content representing cultural identities is diverse in topics as the definition of cultural identity points out. Yet, it is also focused on certain categories **(RQ2)**. By comparing the results for the English Wikipedia with the ones reported by Kittur, Chi, & Suh (2009a), I see that these five categories represented a 82% of the encyclopaedia vs. the 43% they represent in CIRA. The order and proportions in the entire English language edition were also quite different, with Culture (20.2%), People (9.6%), Geography and places (9.5%), Society and Social Sciences (3.6%). Although this change can be due partly to the time passed between the two studies, a strong difference appears between CIRA and the entire encyclopaedia, the first being more distributed into different topics. In fact, the Geography and People categories (whose sum makes 41.4%) are dominant in every language edition's CIRA. This was also expected because of the cultural identity selection criteria.

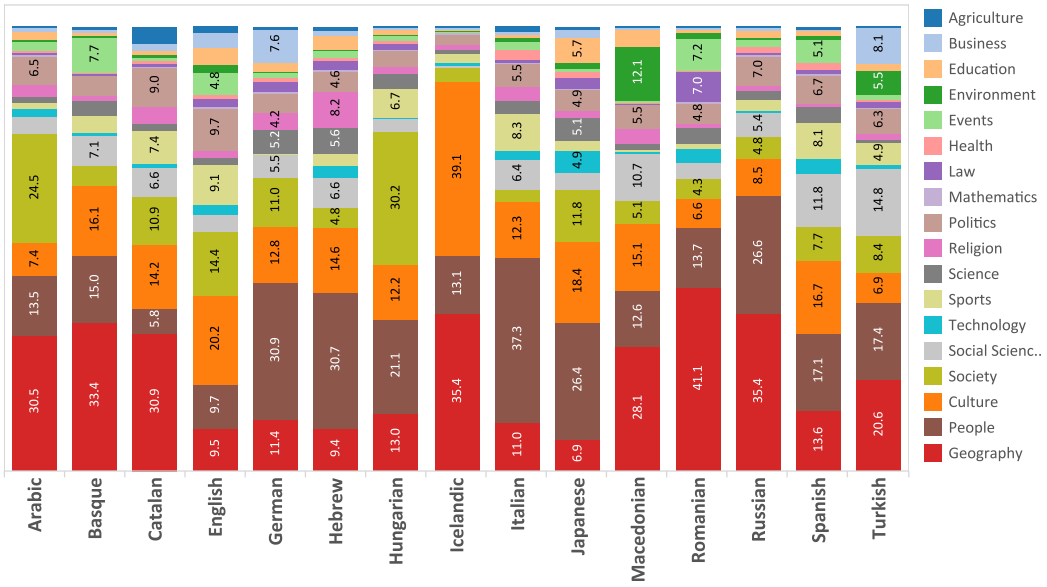


Figure 27. Topical coverage distribution in CIRA.

The cross-cultural comparison of the different CIRA topical coverage shown in Figure 27 allows to see which topics are more represented in each language edition. I note that some patterns seem to confirm common knowledge about cultures. For instance, the Japanese cultural identity appears as the one with the biggest share of articles categorized as technological, while the Hebrew the one with more religion, and the Icelandic has a strong prominence of culture and geography. Across all these data, it is readily apparent that the CIRA from each language edition includes specialized topics as if they were local encyclopaedias placed inside Wikipedia.

Discussion. The application of the method to assign articles to general categories showed that CIRA encompassed all sort of topics, mainly related to people, culture and geography. Like a good representation of the cultural identity, it remains fuzzy in its limits. Because of this, the differences between languages reflect the cultural diversity. The extent of the topics from each Wikipedia language edition reflects the importance they have in the societies editors live in. These results also confirm that the contextualization explored by previous work has effects at a topical level as well. Therefore, it opens the possibility of deepening into the diversity from each Wikipedia language edition.

7.3.5 CIRA Point of View (RQ4)

In previous research, the analysis of the cultural contextualization effects on article text only took into account specific topics. For instance, the same article in different language editions was written giving more prominence to opposite points of view, which is a clear effect caused by the homogeneity of each community. Previous research (Hecht & Gergle, 2010b) considered that each Wikipedia language edition presents constant references to local examples when discussing general topics. In this sense, I propose using CIRA as a specialised dictionary of each Wikipedia language contextual meanings, in order to understand how the CIRA terms are employed by the rest of the Wikipedia, and also by themselves. To do so, I propose a simple calculation by treating CIRA and the rest of Wikipedia as two interconnected meshes and counting their links as references to their articles.

Results. Table 10 shows the percentage of inlinks (incoming links) to CIRA from the same CIRA, the percentage of outlinks from CIRA to CIRA, and the percentage of outlinks to CIRA from the rest of Wikipedia. The first two percentages show the degree of CIRA autoreferentiality. That is, taking into account the 40 Wikipedia language editions, CIRA only directs an average 56.17% of its outlinks to itself (median 57.61% and standard deviation 15.77%), while it receives a 78.87% from itself (median 79.625% and standard deviation 7.16%). This shows that CIRA stands as a structured set of meanings, which tends to be required mostly by itself (and not by the rest of the encyclopaedia), rather than be defined by itself (in some languages the CIRA Outlinks to CIRA% is low, which means that CIRA articles use terms from the rest of the encyclopaedia). Furthermore, the percentage of outlinks from the rest of Wikipedia directed to CIRA are in average a 3.73% (median 2.68% and standard deviation 3.88%). These are low percentages, although some languages like English or German present a 16.11% and 13.79%.

Table 10. Links between CIRA and the rest of Wikipedia. *For each of the 40 Wikipedia editions: Percentage of CIRA inlinks coming from CIRA, Percentage of CIRA Outlinks going to CIRA, Percentage of rest of Wikipedia articles Outlinks going to CIRA.*

ISO code	Language	CIRA Inlinks from CIRA %	CIRA Outlinks to CIRA %	Rest of WP articles Outlinks to CIRA %
af	**Afrikaans**	69.37	57.16	2.56
ar	**Arabic**	83.56	75.99	2.46
eu	**Basque**	76.10	59.03	1.40
ca	**Catalan**	79.56	57.19	0.74
ceb	**Cebuano**	61.32	42.23	0.00
zh	**Chinese**	77.47	37.71	4.38
cs	**Czech**	79.92	40.69	3.25
da	**Danish**	73.86	50.84	8.24
nl	**Dutch**	79.69	58.03	1.55
en	**English**	73.77	68.61	16.11
et	**Estonian**	75.73	43.83	3.70
fi	**Finnish**	82.47	56.62	2.63
fr	**French**	76.49	62.87	7.06
de	**German**	69.48	66.61	13.79
el	**Greek**	84.06	41.10	3.00
gn	**Guarani**	70.15	35.50	3.05
he	**Hebrew**	68.12	55.08	12.14
hu	**Hungarian**	76.16	56.67	2.23
is	**Icelandic**	83.71	54.53	2.73
id	**Indonesian**	85.52	58.47	0.29
it	**Italian**	78.56	60.78	2.93
ja	**Japanese**	80.99	64.01	13.24
ko	**Korean**	74.35	45.98	4.14
mk	**Macedonian**	88.88	66.94	0.74
ms	**Malay**	79.05	17.47	0.98
ne	**Nepali**	94.34	90.67	5.83
no	**Norwegian**	84.54	58.33	2.41
fa	**Persian**	75.89	69.18	3.47
pl	**Polish**	84.39	69.82	3.69
pt	**Portuguese**	86.08	33.25	2.48
ro	**Romanian**	84.67	62.81	0.70
ru	**Russian**	82.36	65.20	4.51
sr	**Serbian**	84.31	62.48	0.36
es	**Spanish**	72.69	47.51	4.48
sw	**Swahili**	82.28	16.86	0.73
sv	**Swedish**	77.07	53.11	1.59
tr	**Turkish**	88.59	66.30	3.37
uk	**Ukrainian**	82.95	93.00	2.26
vi	**Vietnamese**	87.50	71.30	0.17
war	**Waray**	58.94	53.29	0.01
AVG.	***Average***	78.87	56.17	3.73

Discussion. First, by calculating a simple percentage of the incoming and outgoing links I demonstrated that CIRA stands as a set of meanings that are defined among themselves. When CIRA is referenced, it is mostly by the same CIRA, which explains that its nature tends to be isolated and structured. Any CIRA article needs to reference other CIRA to develop its text, but also requires the rest of the Wikipedia topics.

Second, even though the rest of Wikipedia does not address editors' cultural identity meanings, it sometimes employs such meanings to exemplify other topics. However, the percentages are surprisingly lower than expected. More precisely, I would have expected that CIRA articles would be used as examples to illustrate other more universal topics. In some languages, it reaches the 10% although this happens in few cases, and I assume it is because of the contributions of this culture in the universal topics (e.g. German with science) or it could also be a measure of their ethnocentrism view of the world. Likewise, I assume that using the two meshes has some limitations; perhaps an analysis taking into account the article as the measure could inform on when CIRA articles are used as valuable examples in each Wikipedia language edition.

7.3.6 Culture Gap: CIRA Cross-Language Availability (RQ5)

In this section, I examine the cross-language availability of CIRA from the 40-selected languages. In Wikipedia, an article is available in other language editions when it has Interlanguage links (ILL), which can be placed either by an editor of any of the two languages, or by an automatic program (bot). In a way, the bigger encyclopaedias act as leaders and the other editions can copy, translate, and adapt content (Warncke-Wang et al., 2012). An analysis of ILL shows, first and foremost, the degree of uniqueness of the content. Secondly, the analysis shows the relationship between different language editions in integrating one another's specific content, as well as the process of creating content in the overall Wikipedia as a multilingual project. Previous research considered that content available in a language, but not in another, creates a language gap between the two. While this unbalance in content can be due to different reasons, it has been generally assumed that it is due to cultural reasons (Warncke-Wang et al., 2012), although until now, no study has related the exclusiveness to any set of meanings such as CIRA.

a) Interlanguage links analysis

Results. By analysing CIRA interlanguage links I expected to see uniqueness, since cultural identity is defined as shared meanings in a group, but also in terms of *difference* from one another. As seen in Table 11, the average number of ILLs per article is variable across languages, both in CIRA and in the entire Wikipedia. The average in CIRA is 5.4 times lower than in the entire language editions **(RQ-3).** Therefore, CIRA is less shared across languages, and part of the language gap is due to the content representing the cultural identities. Namely, I can affirm that in the language gap there is a **culture gap** (where by culture gap is intended the cultural identity related articles not shared across languages).

Even though in most cases, the average number of ILLs in CIRA is lower than in the entire Wikipedia, the ratio (avg. ILLs CIRA / avg. ILLs WP) is also variable. In fact,

minor language editions like Icelandic, Afrikaans, Estonian and Swahili have between 7 and 11 times less ILLs in CIRA than in the total of their language editions. On the contrary, languages like English, French, Korean, German and Italian show a smaller difference between Wikipedia and CIRA. These latter cases are coincident with some of the biggest Wikipedia language editions, which suggests that both language status and Wikipedia size and development matter also for that language, CIRA being re-created in other languages. In order to further investigate the culture gap in each language edition, I have measured the percentage of articles with no ILLs both in CIRA and in the entire WP, articles that create a gap in relation to all the other language editions. This allows to observe how much the cultural identity representations are responsible for the differences in content imbalance between Wikipedia language editions.

Results show that languages with a high percentage of CIRA also tend to have a high percentage of Wikipedia articles with no Interlanguage Links (WP NO ILLs). CIRA with no ILLs accounts for the majority of Wikipedia content with no ILLs in most of the languages (mean 62.83%, median 63.25% and standard deviation 12.31%, without taking into account the results for languages made by bots like Vietnamese, Waray-Waray and Cebuano). This confirms again that the language gap finds good coincidence with the culture gap. However, this also means that the unique articles in each Wikipedia not belonging to CIRA are unique for other reasons. This may be explained by the way editors structure content or title it, or by the inability to match it with other articles from other languages. For instance, an exception in a mathematical theorem may be considered, by the editors in a certain language community, important enough to deserve an article, while in another language community it is simply added in a new section of an existing article.

b) CIRA cross-language availability

Results. Taking a closer look at CIRA's Interlanguage links, it is possible to obtain a better understanding of the proximity between cultural identities, as well as their coverage or expansion across Wikipedia language editions. To observe this proximity between cultural identities, in Figure 28 I depict a network of languages to show which ones have a higher proportion of articles associated to the cultural identities other languages. More exactly, for each Wikipedia I computed the proportion of articles corresponding to the CIRA of the other languages. Then, for each CIRA, I selected the three languages where it is represented in the highest proportion, and I drew the corresponding edges. Following a standard convention in graph representation, edges are curved and drawn in clockwise direction. Colours are assigned according to the clusters identified by an automatic clustering algorithm (the Louvain method), to highlight groups of language editions that are closer to each other.

Nordic languages form a cluster together with Russian, while Iberic languages are tightly close to each other, as well as Asian languages, and Middle East languages. These results confirm the importance of geographic proximity according to Tobler's Law, which states that things near tend to be similar, and are in line with the results obtained comparing the availability of biographies in different languages (Aragón, Laniado, Kaltenbrunner, & Volkovich, 2012; Eom et al., 2015). However, some less expected relationships also emerge, such as the relevance of Italian CIRA in the Hungarian Wikipedia.

Table 11. CIRA Cross-language coverage. *For each of the 40 Wikipedia editions, columns show: number of article, percentage of CIRA, average number of Interlanguage links per article (ILL WP), average number of Interlanguage links in CIRA (ILL CIRA), percentage of WP articles having no ILLs (No ILL %), percentage of CIRA articles having no ILLs with respect to WP articles having no ILLs (CIRA NO ILL / WP NO ILL).*

ISO Code	Language	WP articles	CIRA %	Avg. ILL WP	Avg. ILL CIRA	WP NO ILL %	CIRA NO ILL %	CIRA NO ILL / WP NO ILL
af	**Afrikaans**	35849	19.199	40.12	4.45	8.68	34.32	76.2
ar	**Arabic**	375282	26.921	12.89	3.55	29.39	59.50	54.5
eu	**Basque**	208631	10.046	14.40	1.28	6.96	50.64	73.09
ca	**Catalan**	467460	16.17	21.52	3.63	17.72	68.71	62.69
ceb	**Cebuano**	1211521	0.03	14.98	1.56	0.44	0.55	0.04
zh	**Chinese**	830671	32.865	6.32	10.89	30.96	58.22	63.36
cs	**Czech**	325342	25.973	4.81	8.85	22.12	60.26	70.95
da	**Danish**	205764	31.696	10.00	2.58	22.61	52.34	73.37
nl	**Dutch**	1828093	7.766	12.96	1.82	22.29	64.43	22.44
en	**English**	4917332	46.838	6.81	1.46	40.77	54.95	63.14
et	**Estonian**	136054	31.055	20.16	1.83	28.71	64.42	69.84
fi	**Finnish**	375348	21.948	6.04	2.92	22.01	70.20	69.99
fr	**French**	1642175	29.004	23.20	4.74	22.63	46.24	59.27
de	**German**	1834107	36.771	15.01	2.48	35.34	60.08	62.50
el	**Greek**	107824	33.548	17.93	4.15	21.58	46.14	71.89
gn	**Guarani**	3032	23.59	82.07	24.18	2.80	6.85	57.64
he	**Hebrew**	174147	31.722	20.02	4.79	20.11	50.27	79.53
hu	**Hungarian**	325234	18.496	16.04	2.92	16.93	54.81	60.05
is	**Icelandic**	39534	30.702	11.97	1.66	27.38	66.11	74.15
id	**Indonesian**	363523	27.015	33.74	2.39	13.45	36.74	73.82
it	**Italian**	1210753	19.242	9.31	3.48	21.70	54.54	48.38
ja	**Japanese**	973935	49.237	7.05	1.15	48.24	75.88	77.45
ko	**Korean**	320742	32.602	14.14	7.76	30.23	69.86	58.64
mk	**Macedonian**	82684	15.876	25.32	3.34	12.61	40.97	51.60
ms	**Malay**	274882	19.469	15.52	1.81	17.54	55.36	61.46
ne	**Nepali**	28948	29.694	22.02	3.30	41.89	41.39	29.50
no	**Norwegian**	415006	26.817	12.40	2.26	20.09	54.29	72.47
fa	**Persian**	460505	11.031	7.64	4.83	19.12	8.16	4.71
pl	**Polish**	1122180	23.152	9.35	1.29	24.86	58.14	54.15
pt	**Portuguese**	878629	19.05	11.23	2.43	21.08	64.53	58.44
ro	**Romanian**	326904	20.744	16.89	3.45	9.37	32.57	72.74
ru	**Russian**	1232891	31.232	8.25	2.20	24.64	44.60	56.72
sr	**Serbian**	321604	12.051	16.04	4.72	5.98	24.89	50.25
es	**Spanish**	1147690	27.652	9,32	3,37	18.89	44.34	64.90
sw	**Swahili**	29153	18.297	39,97	3,67	11.60	46.77	73.82
sv	**Swedish**	1969513	11.415	5,98	1,45	12.55	72.49	65.99
tr	**Turkish**	249049	33.897	16,21	3,38	17.62	36.36	69.94
uk	**Ukrainian**	581695	24.838	12,88	2,41	18.78	43.08	56.98
vi	**Vietnamese**	1137180	2.466	7,36	1,45	10.62	72.76	16.9
war	**Waray**	1259262	0.037	6.32	10.89	0.24	12.63	1.94
AVG.	***Average***	735753.2	21783.39	16.6	4.02	20.01	48.98	57.14

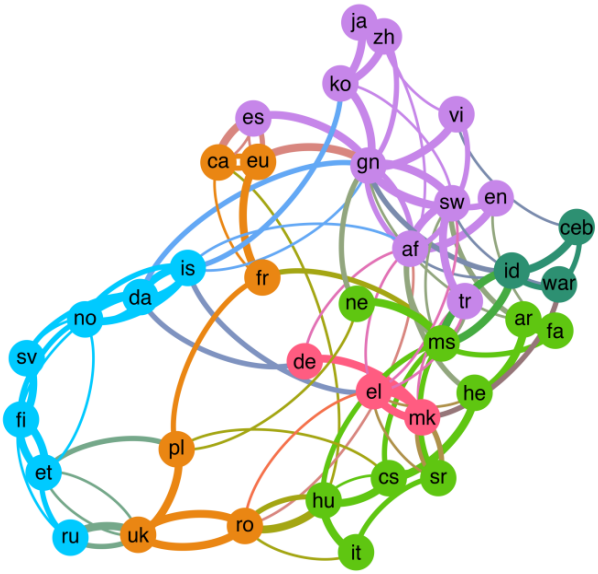


Figure 28. Network graph with CIRA across languages. Colours represent the proximity between languages in number of shared articles.

c) Mapping the culture gap

Results. To see how well each Wikipedia language edition covers other language CIRA, I have created Table 12, Table 13, Figure 29 and Figure 30.

Table 12 shows *the coverage of languages' CIRA by the other Wikipedia language editions* (i.e. the percentage of a language CIRA – column – covered by a Wikipedia language edition – row). Hence, the entire table allows to see the culture gap of each language edition, and how this also depends on the linguistic and geographical proximities. However, it seems the factor of scale is more important, since wide language editions (in number of articles and created by large communities such as English, German, French, etc.) cover a higher percentage of the other CIRA.

Table 13 shows *the extent of one CIRA in the other Wikipedia language editions* (i.e. the percentage a CIRA – column – occupies in terms of articles in other Wikipedia language editions – row). This allows to see the impact or spread of some cultural identities in other languages. English CIRA is by far the most expanded in the other Wikipedia language editions, followed by the German and French. At another level, there is Spanish, Russian, Italian, Japanese, and Chinese. However, it is interesting to note that some large language editions like Dutch, which has a larger number of articles than Arabic, do not even occupy a 1% in most of the other language editions. In particular, Dutch only occupies a greater extent of articles in Afrikaans Wikipedia, which is the edition of a language spoken in South Africa that evolved from Dutch. In turn, Arabic is perhaps the language with the strongest demography (420 million in 2016[74]) but its CIRA only occupies in average 1.35% of the articles in the other language editions (median 1.21%, standard deviation 0.88%). Therefore, Table 13 depicts a portrait of the relevance of cultural identities in Wikipedia, but not of their relevance in the world.

---

[74] http://www.unesco.org/new/en/unesco/events/prizes-and-celebrations/celebrations/international-days/world-arabic-language-day/

Table 12. Culture gap: 40 Wikipedia language editions coverage (% articles) of 40 Wikipedia language editions CIRA. *Each row shows the coverage of each Wikipedia language editions' CIRA. The coverage is calculated as the number of articles in a Wikipedia language edition (row) which belong to a Wikipedia language edition CIRA (column) divided by the total number of articles in the Wikipedia language edition CIRA (column). For an easy identification of values, cells are coloured in red to indicate a percentage lower than 1%, and in green in a continuum until 93.67% (the highest value).*

Wikipedia Language Edition CIRA

Lang	af	ar	ca	ceb	cs	da	de	el	en	es	et	eu	fa	fi	fr	gn	he	hu	id	is	it	ja	ko	mk	ms	ne	nl	no	pl	pt	ro	ru	sr	sv	sw	tr	uk	vi	war	zh
af		1.05	0.11	0.63	0.21	1.57	0.67	1.34	0.62	0.41	0.22	0.75	0.11	0.17	0.53	13.99	1.49	0.24	0.21	0.53	0.41	0.22	0.55	0.52	0.46	0.52	0.98	0.29	0.14	0.40	0.46	0.36	0.43	0.23	1.37	0.71	0.40	0.37	1.71	0.62
ar	4.13		2.43	6.96	1.60	5.57	3.46	7.99	4.21	5.17	0.90	4.94	30.45	1.47	4.34	18.74	12.15	2.57	1.31	1.37	4.05	1.61	2.90	4.64	2.17	1.89	1.90	1.21	0.49	2.56	3.48	1.91	3.20	1.33	3.32	6.57	1.88	1.60	4.71	3.29
ca	5.89	7.57		16.46	2.33	7.00	4.88	13.46	4.26	18.10	1.47	19.08	1.04	1.25	15.28	31.61	8.65	3.53	0.94	2.96	7.06	1.31	2.40	3.30	1.33	2.13	3.14	1.97	0.88	3.96	3.96	2.42	4.06	1.75	5.32	5.33	2.30	1.33	8.78	3.43
ceb	0.43	0.80	0.39		0.09	0.28	0.59	0.19	1.86	2.14	2.97	1.65	0.62	0.16	8.81	9.65	1.40	0.38	1.19	0.12	0.22	0.19	0.34	2.77	0.74	0.82	0.26	0.76	0.09	3.71	0.33	0.30	3.06	0.36	0.81	0.46	0.60	0.81	78.59	4.45
cs	3.77	5.14	1.08	3.80		6.84	5.22	6.75	3.13	2.85	2.72	3.29	0.73	2.04	5.46	17.34	14.57	6.62	0.60	3.95	3.75	1.18	2.60	2.91	1.18	1.75	1.97	2.11	2.45	2.05	4.27	3.23	6.39	2.09	2.94	3.70	3.39	1.51	1.50	2.97
da	4.58	3.20	0.93	0.63	1.53		3.90	3.90	2.22	2.17	1.69	2.62	0.51	1.66	2.69	16.92	5.51	2.41	0.46	5.34	2.05	0.80	1.99	1.33	0.94	2.36	1.58	6.53	0.68	1.64	2.18	1.29	2.08	4.43	4.24	2.36	1.40	1.02	0.86	2.05
de	28.20	13.13	5.60	58.65	21.21	22.29		22.10	14.79	14.17	12.10	9.91	2.47	7.62	22.87	45.32	21.73	14.11	2.63	17.09	18.96	4.43	6.11	8.18	4.00	7.75	11.32	9.46	8.81	12.26	13.43	9.13	14.39	10.04	13.88	16.29	8.28	4.01	56.10	9.43
el	1.46	2.88	0.57	0.84	0.82	2.93	1.64		1.05	1.33	0.49	1.73	0.34	0.45	1.53	9.51	3.96	1.18	0.23	1.31	1.78	0.31	1.01	8.89	0.54	1.13	0.76	0.52	0.27	0.97	2.28	0.83	3.92	0.58	2.77	3.68	0.89	0.48	1.50	0.98
en	62.36	33.63	12.69	67.30	24.34	29.75	28.27	42.22		38.12	22.12	22.34	91.36	18.45	40.23	86.29	41.85	24.94	10.00	22.95	30.58	16.86	17.75	25.69	15.40	44.74	22.30	28.27	31.79	25.26	39.53	17.58	49.62	18.87	50.63	31.62	14.40	20.31	68.52	25.65
es	14.27	11.30	26.04	28.27	4.50	11.12	9.56	18.50	12.62		2.94	39.36	1.66	3.26	19.69	91.05	15.37	6.98	2.38	10.77	14.38	4.28	5.58	6.04	3.43	4.06	5.39	4.47	2.38	16.12	7.45	5.34	10.02	4.08	9.35	7.64	5.11	3.79	20.56	7.30
et	2.09	1.90	0.30	1.06	1.06	4.52	1.95	3.54	1.08	0.99		1.78	0.45	3.02	1.43	16.92	3.13	1.53	0.35	1.88	1.67	0.42	1.39	1.06	0.74	0.80	0.68	1.86	0.62	0.88	1.65	2.49	1.61	1.74	2.42	1.84	2.11	0.66	2.14	1.48
eu	2.95	2.96	4.58	16.88	0.74	3.34	1.67	4.10	1.63	8.47	0.68		0.46	0.84	10.81	17.20	3.89	1.34	0.83	2.38	1.75	0.56	1.19	1.07	0.75	2.02	0.93	0.74	0.24	1.60	1.84	0.70	1.30	0.67	2.17	1.85	0.97	0.83	4.50	2.46
fa	4.87	10.32	1.79	11.39	1.71	5.76	4.12	6.29	6.28	4.37	2.49	3.01		0.97	4.63	26.29	9.04	3.43	1.13	1.57	3.72	1.95	3.15	2.09	2.00	2.82	1.48	1.59	1.02	3.62	3.04	2.38	3.81	1.57	6.28	6.93	2.06	2.18	8.57	3.88
fi	7.71	5.42	1.52	3.38	3.14	8.25	4.58	12.19	4.58	3.84	5.50	4.21	0.83		4.98	27.55	8.49	4.50	1.01	5.35	4.23	1.67	2.77	3.02	1.88	3.53	2.18	4.61	1.13	3.19	3.47	4.39	4.11	8.31	6.54	4.05	2.76	1.89	3.85	3.62
fr	24.92	20.66	10.19	41.35	13.10	15.04	17.07	22.79	16.26	20.65	6.83	24.01	2.66	8.49		56.64	21.30	18.16	3.30	12.67	22.53	7.78	7.42	21.85	4.14	15.92	14.65	7.90	12.42	14.88	12.61	8.43	41.58	7.46	14.92	12.05	6.52	6.10	28.69	10.84
gn	0.12	0.09	0.02	0.21	0.01	0.16	0.04	0.03	0.04	0.24	0.02	0.19	0.00	0.01	0.03		0.10	0.02	0.02	0.04	0.02	0.01	0.09	0.04	0.03	0.12	0.01	0.02	0.01	0.27	0.00	0.02	0.04	0.01	0.26	0.06	0.03	0.05	0.00	0.07
he	3.84	7.21	0.80	2.11	1.60	5.04	3.09	5.67	2.43	2.14	0.86	2.23	0.63	0.69	2.43	16.50		2.74	0.43	1.26	2.27	0.81	2.04	2.23	0.96	1.34	1.06	1.01	0.97	1.59	3.44	1.88	2.88	1.04	3.00	3.41	2.10	0.93	0.86	2.03
hu	3.35	4.10	3.72	4.01	11.02	5.75	5.58	6.44	2.69	6.15	1.53	6.69	0.51	1.53	11.65	15.39	5.89		0.62	4.12	8.28	1.05	2.32	4.02	1.00	1.46	2.01	1.87	1.25	2.51	17.03	2.40	9.69	1.72	2.68	3.84	2.88	0.91	1.71	2.57
id	4.24	5.61	0.79	53.38	1.16	4.45	2.60	4.74	2.51	2.29	0.58	2.43	0.87	0.69	4.83	19.16	6.48	2.02		1.65	3.47	2.04	3.80	1.61	35.96	2.08	2.24	0.92	0.45	2.13	2.41	1.13	2.37	0.95	3.37	3.22	1.04	2.60	53.96	4.10
is	0.65	0.76	0.09	0.00	0.24	2.71	0.63	1.54	0.42	0.31	0.30	0.86	0.07	0.24	0.44	8.81	1.39	0.71	0.14		0.42	0.16	0.60	0.25	0.33	0.28	0.24	0.94	0.11	0.37	0.49	0.20	0.43	0.63	0.90	0.58	0.34	0.31	0.86	0.48
it	17.03	11.98	7.26	58.86	14.27	12.23	13.66	23.17	11.86	16.83	4.27	12.68	3.78	4.86	22.95	66.57	15.13	14.49	1.52	9.56		5.24	5.14	7.19	2.65	5.18	6.54	6.52	4.02	11.96	15.00	6.97	12.78	5.35	10.01	9.73	4.87	3.44	54.60	6.35
ja	7.26	6.66	1.73	6.33	3.59	7.71	6.14	8.71	6.80	5.21	1.59	6.05	1.05	2.10	6.35	24.62	9.99	4.76	1.64	4.08	8.35		12.68	2.99	3.63	4.20	2.56	2.29	1.23	3.93	4.16	3.77	5.06	2.37	8.39	4.89	3.10	4.76	9.85	16.28
ko	4.66	4.73	0.71	15.61	1.51	5.36	2.93	6.30	3.15	2.86	0.90	2.41	0.74	0.95	2.92	26.01	6.01	2.47	0.98	2.55	3.12	7.10		2.70	2.57	1.96	1.20	1.14	0.53	2.33	2.63	1.82	2.71	1.21	6.46	3.49	1.51	3.29	8.78	9.45
mk	1.12	1.48	0.17	2.95	0.43	2.23	1.67	5.02	0.56	0.93	0.32	1.26	0.29	0.22	0.69	15.94	2.04	0.86	0.17	0.96	0.97	0.17	0.59		0.50	0.54	0.29	0.35	0.18	0.45	1.70	0.74	4.79	0.36	1.31	2.15	0.86	0.34	5.78	0.71
ms	1.78	4.14	1.92	2.11	7.89	2.17	2.95	1.02	1.56	3.77	0.34	4.37	18.92	0.30	9.14	11.05	2.81	6.70	26.07	0.54	4.12	1.08	1.59	0.88		30.32	0.75	2.18	0.64	1.40	5.57	0.50	12.87	0.41	1.54	9.47	0.56	1.41	5.14	2.06
ne	0.46	0.25	0.02	0.00	0.03	0.43	0.17	0.09	0.27	0.09	0.09	0.12	0.09	0.03	0.10	2.24	0.38	0.09	0.12	0.06	0.05	0.06	0.28	0.08	0.15		0.04	0.07	0.01	0.11	0.08	0.05	0.04	0.06	0.32	0.17	0.10	0.14	0.00	0.40
nl	27.27	8.16	5.55	90.72	11.98	17.62	12.28	12.48	9.01	11.68	5.97	10.18	1.35	3.34	21.07	52.17	11.91	11.45	33.86	8.28	11.37	2.55	3.91	4.77	10.96	5.00		6.13	6.79	12.34	10.81	3.91	12.84	6.35	10.72	7.83	3.68	7.82	60.39	6.66
no	8.95	5.47	1.43	18.14	2.73	18.66	5.27	7.47	4.54	4.48	3.26	3.69	0.89	4.52	4.58	41.96	8.30	3.83	0.84	11.84	4.49	1.35	3.37	2.63	1.85	3.61	2.79		1.23	5.92	3.08	2.88	4.10	8.20	9.07	4.06	2.28	2.18	11.78	5.66
pl	17.57	11.02	5.96	9.49	16.73	12.46	13.03	16.04	9.76	12.10	14.27	10.38	2.36	5.07	17.90	55.11	17.66	12.99	1.75	8.72	13.28	2.92	4.52	9.76	3.01	34.85	5.88	5.81		8.97	9.23	10.28	12.24	5.61	11.41	7.82	15.26	3.23	9.42	6.02
pt	9.72	8.90	4.97	15.82	10.42	9.87	8.73	13.92	9.07	16.59	2.86	10.37	1.39	2.70	14.82	72.45	12.20	5.04	1.38	5.23	10.79	3.36	4.17	4.09	2.29	4.86	4.00	3.54	2.73		13.71	3.34	16.60	3.78	6.73	5.94	2.87	2.72	6.85	5.16
ro	11.75	5.61	1.93	2.11	1.97	5.53	4.83	6.75	2.07	3.13	1.52	3.32	0.89	1.28	10.37	43.64	5.76	11.01	0.44	1.80	5.89	0.95	1.63	4.53	0.87	0.97	1.66	2.31	1.64	4.80		2.85	5.30	1.57	7.95	3.25	23.21	0.91	2.36	1.76
ru	15.09	11.83	5.01	14.56	8.89	11.42	13.25	20.31	9.23	12.76	9.90	9.96	3.42	7.19	14.19	52.87	19.27	9.04	2.10	8.42	12.51	4.15	5.57	11.23	3.25	5.70	5.27	5.10	8.23	12.43	12.73		18.75	5.35	10.17	10.56	46.61	4.53	8.57	8.46
sr	3.22	3.47	0.64	4.85	6.47	4.20	3.79	8.67	2.89	1.85	0.76	2.37	0.50	0.77	4.52	21.54	4.63	4.20	0.35	1.62	5.93	0.52	1.35	22.98	0.72	0.96	0.69	0.74	1.01	1.15	3.73	2.06		0.89	3.09	3.25	1.89	1.12	2.78	1.68
sv	9.56	6.87	1.95	91.56	4.28	20.62	8.64	9.18	7.97	6.60	4.59	5.58	1.44	12.87	15.06	35.94	10.33	6.88	1.48	12.39	5.57	2.06	3.75	6.20	2.01	2.70	3.89	10.91	1.88	6.65	4.79	3.97	9.55		11.58	5.11	3.45	3.58	70.88	9.07
sw	2.95	1.03	0.06	0.21	0.09	0.65	0.35	0.35	0.49	0.23	0.04	0.55	0.07	0.06	0.60	12.45	1.06	0.21	0.12	0.21	0.37	0.10	0.46	0.15	0.33	0.28	0.13	0.11	0.03	0.27	0.30	0.09	0.25	0.16		1.51	0.11	0.46	1.71	0.36
tr	3.88	6.35	0.91	2.53	1.57	4.88	3.24	8.14	2.42	3.19	1.12	3.05	4.42	1.13	2.81	18.04	7.38	2.85	0.57	1.79	2.62	1.07	2.09	4.22	1.21	1.08	1.12	1.12	0.56	3.34	3.05	2.03	5.12	1.31	2.87		1.66	1.11	5.78	2.39
uk	6.11	6.28	3.00	6.96	4.07	6.77	5.33	12.11	3.65	5.96	5.23	6.73	1.11	1.96	11.94	27.27	8.51	4.59	0.98	3.56	7.51	2.00	2.83	9.27	1.74	2.29	1.86	1.92	5.37	3.08	32.02	15.42	6.82	2.54	3.47	5.10		1.75	3.43	3.80
vi	2.62	4.40	1.99	93.67	8.48	3.65	3.74	3.25	4.02	5.99	1.07	4.89	0.93	0.96	10.32	37.34	3.96	6.70	1.57	0.97	4.28	1.84	4.23	1.52	2.55	3.16	0.97	1.31	0.86	5.35	5.07	1.91	1.73	1.12	5.23	35.56	2.34		58.03	10.98
war	1.23	1.19	2.58	93.55	0.41	1.19	2.92	1.08	1.64	5.02	0.77	4.49	0.48	0.26	9.01	43.08	1.65	0.81	0.87	0.26	3.94	0.44	0.68	0.61	0.58	1.15	0.51	0.79	1.32	4.73	0.53	0.68	0.52	0.45	1.18	0.89	0.99	1.46		5.00
zh	9.86	6.71	3.70	7.38	2.91	7.37	6.85	8.09	6.42	9.31	1.71	6.74	1.72	1.88	11.35	60.00	8.18	8.55	2.13	3.93	7.24	11.05	11.96	4.08	5.33	4.22	1.94	2.32	1.31	7.35	4.03	3.60	5.06	2.01	9.86	5.17	4.76	9.48	9.64	

Table 13. Culture spread: 40 Wikipedia Language editions CIRA extent (% articles) in 40 Wikipedia language editions. *Each row shows the extent of each Wikipedia languages' CIRA. The extent is calculated as the number of articles from a Wikipedia language CIRA (column) which are available in a Wikipedia language edition (row) divided by the total number of articles in the Wikipedia language edition (row). For an easy identification of values, cells are coloured in brown to indicate a percentage lower than 1%, and in green in a continuum until 37.13% (the highest value).*

Wikipedia Language Edition CIRA

Lang	af	ar	ca	ceb	cs	da	de	el	en	es	et	eu	fa	fi	fr	gn	he	hu	id	is	it	ja	ko	mk	ms	ne	nl	no	pl	pt	ro	ru	sr	sv	sw	tr	uk	vi	war	zh
af		2.77	0.22	0.01	0.47	2.68	11.75	1.27	37.13	3.43	0.25	0.41	0.15	0.37	6.64	0.26	2.16	0.39	0.53	0.17	2.48	2.76	1.44	0.18	0.65	0.12	3.46	0.84	0.98	1.67	0.79	3.61	0.43	1.37	0.19	1.58	1.53	0.28	0.02	4.54
ar	0.07		0.45	0.01	0.33	0.90	5.77	0.72	23.96	4.06	0.09	0.26	3.83	0.30	5.11	0.03	1.67	0.38	0.32	0.04	2.33	1.91	0.72	0.15	0.29	0.04	0.64	0.33	0.32	1.02	0.57	1.83	0.31	0.74	0.04	1.37	0.67	0.11	0.01	2.28
ca	0.08	1.55		0.02	0.40	0.93	6.68	0.99	19.92	11.67	0.13	0.81	0.11	0.21	14.79	0.05	0.97	0.43	0.19	0.07	3.34	1.28	0.50	0.09	0.14	0.04	0.88	0.45	0.46	1.31	0.53	1.90	0.32	0.80	0.06	0.91	0.67	0.08	0.01	1.95
ceb	0.00	0.05	0.02		0.00	0.01	0.22	0.00	2.36	0.37	0.07	0.02	0.02	0.01	2.32	0.00	0.04	0.01	0.06	0.00	0.03	0.05	0.02	0.02	0.02	0.00	0.02	0.05	0.01	0.27	0.01	0.07	0.07	0.05	0.00	0.02	0.05	0.01	0.02	0.69
cs	0.08	1.52	0.24	0.01		1.30	10.28	0.71	21.06	2.64	0.34	0.20	0.11	0.49	7.59	0.04	2.36	1.17	0.17	0.14	2.55	1.66	0.77	0.11	0.18	0.04	0.79	0.69	1.86	0.97	0.83	3.64	0.72	1.38	0.05	0.91	1.43	0.12	0.00	2.43
da	0.15	1.52	0.33	0.00	0.61		12.34	0.66	23.97	3.23	0.34	0.26	0.12	0.64	6.00	0.06	1.43	0.68	0.21	0.30	2.24	1.81	0.96	0.08	0.24	0.10	1.03	3.40	0.83	1.27	0.69	2.34	0.38	4.67	0.11	0.93	0.95	0.13	0.00	2.69
de	0.10	0.70	0.22	0.02	0.95	0.77		0.42	17.96	2.37	0.27	0.11	0.07	0.33	5.75	0.02	0.64	0.45	0.14	0.11	2.33	1.12	0.33	0.06	0.11	0.04	0.83	0.56	1.21	1.06	0.47	1.86	0.29	1.19	0.04	0.73	0.63	0.06	0.01	1.39
el	0.09	2.57	0.38	0.00	0.61	1.68	9.77		21.32	3.73	0.18	0.32	0.15	0.33	6.44	0.06	1.93	0.63	0.20	0.14	3.66	1.31	0.90	1.03	0.25	0.09	0.92	0.51	0.61	1.38	1.33	2.84	1.34	1.15	0.13	2.74	1.13	0.12	0.01	2.43
en	0.09	0.67	0.19	0.01	0.41	0.38	3.77	0.30		2.39	0.19	0.09	0.92	0.30	3.79	0.01	0.46	0.30	0.19	0.06	1.41	1.60	0.36	0.07	0.16	0.08	0.62	0.62	1.63	0.82	0.53	1.34	0.38	0.84	0.05	0.53	0.41	0.11	0.01	1.42
es	0.08	0.96	1.66	0.01	0.32	0.61	5.44	0.57	24.54		0.11	0.70	0.07	0.23	7.92	0.06	0.72	0.36	0.20	0.11	2.83	1.73	0.48	0.07	0.16	0.03	0.63	0.42	0.52	2.23	0.42	1.74	0.33	0.78	0.04	0.54	0.62	0.09	0.01	1.73
et	0.10	1.36	0.16	0.00	0.63	2.08	9.29	0.91	17.61	2.23		0.26	0.16	1.76	4.80	0.09	1.23	0.65	0.24	0.16	2.76	1.44	1.00	0.10	0.28	0.05	0.66	1.46	1.14	1.01	0.77	6.81	0.44	2.76	0.09	1.10	2.16	0.13	0.01	2.94
eu	0.09	1.35	1.56	0.04	0.28	0.98	5.10	0.67	16.96	12.14	0.13		0.10	0.31	23.25	0.06	0.97	0.37	0.37	0.13	1.84	1.21	0.51	0.06	0.18	0.08	0.54	0.37	0.28	1.09	0.51	1.22	0.23	0.68	0.05	0.71	0.64	0.11	0.01	3.11
fa	0.07	2.17	0.28	0.01	0.30	0.78	5.79	0.48	30.14	2.89	0.22	0.13		0.17	4.59	0.04	1.04	0.43	0.23	0.04	1.81	1.95	0.67	0.06	0.22	0.05	0.43	0.37	0.55	1.24	0.42	1.92	0.31	0.73	0.07	1.22	0.62	0.13	0.01	2.27
fi	0.14	1.42	0.30	0.00	0.69	1.39	8.00	1.15	27.30	3.16	0.60	0.23	0.11		6.14	0.05	1.22	0.70	0.26	0.17	2.55	2.08	0.74	0.10	0.26	0.08	0.79	1.33	0.76	1.36	0.60	4.39	0.41	4.84	0.09	0.89	1.03	0.14	0.01	2.62
fr	0.10	1.22	0.45	0.01	0.65	0.57	6.72	0.48	21.85	3.83	0.17	0.29	0.08	0.41		0.02	0.69	0.64	0.19	0.09	3.06	2.18	0.44	0.17	0.13	0.08	1.19	0.51	1.88	1.42	0.49	1.90	0.94	0.98	0.05	0.59	0.55	0.10	0.01	1.77
gn	0.26	2.80	0.36	0.03	0.26	3.35	9.70	0.39	31.90	24.28	0.23	1.27	0.00	0.33	4.26		1.76	0.39	0.65	0.16	1.73	2.12	2.85	0.16	0.52	0.33	0.32	0.88	0.72	14.63	0.00	2.02	0.52	0.75	0.46	1.66	1.50	0.49	0.00	6.35
he	0.14	3.97	0.33	0.01	0.74	1.79	11.36	1.12	30.53	3.71	0.20	0.26	0.18	0.31	6.31	0.06		0.90	0.23	0.08	2.88	2.11	1.13	0.16	0.28	0.06	0.79	0.61	1.37	1.41	1.24	3.96	0.61	1.27	0.09	1.57	1.65	0.14	0.00	3.09
hu	0.06	1.10	0.74	0.01	2.47	0.99	9.95	0.62	16.38	5.16	0.17	0.37	0.07	0.33	14.67	0.03	0.86		0.16	0.13	5.10	1.33	0.63	0.14	0.14	0.03	0.73	0.55	0.86	1.09	3.01	2.45	0.99	1.02	0.04	0.86	1.10	0.07	0.00	1.90
id	0.08	1.52	0.16	0.07	0.26	0.78	4.69	0.46	15.42	1.94	0.07	0.14	0.12	0.15	6.14	0.04	0.96	0.33		0.05	2.16	2.61	1.04	0.06	5.15	0.05	0.84	0.27	0.31	0.94	0.43	1.17	0.25	0.57	0.05	0.73	0.40	0.20	0.07	3.07
is	0.11	1.90	0.16	0.00	0.50	4.39	10.60	1.39	23.92	2.42	0.32	0.45	0.09	0.50	5.17	0.16	1.91	1.06	0.35		2.44	1.95	1.54	0.08	0.43	0.06	0.84	2.59	0.72	1.51	0.83	1.94	0.41	3.52	0.12	1.21	1.22	0.22	0.01	3.32
it	0.09	0.97	0.44	0.02	0.97	0.64	7.39	0.67	21.92	4.28	0.15	0.21	0.15	0.32	8.77	0.04	0.67	0.70	0.12	0.09		2.02	0.42	0.08	0.11	0.04	0.73	0.58	0.84	1.58	0.81	2.16	0.40	0.97	0.04	0.66	0.57	0.08	0.02	1.43
ja	0.05	0.67	0.13	0.00	0.30	0.50	4.15	0.32	15.68	1.65	0.07	0.13	0.05	0.17	3.03	0.02	0.55	0.29	0.16	0.05	1.95		1.31	0.04	0.19	0.04	0.36	0.26	0.32	0.65	0.28	1.46	0.20	0.53	0.05	0.41	0.45	0.13	0.01	4.56
ko	0.10	1.41	0.16	0.02	0.38	1.03	5.84	0.68	21.42	2.69	0.11	0.15	0.11	0.23	4.10	0.06	0.98	0.44	0.29	0.09	2.15	10.07		0.11	0.41	0.05	0.49	0.38	0.40	1.13	0.52	2.07	0.31	0.80	0.10	0.87	0.65	0.27	0.01	7.82
mk	0.09	1.76	0.15	0.02	0.43	1.71	13.28	2.14	15.18	3.48	0.16	0.31	0.17	0.21	3.85	0.13	1.33	0.61	0.20	0.14	2.66	0.98	0.72		0.32	0.06	0.47	0.46	0.54	0.87	1.36	3.37	2.19	0.95	0.08	2.13	1.47	0.11	0.03	2.34
ms	0.04	1.49	0.52	0.00	2.38	0.50	7.08	0.13	12.76	4.26	0.05	0.33	3.42	0.09	15.50	0.03	0.56	1.44	9.12	0.02	3.42	1.84	0.59	0.04		0.93	0.38	0.87	0.59	0.83	1.34	0.69	1.78	0.33	0.03	2.85	0.29	0.14	0.01	2.06
ne	0.11	0.83	0.05	0.00	0.10	0.93	3.72	0.11	20.88	0.99	0.12	0.09	0.14	0.09	1.51	0.05	0.71	0.17	0.41	0.02	0.37	0.97	0.95	0.04	0.26		0.19	0.27	0.12	0.61	0.18	0.61	0.05	0.46	0.06	0.47	0.48	0.13	0.00	3.74
nl	0.10	0.45	0.23	0.02	0.55	0.62	4.48	0.24	11.21	2.00	0.14	0.12	0.04	0.15	5.42	0.02	0.36	0.37	1.80	0.05	1.43	0.66	0.22	0.03	0.32	0.02		0.37	0.95	1.11	0.40	0.82	0.27	0.77	0.03	0.36	0.29	0.12	0.02	1.01
no	0.14	1.28	0.25	0.02	0.54	2.83	8.27	0.63	24.28	3.30	0.32	0.18	0.11	0.87	5.07	0.07	1.07	0.54	0.19	0.33	2.43	1.51	0.80	0.08	0.23	0.07	0.90		0.74	2.27	0.48	2.58	0.37	4.29	0.11	0.80	0.77	0.14	0.01	3.68
pl	0.11	0.97	0.39	0.00	1.23	0.71	7.63	0.51	19.51	3.33	0.53	0.19	0.10	0.36	7.40	0.03	0.85	0.68	0.15	0.09	2.69	1.22	0.41	0.11	0.14	0.26	0.72	0.56		1.29	0.54	3.45	0.41	1.10	0.05	0.57	1.92	0.08	0.00	1.46
pt	0.07	0.99	0.42	0.01	0.98	0.71	6.51	0.56	23.09	5.82	0.13	0.24	0.08	0.25	7.80	0.06	0.75	0.34	0.15	0.07	2.78	1.78	0.47	0.06	0.14	0.05	0.62	0.44	0.78		1.02	1.43	0.71	0.94	0.04	0.55	0.46	0.08	0.00	1.60
ro	0.22	1.54	0.40	0.00	0.45	0.98	8.88	0.67	13.01	2.70	0.18	0.19	0.12	0.29	13.45	0.09	0.87	1.81	0.12	0.06	3.74	1.24	0.46	0.16	0.13	0.02	0.64	0.70	1.16	2.18		3.00	0.56	0.96	0.12	0.75	9.14	0.07	0.00	1.34
ru	0.08	0.93	0.30	0.01	0.59	0.58	6.97	0.58	16.59	3.16	0.33	0.16	0.14	0.46	5.27	0.03	0.83	0.43	0.16	0.08	2.28	1.55	0.45	0.12	0.14	0.04	0.57	0.44	1.67	1.59	0.67		0.57	0.94	0.04	0.70	5.26	0.10	0.00	1.85
sr	0.07	1.06	0.15	0.01	1.66	0.83	7.74	0.95	20.12	1.78	0.10	0.15	0.08	0.19	6.52	0.05	0.78	0.77	0.11	0.06	4.18	0.76	0.42	0.91	0.12	0.03	0.29	0.25	0.79	0.58	0.76	2.41		0.60	0.05	0.83	0.83	0.10	0.00	1.42
sv	0.03	0.27	0.06	0.03	0.14	0.53	2.30	0.13	7.26	0.83	0.08	0.05	0.03	0.42	2.83	0.01	0.23	0.16	0.06	0.06	0.51	0.39	0.13	0.03	0.04	0.01	0.18	0.48	0.19	0.36	0.11	0.61	0.15		0.02	0.17	0.20	0.04	0.01	1.00
sw	0.65	3.30	0.14	0.00	0.24	1.35	7.46	0.40	35.52	2.27	0.05	0.37	0.12	0.16	9.02	0.28	1.86	0.40	0.36	0.08	2.70	1.57	1.46	0.06	0.57	0.08	0.56	0.38	0.28	1.37	0.62	1.05	0.30	1.14		4.05	0.51	0.41	0.03	3.18
tr	0.10	2.47	0.26	0.01	0.51	1.23	8.41	1.14	21.49	3.90	0.18	0.25	0.87	0.36	5.16	0.05	1.58	0.66	0.22	0.08	2.35	1.97	0.81	0.21	0.25	0.04	0.59	0.48	0.56	2.07	0.77	3.02	0.77	1.13	0.06		0.93	0.12	0.01	2.58
uk	0.07	1.04	0.37	0.01	0.56	0.72	5.86	0.72	13.71	3.09	0.36	0.23	0.09	0.26	9.29	0.03	0.77	0.45	0.16	0.07	2.85	1.57	0.47	0.20	0.15	0.03	0.42	0.35	2.28	0.81	3.46	9.73	0.43	0.93	0.03	0.70		0.08	0.00	1.74
vi	0.02	0.39	0.13	0.04	0.63	0.21	2.21	0.10	8.10	1.66	0.04	0.09	0.04	0.07	4.30	0.02	0.19	0.35	0.14	0.01	0.87	0.77	0.39	0.02	0.12	0.02	0.12	0.13	0.20	0.78	0.30	0.65	0.06	0.22	0.02	2.63	0.30		0.02	2.69
war	0.01	0.10	0.16	0.05	0.03	0.06	1.57	0.03	3.00	1.26	0.03	0.08	0.02	0.02	3.41	0.02	0.07	0.04	0.07	0.00	0.73	0.17	0.06	0.01	0.02	0.01	0.06	0.07	0.27	0.63	0.03	0.21	0.02	0.08	0.01	0.06	0.11	0.03		1.11
zh	0.08	0.79	0.33	0.00	0.29	0.56	5.39	0.34	17.27	3.45	0.08	0.17	0.10	0.18	6.31	0.05	0.53	0.60	0.24	0.06	1.97	6.19	1.42	0.06	0.33	0.04	0.31	0.30	0.40	1.40	0.31	1.63	0.23	0.53	0.06	0.51	0.80	0.31	0.01	

Figure 29. Culture spread in 293 Wikipedia language editions (1/2). *Each row shows the extent a Wikipedia language CIRA occupies (% articles) in other Wikipedia languages (located in the x axis) calculated like in Table 13. The shape and colour of each Wikipedia language edition indicates its status (official or not official) and territorial relationship with the language CIRA (overlapping, neighbouring and none).*

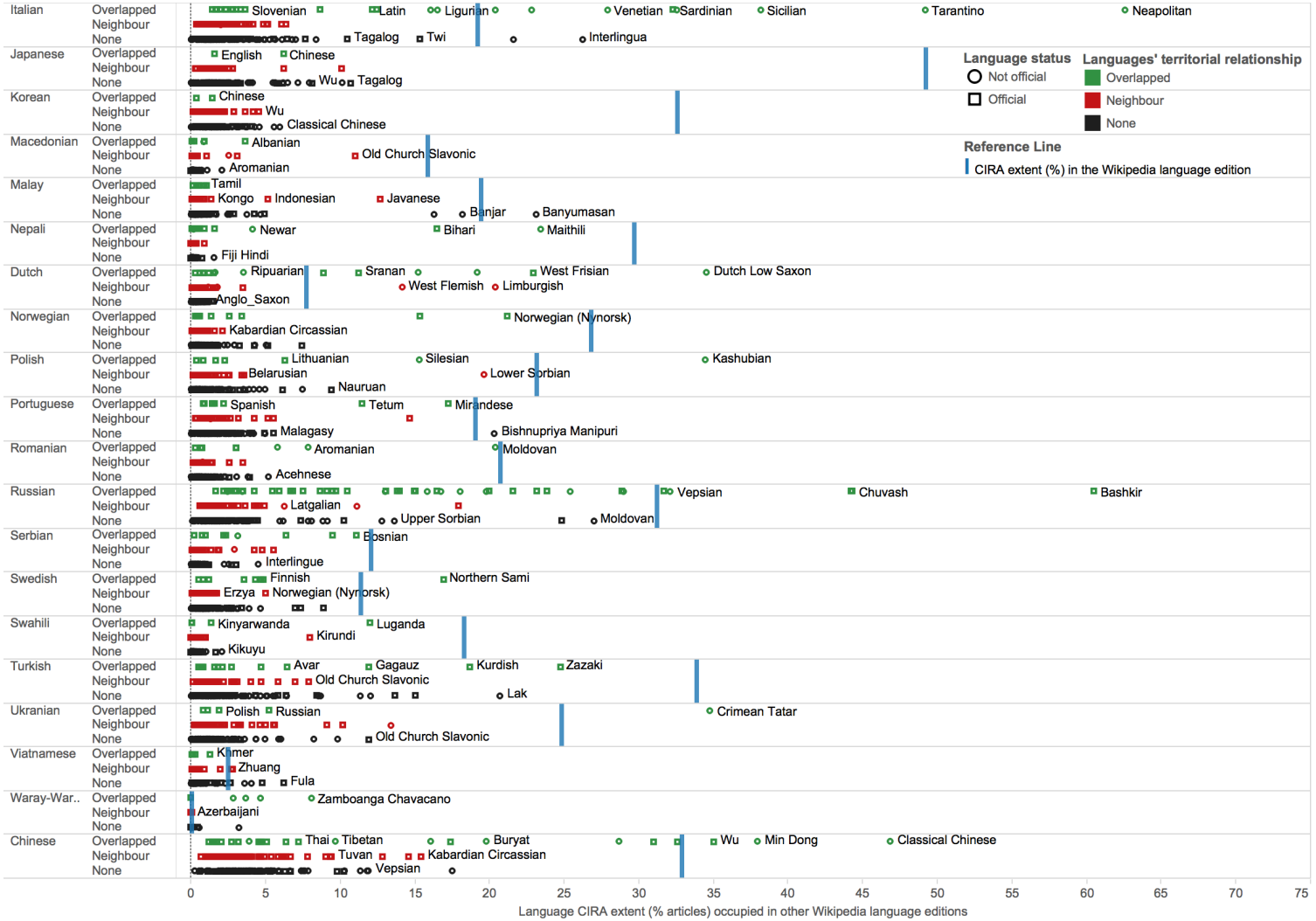


Figure 30. Culture spread in 293 Wikipedia language editions (2/2). *Each row shows the extent a Wikipedia language CIRA occupies (% articles) in other Wikipedia languages (located in the x axis) calculated like in Table 13. The shape and colour of each Wikipedia language edition indicates its status (official or not official) and territorial relationship with the language CIRA (overlapping, neighbouring and none).*

In Figure 29 and Figure 30, I repeat the same calculation as in Table 13 but instead of showing the relationship between 40 language editions, I have considered showing the percentage that 40 Wikipedia language edition CIRA occupy in all the existing 293 Wikipedia language editions. With the entire list of Wikipedias it is possible to see the relationship between the 40 Wikipedia language edition CIRA and small languages I did not include in the study. The precise percentage a language's CIRA occupies in other languages is generally several orders of magnitude smaller than in their own's. These figures show that this is not the case for these small languages which coexist in the same territory. In fact, the colour of each Wikipedia language edition indicates territorial relationship with the language CIRA (overlapping, neighbouring, none), while the circles or squares indicate language status (official, unofficial). Usually, the CIRA from big languages that have a territorial relationship with smaller languages, tend to occupy an important extent of the latter's overall content. This is the case of Italian and its dialects (62.49% of Neapolitan language), but also of Chinese with Classical Chinese (46.85%) or Afrikaans with Venda (34.47%). Possibly, in these cases, some of the meanings from the small and the big languages are shared and considered part of their CIRA. Hence, this shows how some cultural identities can coincide with or integrate others.

a) Topics across languages

Finally, to further investigate the CIRA availability across languages, I have calculated the topical coverage of the articles existing across languages. Thus, I have weighted the assignation of each CIRA article to the main categories (i.e. Geography, People, Religion, Sports) on the basis of the number of Interlanguage Links, and computed the total for the entire set of articles. This allows to observe whether certain topics that belong to the cultural identity of a language appear more relevant to other cultures.

Results. In Figure 31, I present a classification of topics in terms of percentage of articles and percentage of ILLs. The most representative category is again Geography, which exhibits an even higher proportion of ILLs than number of articles (26.1% vs 22.0%), while the category People has a slightly lower percentage (17.4% vs 19.4%). This suggests that when editors from a language edition import content from the CIRA of another language, they consider these topics as the most noteworthy, to be disseminated first. Some remarkable differences can be noticed for some categories, such as Religion in the Arabic CIRA, that contains few articles, but has a much higher proportion of ILLs, indicating that these articles are often shared with other language editions. A similar effect can be observed in the category Sports of the Spanish Wikipedia.

Discussion. The culture gap is a problem for Wikipedia language editions, since editors are not able to cover the concepts from other language cultures. Few languages cover a good percentage of the other languages' CIRA. English is an exception, but still it only covers in average a 33.71% of other languages CIRA (median 28.27%, standard deviation 19.36%). Likewise, only the CIRA from languages such as English or German occupy an average percentage higher than 5% of the articles of other Wikipedia language editions. Such a gap is not surprising, because of the very definition of cultural identity. However, bridging it could help achieving the goal of gathering the sum of human knowledge that Wikipedia advocates. The results have also shown that CIRA articles about geography and people tended to be among the most shared across languages.

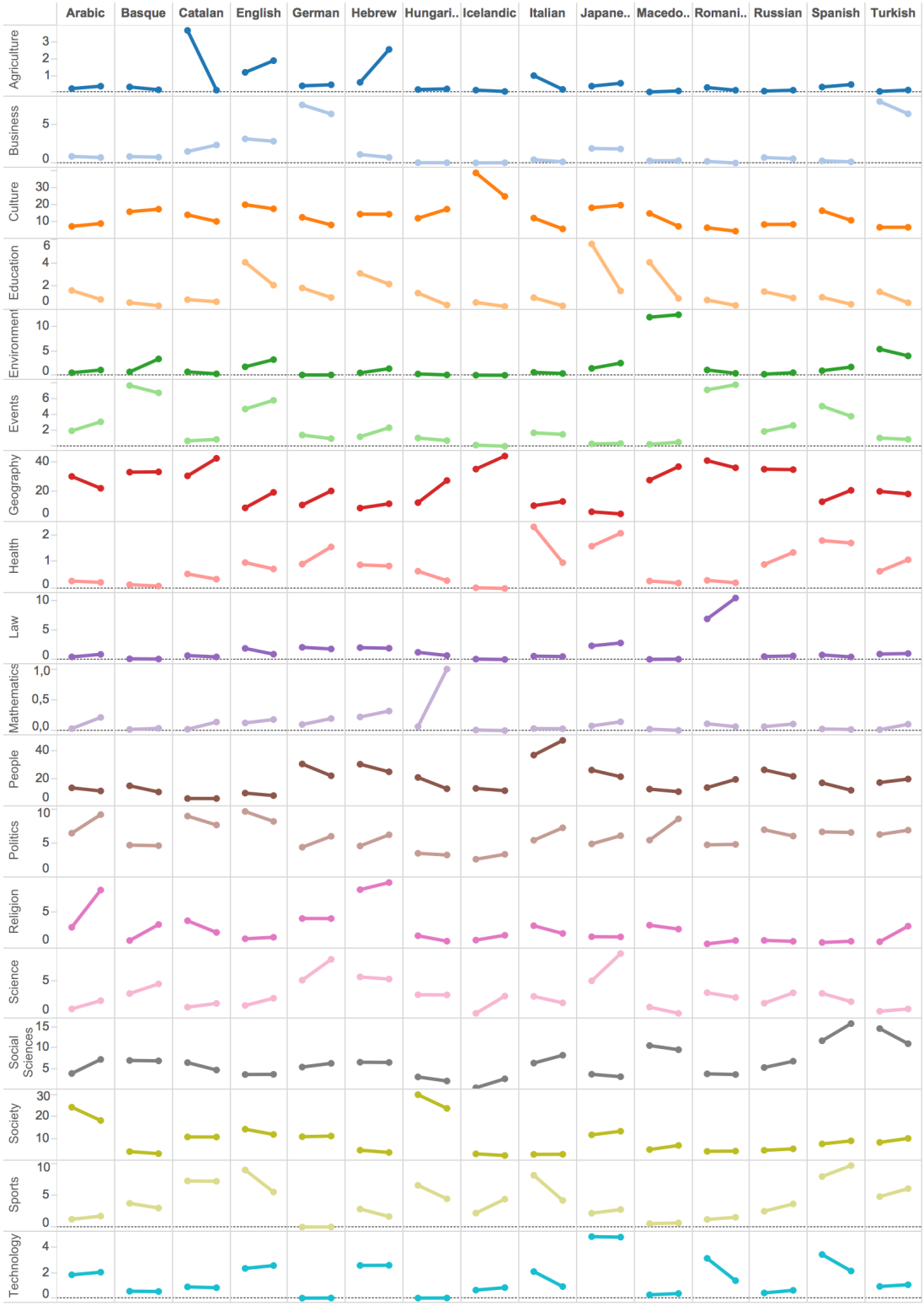


Figure 31. Comparison between CIRA topical coverage in its Wikipedia language edition and in the other editions. *For each topic, the first point is the percentage it occupies in the local Wikipedia language edition CIRA, while the second point corresponds to the topic coverage weighted with the number of Interlanguage Links of its articles. Any slope variation implies that articles linked to this topic are more or less available in other languages.*

7.3.7 Summary of Results

This chapter showed that each Wikipedia contains a non-negligible amount of content representing editors' cultural identities. Previous work analysed the effects of the context over the content, but did not aim at selecting specifically all the articles describing the meanings associated to the editor's values, traditions, history or geography, in other words, their cultural identity. The large extent of this content indicates that editors are motivated to contribute and create articles related to very specific meanings. Therefore, it is consistent with the idea that editors may feel motivated to contribute to those topics as identity-congruent choices.

The analysis of the 40 languages proved that the concepts related to editors' cultural identities range from 7% to 49% of the total number of articles **(RQ1)**. CIRA have been produced with no specific plan, policy or guideline recommending it, but as an effect of editors' preferences. Even though the analysis has been run on very different language editions, the relative size of CIRA slightly correlates with the total number of editors and not with the current active editors. This is in agreement with the concept of cultural identity, which relates to all editors' shared meanings, independently of their level of participation in Wikipedia.

An analysis of the creation of CIRA over time showed that this content grew constantly along with Wikipedia **(RQ2)**. In particular, both CIRA and its segment of Geolocated articles had grown more from 2003 to 2007. This is understandable because these articles comprise the most relevant geographical places for the editors: their cities, towns, rivers, and they can be finite. However, the degree of specialisation that CIRA can reach through very different topics implies that new content can continually appear, and article creation can serve as motivating choices.

Three different analyses (topics, inner relationships and cross-language coverages) helped in understanding the nature of the representation of cultural identities in Wikipedia. As far as topics are concerned, cultural identities contain all sort of meanings relevant to the historical context where people live. In a similar way, CIRA is composed by articles which can be assigned to all the general categories, just like the remaining part of Wikipedia **(RQ3)**. I found that the categories Geography and People are more relevant. However, other categories also play a role in expressing the diversity within the group of CIRA. Cultural identity has been evolved in relation to territory and power. Editors need to understand that their nearby environment and their meanings are all reflected in Wikipedia. In fact, the CIRA topical coverage reminds of a local specialised version of an encyclopaedia.

Even though CIRA is topically diverse, it is also characterised by a sense of unity. By analysing the links located in CIRA, I found that it is employed to define itself: an average of 78.87% of the inlinks come from the same CIRA, while only a 56.17% is directed to it **(RQ4)**. This shows that a cultural identity is represented through a self-referential structure of meanings. When analysing the references from the rest of the Wikipedia directed at CIRA, I saw that they account for an average of 3.73% of the outlinks - with exceptions like English or German, which were over 10%. Even though editors may have

used CIRA as examples in the text to illustrate other Wikipedia topics, this occurs less than I would have expected.

According to the analysis based on ILLs in the 40 languages, CIRA articles are 4.5 times less shared than the average content of each language edition **(RQ5)**. Since content representing cultural identities responds to editors' identity-congruent choices, it was expected that these articles would find no equivalence in other language editions. This shows that the lack of correspondence of content between languages corresponds mainly to articles in CIRA. The graphs provided to illustrate this culture gap could be useful to show editors the content from the cultural identities of other languages that could be imported, worked on and extended. Currently, only large Wikipedia language editions partially cover other languages' CIRA, and, the most shared articles tend to be about geography and people.

All in all, analysis the representation of cultural identities in Wikipedia contributed with new findings about how each language edition is culturally contextualised, especially the nature of the content. The results show evidence for the representation of cultural identity in Wikipedia for very different and distant societies. CIRA extent in terms of articles and the fact that it has been created over time could suggest that it responds to the influence of an identity-based motivation. It is possible to speculate that without an identity-based motivation that fostered editors to create the cultural identities representations in each language, its extent would not be that large when compared to the representations of the rest of language associated cultural identities in that language edition. Likewise, it would have been mostly created during the first years of Wikipedia. In order to scrutinise the influence of cultural identity on editor participation, the following chapter will propose new analysis of the editors and the content characteristics.

Chapter 8. Cultural Identities and Engagement

Once obtained these articles, in this chapter I propose two blocks of research questions in order to estimate the influence of cultural identity in Wikipedia language editions, both in content and in Wikipedia editor engagement.

Firstly, I measure engagement in cultural identity representations and I propose a solution in order to improve cross-language coverage (8.1). Secondly, I analyse participation from the editor perspective to explore how cultural identity representations are created (8.2).

8.1 Participation in Cultural Identity representations

In this section, I present the results of the analysis of the participation in cultural identity representations.

8.1.1 Research Questions

Cultural Identity Related Articles have proved to cover a considerable proportion in each Wikipedia. This confirms that when editors want to contribute they often create and edit those articles whose meanings are identity-congruent. However, the extent in number of articles is a partial indicator, whereas the number of edits received by these articles compared to the total amount each Wikipedia receives could be a complementary and better proxy for the influence of cultural identity. This leads to the first research question:

RQ1. *Does content representing editors' cultural identities reflect a higher participation than the rest of the Wikipedia language edition content?*

Wikipedia articles cover all topics of readers' interest. That is, Wikipedia is a powerful tool for people in order to understand the main concepts required to follow any kind of event (Ciampaglia, Flammini, & Menczer, 2015; Keegan et al., 2013). However, some studies show a certain misalignment between readers demand and article production (Lehmann, Müller-Birn, Laniado, Lalmas, & Kaltenbrunner, 2014; Warncke-Wang, Ranjan, Terveen, & Hecht, 2015) (e.g. articles with higher quality do not imply more attention from readers). Then, I wonder about the interest cultural identity representations might draw from readers, considering these articles may be relevant in order to be informed of their same immediate environment.

RQ2. *Does content representing editors' cultural identities respond to readers' information demand?*

Articles can be analysed using their characteristics, ranging from the number of categories in which they are members, images, or external references. After sufficient participation, articles tend to be more developed in article characteristics and communicate better its

content. Cultural identity representations provide a wide variety of topics, which may require different configurations of these features. Therefore, I ask:

RQ3. *Does content representing editors' cultural identities reflect a greater level of detail than the rest of the Wikipedia language edition content?*

As seen in Section 7.3.6, good part of the content not shared across languages belongs to cultural identity representations. Some language editions present a strong isolation, despite these languages may have a large number of speakers. Hence, considering the informational role of Wikipedia in the current society, this carries a problem of disinformation to readers who do not have content about other cultures. To find a solution and bridge this culture gap, I propose to study the characteristics of shared content. For instance, as seen in Section 7.3.6, articles about topics related to geography and people tended to exist across languages. The fourth research question is:

RQ4. *Does content representing editors' cultural identities which exist across different Wikipedia language editions exhibit higher levels of participation?*

Once many Wikipedia language editions have grown in number of articles and also in many topics, bridging the culture gap by obtaining the content representing other cultural identities is the opportunity for having more multicultural Wikipedia language editions. However, editors might find it difficult to guess which content can be more relevant to import from other language editions cultural identity representations. Perhaps, it would be useful to develop a method to select the concepts which should receive more priority in translation. Therefore, in this regard it would be necessary to find at least a criterion.

RQ5. *Which articles in the content representing cultural identity should become a priority to be translated into different Wikipedia language editions?*

8.1.2 Editor Interactions in CIRA (RQ1)

Cultural identities are often situationally cued in the process of contributing to Wikipedia. The extent of CIRA confirms that the creation of articles is imbued with identity-based meanings. To have a more detailed view, I quantified the number of edits in the different article types. I make a distinction between human editors and bots, because bots' behaviour is directed by the algorithms coded by small number of editors, and since the quantity of edits they perform is much higher than human, they could mask human edits. In further subsections, when I say edits I only refer to human edits.

Results. Figure 32 shows the percentage of edits human editors made in CIRA, the percentage of edits made by all kind of editors (including bots) in CIRA, and the percentage of articles previously calculated. Results show that the share of human edits in CIRA is much higher in twelve of fifteen cases. In some of them like Icelandic or English, it is almost equivalent, although for others the percentage doubles (Arabic grows from a 26.92% to a 49.95% and Catalan from a 16.17% to a 28.57%). By taking into account the total number of edits which includes bot edits, the percentage is nonetheless higher than the percentage computed with number of articles **(RQ1)**.

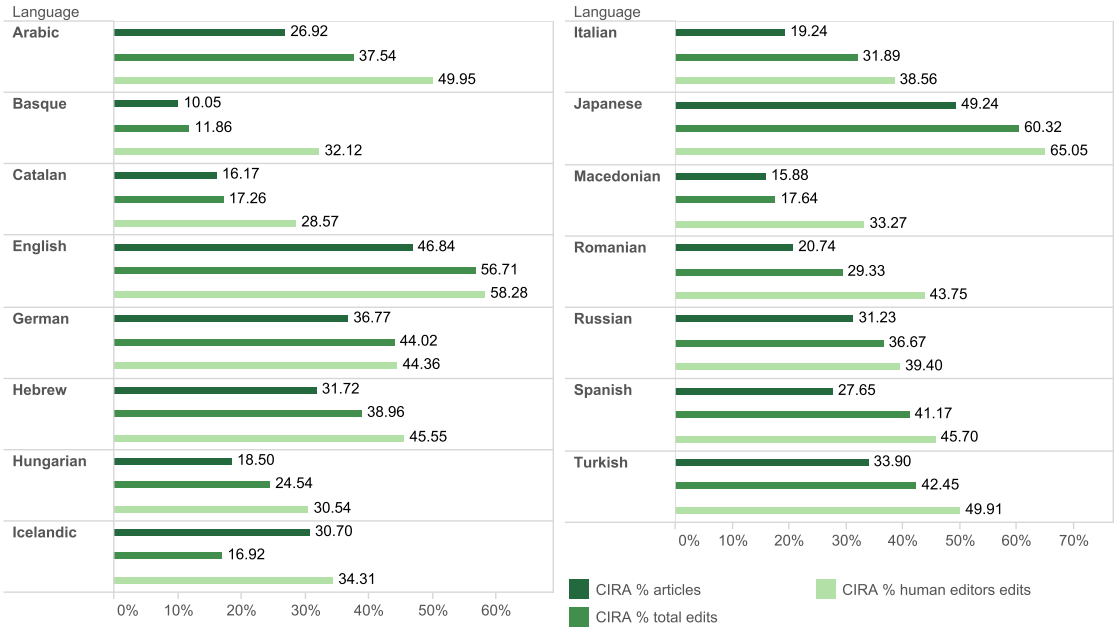


Figure 32. Comparison of CIRA extent in percentage of articles, human edits and total edits (including bot edits).

Discussion. By comparing the extent in articles and in edits, it is possible to state more safely that CIRA may respond to the influence of cultural identity. Because an edit is the minimal interactions in order to modify the content of an article, and each of them can be driven to find identity-congruence. These percentages imply for several languages that an important part (and sometimes a majority) of their interactions is driven at some level by this motivation type.

I conjecture that other factors multiply the effects of cultural identity (such as community dynamics like peer coordination, vandalization which require revert, among others) or ease the participation in this content. Previous research has demonstrated that articles with more edits encourage more editors to continue contributing to them (Aaltonen & Seiler, 2015). This could be more intense in articles concerning the editors' cultural identities, which by definition relate the entire community. Likewise, interactions by editors or the community maintance tasks by functional roles could also create a cumulative effect. The cumulative effect would be likely to happen. Either with the intent of completing this content, or engaging on a controversial topic, in both cases, editors may be more prone to edit because they share the same values. Some of these considerations will also be taken into account in Section 8.2, when I will analyse the influence of cultural identity on editor participation.

8.1.3 Editor and Reader Engagement with CIRA (RQ2)

a) CIRA segments and the rest of Wikipedia
In the previous section I have confirmed that CIRA is devoted a big part of the editing work. However, comparisons between CIRA and the rest of Wikipedia as groups of articles may be insufficient: there could be an imbalance of edits in some very popular articles (e.g. capital cities, celebrities or political figures), while the rest of CIRA could receive less participation than the average of Wikipedia. For this reason, it would be necessary to compare CIRA and the rest of Wikipedia at article level.

Likewise, I argue that the demand from readers could increase the saliency of cultural identities in each Wikipedia language edition. Some articles are constantly created in order to explain previously unavailable concepts, places or people which appear in the news. Some others respond to a lonely initiative of a highly-motivated editor. Then, by comparing reader to editor engagement in CIRA, it would be easier to see in what way editors are also reinforced by outside the project expectations. An editor may decide to edit an article because its topic is identity-congruent, but acknowledging at the same time its topic popularity. In this sense, the different segments of CIRA presented in the previous chapter (CIRA geolocated articles, CIRA with keywords on title and the rest of CIRA) represent different sorts of information with very different levels of popularity (e.g. a geolocated article on a city may be very different from a summary of a literature genre). Then, it would be interesting to compare the CIRA segments to see how editing and popularity differs in them.

Taking this into account, I propose establishing a double-level comparison between reader engagement (using the number of page views for each article page from May 2015 until January 2016) and editor engagement (using the number of edits during the entire history until January 2016), in the different segments of CIRA and in the rest of Wikipedia. In this regard, I propose using the Kruskal-Wallis test in order to determine whether there exist significant changes. However, this test requires the different groups to compare to have exclusive members. Since the CIRA segments have an overlap in some articles from CIRA Geolocated and CIRA Keywords, I propose dividing them into the smaller segments: CIRA Keywords – CIRA Geolocated as those coincident in both

segments, CIRA Keywords without the previously selected, CIRA Geolocated without the previously selected, the rest of articles which compose CIRA.

Results. Table 14 shows the mean ranks for the number of edits and the number of page views for each CIRA segment. Results from the Kruskal-Wallis test show that differences between segments are all significant for a p-value lower than 0.001. However, since the test does not provide information regarding which content type is significantly different from each other, a Dunn's test (1964) procedure with a Bonferrroni have also been conducted in order to conduct a pairwise comparison. Extended results with each pair p-values are located in Table 34 (page views) and Table 35 (edits) in Appendix 3.

The results are significant for a vast majority of the cases; only in few languages like the Macedonian Wikipedia it is possible to see that differences CIRA rest and CIRA GL are not significant in terms of edits and page views, or in the Icelandic Wikipedia between CIRA GL and CIRA KW-GL in terms of edits. Mainly, all the CIRA segments have higher values for both edits and page views than rest of articles from Wikipedia (**RQ2**). In Table 14, a pattern transition can be seen in the mean ranks in almost all languages: CIRA Keywords-Geolocated obtains the highest mean ranks, followed by either CIRA Geolocated or CIRA Keywords. In almost all cases, the rest of CIRA is lower than the other segments, but still higher than the rest of Wikipedia. Since the test has been applied to both edits and page views with the same population of articles, it is possible to compare the mean ranks obtained for two metrics for each group of articles. For almost all languages and in any of the CIRA segments, mean ranks computed for edits are higher than for page views.

For illustrative purposes, I included Figure 33, which shows the average values for edits and page views for each segment of CIRA (including the overlapped CIRA Keywords-Geolocated). The figure shows the edits (as bars) and the pageviews (as horizontal lines) both depicted all over their range in the dual-axis y. Hence, it is possible to see again that for the same segments of CIRA the average number of edits is larger than the average number of page views in line with the Kruskal-Wallis and Dunn's statistical significant results. Figure 33 is useful to see the differences in absolute values for the Wikipedia language editions; for instance, while the range of page views for the Japanese Wikipedia is up to 40k, for the Basque Wikipedia it is 400.

Table 14. Mean ranks for the number of edits and number of page views in different segments and intersections of CIRA and the rest of Wikipedia. *Darker colours represent higher mean ranks, indicating higher number of edits and page views in that content type. CIRA KW-GL stands as the intersection of articles with keywords on title and geolocation, CIRA KW as the articles with keywords on title and without CIRA KW-GL, CIRA GL as the articles with geolocation without the CIRA KW-GL, CIRA REST as those articles in CIRA but not included in the previous selections, and WP REST as the rest of Wikipedia without articles from CIRA. Results of a Kruskal-Wallis test are statistically significant for all languages with p-values always lower than 0.001.*

Language	Metric	CIRA KW-GL	CIRA KW	CIRA GL	CIRA REST	WP REST
Arabic	Edits	290,758	246,308	266,367	216,993	173,681
	PV	298,913	249,623	270,021	211,799	174,928
Basque	Edits	197,027	166,467	172,850	160,888	97,755
	PV	185,864	149,178	144,771	139,770	100,220
Catalan	Edits	382,121	349,292	259,115	334,169	221,303
	PV	372,024	344,451	221,513	320,783	226,112
English	Edits	2,498,244	2,623,270	2,562,090	2,795,003	2,211,878
	PV	2,621,479	2,714,416	2,566,498	2,799,549	2,203,191
German	Edits	1,145,545	1,003,962	1,025,320	1,027,417	853,503
	PV	1,191,907	980,990	1,051,316	1,027,248	850,468
Hebrew	Edits	129,083	112,869	125,163	102,682	78,834
	PV	121,954	103,719	127,252	94,977	82,149
Hungarian	Edits	279,864	193,815	261,092	210,794	150,715
	PV	267,768	196,386	266,990	202,851	152,022
Icelandic	Edits	32,655	26,267	28,703	20,042	19,240
	PV	32,894	24,117	29,157	15,083	21,209
Italian	Edits	1,055,822	833,045	888,808	805,224	553,683
	PV	1,005,778	755,232	839,191	775,189	562,063
Japanese	Edits	761,386	577,648	686,429	545,199	420,352
	PV	752,618	555,455	640,870	513,959	451,398
Macedonian	Edits	70,510	64,058	63,015	52,545	38,754
	PV	44,326	43,466	43,001	41,377	41,257
Romanian	Edits	261,977	249,560	230,016	230,892	145,402
	PV	237,757	250,070	213,301	210,767	150,158
Russian	Edits	583,347	804,657	515,294	754,215	591,888
	PV	618,330	768,412	513,234	710,884	604,667
Spanish	Edits	833,199	756,783	783,634	743,856	505,647
	PV	836,554	728,382	730,250	707,357	520,390
Turkish	Edits	158,900	129,431	166,207	150,433	110,832
	PV	155,726	136,848	128,604	136,489	118,873

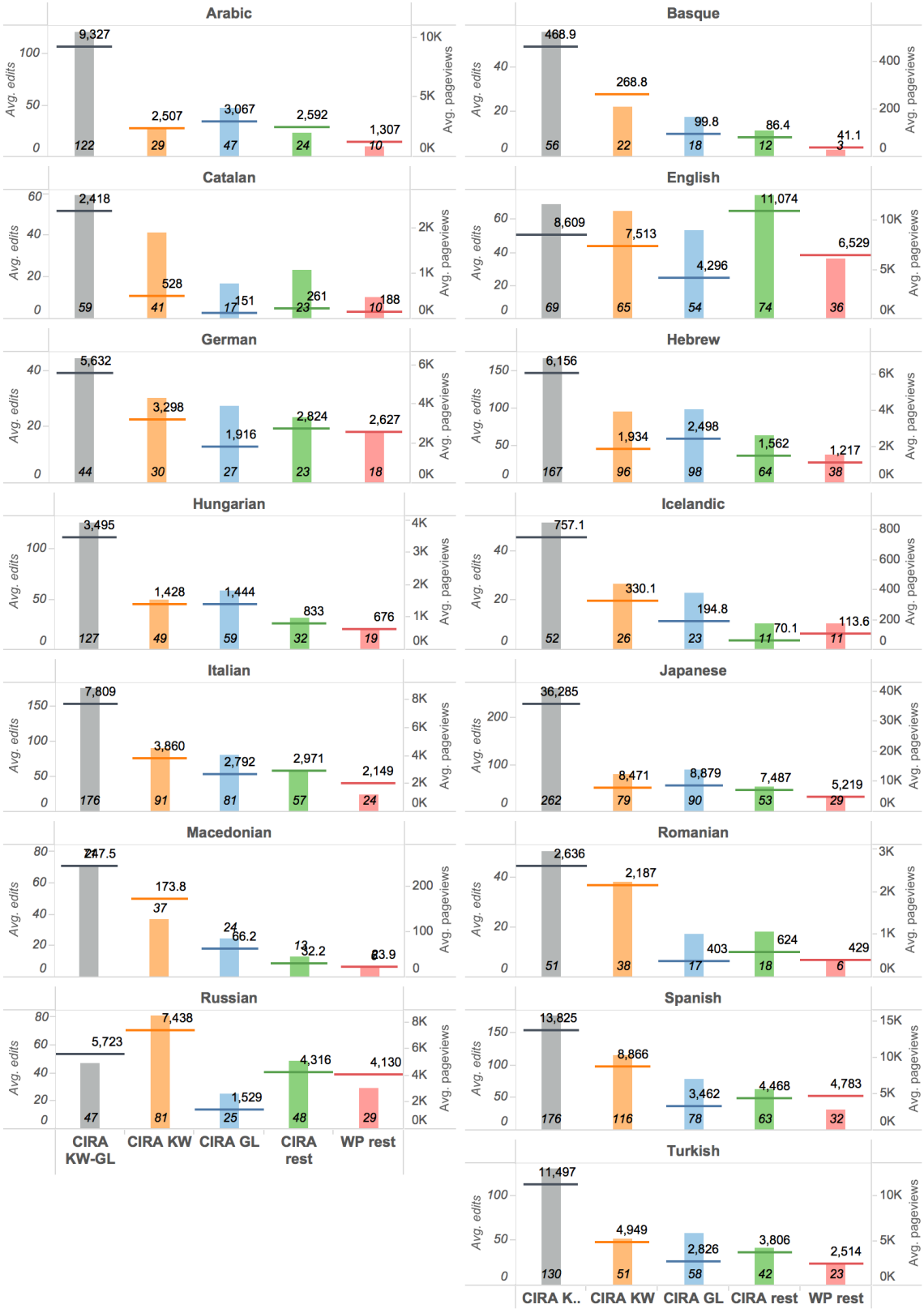


Figure 33. Reader and Editor Engagement in CIRA compared to the rest of Wikipedia. *Average absolute values for the number of edits (bar) and the number of page views (line) in the different article types: CIRA Keywords-Geolocated (CIRA KW-GL), CIRA Keywords (CIRA KW), CIRA Geolocated (CIRA GL), rest of CIRA (CIRA rest), rest of Wikipedia (WP rest).*

b) Coincidence between Editor and Reader Engagement in the Territory

The comparison between CIRA segments and the rest of Wikipedia revealed that cultural identity representations are a shared interest between readers and editors. In fact, showed significantly higher values than for the rest of Wikipedia content, especially those segments such as CIRA Keywords or CIRA Geolocated articles. In this sense, examining in further detail the CIRA Geolocated articles can reveal how the interest is distributed among the concepts they include: namely, the cities, physical geographical elements, among others.

Results. Figure 34 presents a map for the Catalan and English Wikipedia - the rest of the maps for the 13 remaining Wikipedia language editions are located in Section 2.2 Appendix 2, which I encourage the reader to check them. In the map, each article is depicted with a dot. The size of the dot represents the number of page views, and the colour is the number of edits presented as a divergence continuum red-green where the middle point is 250 edits in beige. This way, it will be easy to perceive when an article has been created through many edits, and at the same time, it is popular in terms of page views. Additionally, among the articles where editors and readers find coincidence, I selected a few to provide details.

At first glance, a minority of articles obtain many more edits and page views in than the others. These are usually the main cities from each territory or special monuments within them. For instance, in the Catalan Wikipedia the article 'Barcelona' is among the most edited with 1,833 edits, while for the English, 'New York City' and 'Buckingham Palace' are articles which exceed the 8,851 and 2,583 edits respectively. Each Wikipedia presents a different scale, but the imbalances are similar. Likewise, the map also shows that an important density of articles in the territory revolves around biggest cities. Since the urbanisation in the different countries varies a lot, some languages like the German and Catalan present a very strong density, while others like the Hebrew or Arabic show many empty areas. There are exceptions, but generally the disposition of points is somewhat similar to a population map. In fact, besides the articles about cities, the rest of concepts range from a company to an historical event or monument, which usually take place or are located in urban field.

Since the big dots (which implies more page views) tend to be greener than the rest (which implies more edits), it is possible to state that editors and readers are both interested in these articles. In order to assess the degree of coincidence between these activities, a Spearman correlation has been computed between these two metrics for articles in the entire Wikipedia, Wikipedia without CIRA, CIRA and only CIRA Geolocated articles. Therefore, it is possible to compare the coefficient values obtained from these groups and determine which content type shows a greater editor - reader coincidence (and consequently, hint the possible influence of reader interest on editor participation).

Table 15 shows that coefficients tend to be higher for CIRA than for the rest of Wikipedia or the entire Wikipedia. Although, for many cases CIRA Geolocated is even higher than the other groups in ten out of fifteen cases. This confirms then the coincidence between editors and readers in CIRA and more strongly in the CIRA Geolocated articles.

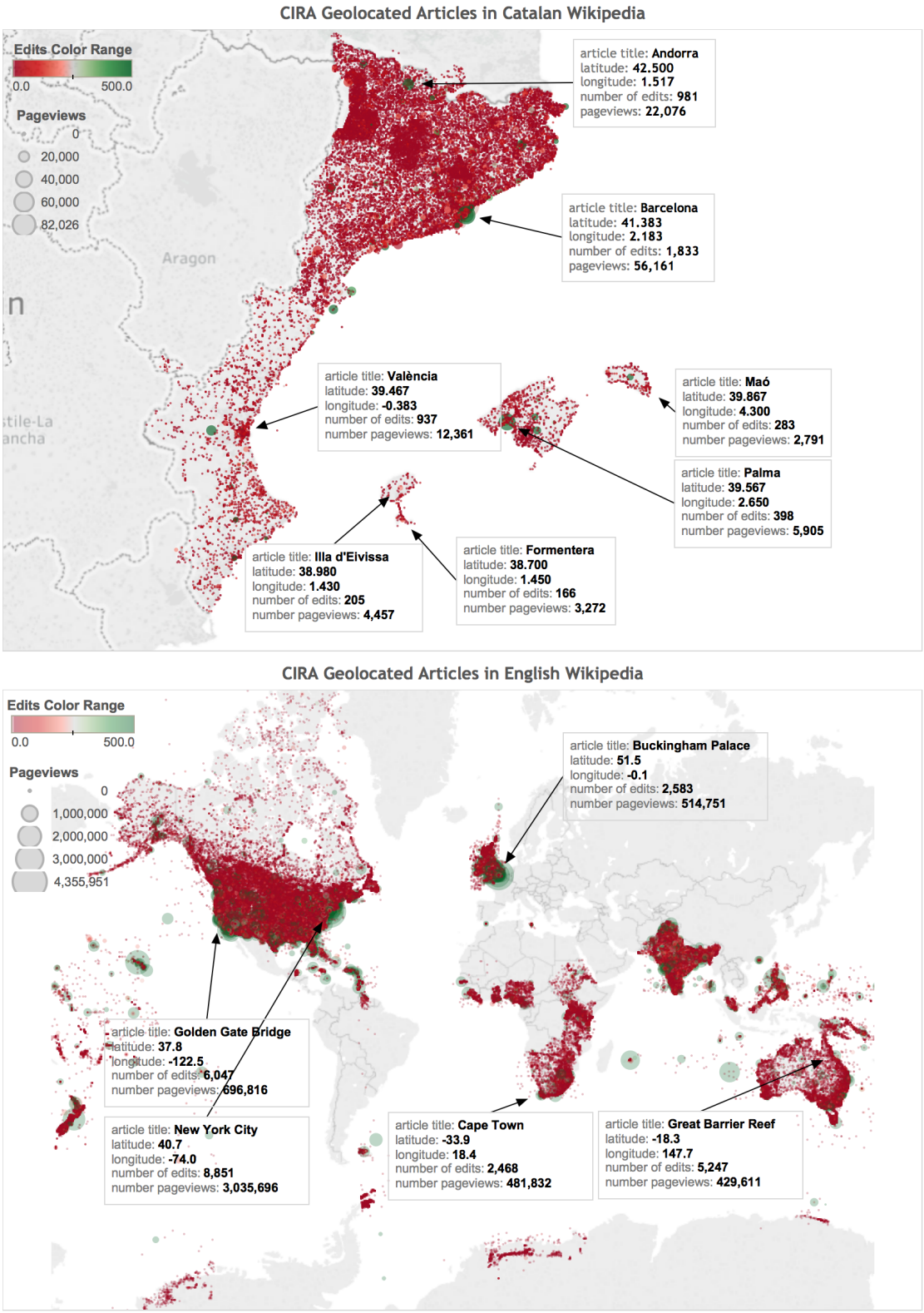


Figure 34. Editor and reader engagement in CIRA Geolocated articles from the Catalan and English Wikipedia (top and bottom respectively). *Each point is a CIRA geolocated article. Colour represents the number of edits, depicted as a continuum from red to green with a middle point of 250 edits in colour beige. Size represents the number of page views. Important geolocated articles are marked with infoboxes.*

Table 15. Spearman correlation for number of edits and page views in different article groups for each language edition. *Columns show the coefficient for the entire Wikipedia (All WP), Wikipedia without CIRA (WP Rest), CIRA and CIRA Geolocated articles (CIRA GL). Correlations are significant for all values at the level p-value<0.01.*

ISO code	Language	All WP	WP Rest	CIRA	CIRA GL
ar	**Arabic**	0.785	0.758	0.802	0.782
eu	**Basque**	0.715	0.706	0.645	0.721
ca	**Catalan**	0.656	0.638	0.714	0.728
en	**English**	0.648	0.597	0.665	0.723
de	**German**	0.667	0.654	0.664	0.727
he	**Hebrew**	0.646	0.618	0.668	0.688
hu	**Hungarian**	0.735	0.729	0.660	0.703
is	**Icelandic**	0.475	0.499	0.497	0.652
it	**Italian**	0.702	0.682	0.661	0.760
ja	**Japanese**	0.685	0.664	0.699	0.819
mk	**Macedonian**	0.063	0.062	0.082	0.108
ro	**Romanian**	0.766	0.754	0.683	0.692
ru	**Russian**	0.732	0.724	0.760	0.737
es	**Spanish**	0.768	0.761	0.691	0.769
tr	**Turkish**	0.610	0.660	0.488	0.554
AVG.	***Average***	0.644	0.634	0.625	0.678

Discussion. In this section, I have presented a comparison between the different segments of CIRA at article level for participation and readership. In first place, results confirmed a higher level of participation in CIRA than in the rest of Wikipedia, especially in articles from the group with both keywords in the title and geolocation tag, and the rest of articles with keywords on the title. This pattern was equivalent for the readership, with a higher number of page views in the segments of CIRA than in the rest of Wikipedia.

I am especially cautious in establishing a causal effect between readers' page views and editors' edits, since previous research showed there exists a misalignment between demand and supply, in other words, that different factors intervene. It is only possible to say that according to the results from Spearman correlation in CIRA there exists a coincidence between readers and editors. The results from this latter analysis comparing geolocated articles, CIRA and the entire Wikipedia showed that the coincidence is mostly focused on the territory. Perhaps the use of Wikipedia as background information for any fact checking, or understanding the news, could explain why readers continually consult articles about their immediate environment.

Because of this, I assume that the readers' demand for content related to cultural identities make Wikipedia a context where cultural identities are more relevant. However, the results from the overall comparison explain that, for each of the CIRA segments, the results from readers' page views were lower than those from editors' edits. Hence, editors engage more in participating into CIRA than readers viewing it. In other words, editors are motivated to edit CIRA even over the possible demand by readers.

8.1.4 CIRA Article Features Analysis (RQ3)

One of the possible additional factors that may explain contributions in cultural identity representations is the editors' prior knowledge on the topic. The initial understanding of the topics could facilitate contributing into Wikipedia. Also, the previous section has shown that CIRA has been dedicated a larger number of contributions at an article level for each of its segments (from keywords to geolocated) than the rest of Wikipedia. Therefore, I wonder whether these articles have also been created at greater detail. In other words, if there is a direct translation from prior knowledge and from the engagement into more detailed articles.

In order to explore CIRA articles, it is necessary to consider the article characteristics defined in Section 5.1, and quantify them into features. I selected: the number of Bytes as a proxy for page length, the number of Bytes in Discussion page, the number of Images, the number of External References, the number of Redirects and the number of Categories. Before proceeding, I have calculated their Spearman correlation between the different features in order to detect if there exists any redundancy.

Table 16. Spearman correlation for the different article features. *The values provided are calculated as the mean of the correlation for each value in the 15 Wikipedia language editions. All the correlations are significant for all values at the level of p-value<0.01.*

Correlations	Bytes	Discussion Bytes	Images	External Ref.	Redirects	Categories
Bytes	1	0.297	0.264	0.549	0.234	0.317
Discussion Bytes	0.297	1	0.028	0.255	0.249	0.272
Images	0.264	0.028	1	0.289	0.101	0.005
External Ref.	0.549	0.255	0.289	1	0.125	0.258
Redirects	0.234	0.249	0.101	0.125	1	0.136
Categories	0.317	0.272	0.005	0.258	0.136	1

Results. Table 16 shows the average value for each coefficient in the 15 Wikipedia language editions. From all the different features, only External References have shown a moderate correlation with Bytes of 0.549 ($p < 0.01$). This means that the rest of features are very different among each other. To study how CIRA is characterised by the different features, I propose comparing the different segments of CIRA and the rest of Wikipedia. I rely again on the Kruskal-Wallis and the Dunn's test in order to evaluate whether there are significant differences between the different content types. Extended results with the Dunn's p-values are located in Section 3.5 in Appendix 3.
Table 17 presents the mean ranks for the different segments of CIRA and the rest of Wikipedia. Results from the Kruskal-Wallis test show that differences between segments are all significant for a p-value lower than 0.001. The vast majority of the pairwise comparison results are significant at a p-value < 0.000, being the most common exception the comparison of CIRA GL with CIRA KW-GL, which can be explained because they share the geolocation characteristic. To complement the mean ranks, Figure 35 shows a visualization of the average value for every feature and for every segment of CIRA, which is consistent with the test results and illustrates differences between languages.

Table 17. Mean ranks to the article features in different segments and intersections of CIRA and the rest of Wikipedia. *Darker colours represent higher mean ranks, indicating higher value of a feature in articles from that content type. CIRA KW-GL stands as the intersection of articles with keywords on title and geolocation, CIRA KW as the articles with keywords on title without CIRA KW-GL, CIRA GL stands as the articles with geolocation without CIRA KW-GL, CIRA rest as the articles not included in the previous selections, and WP REST as the rest of Wikipedia without CIRA. Results of the Kruskal-Wallis test are statistically significant for all languages with p-values always lower than 0.001.*

Language	Feature	CIRA KW-GL	CIRA KW	CIRA GL	CIRA REST	WP REST
Arabic	Bytes	244,726	216,901	211,588	203,973	180,807
	Categories	224,048	192,144	192,134	197,415	184,357
	Disc. Bytes	213,184	196,210	200,803	197,852	183,770
	Ext. Ref.	257,558	170,680	244,129	175,317	189,238
	Images	243,895	173,691	231,585	186,294	186,487
	Redirects	233,883	188,709	188,173	158,634	195,966
Basque	Bytes	125,051	144,308	158,247	117,786	102,555
	Categories	120,302	125,924	145,966	155,201	99,409
	Disc. Bytes	109,265	115,371	124,462	108,668	103,787
	Ext. Ref.	104,692	71,431	141,523	65,819	107,865
	Images	143,115	66,342	135,567	67,213	107,063
	Redirects	116,968	129,276	171,664	121,190	102,467
Catalan	Bytes	316,679	307,325	196,526	248,388	235,238
	Categories	298,471	281,146	268,985	342,367	220,344
	Disc. Bytes	263,869	252,938	212,407	228,217	236,036
	Ext. Ref.	316,163	202,655	251,907	185,404	236,391
	Images	295,188	189,358	236,125	150,451	241,104
	Redirects	313,036	269,525	208,602	268,619	232,642
English	Bytes	2,955,132	3,008,833	2,753,012	2,897,407	2,090,731
	Categories	2,265,595	2,263,784	2,381,995	3,169,861	2,017,590
	Disc. Bytes	2,617,742	2,695,197	2,694,333	2,855,946	2,144,559
	Ext. Ref.	3,402,319	2,436,205	3,315,235	2,745,786	2,111,299
	Images	3,285,670	2,515,258	3,259,721	2,224,949	2,457,625
	Redirects	2,670,577	2,575,053	2,427,049	2,377,581	2,508,588
German	Bytes	1,212,649	1,093,038	1,106,223	1,077,530	819,056
	Categories	1,031,234	958,757	992,115	1,246,658	768,812
	Disc. Bytes	1,038,480	987,030	980,571	983,864	878,448
	Ext. Ref.	1,268,672	918,187	1,205,243	1,093,126	803,412
	Images	1,324,662	850,571	1,336,484	749,425	929,258
	Redirects	978,657	931,976	855,898	904,170	930,261
Hebrew	Bytes	118,740	114,617	97,780	96,636	82,153
	Categories	116,838	93,850	112,119	106,817	77,991
	Disc. Bytes	122,987	108,239	114,537	96,811	81,694
	Ext. Ref.	144,099	110,351	135,107	104,819	77,688
	Images	127,907	87,609	135,021	77,912	89,282
	Redirects	107,731	98,690	97,172	87,138	86,448
Hungarian	Bytes	242,347	223,555	215,896	189,917	155,182
	Categories	157,387	178,695	144,180	241,706	148,029
	Disc. Bytes	205,322	205,414	199,616	174,805	158,717
	Ext. Ref.	266,776	142,403	251,937	163,278	160,679
	Images	270,199	141,505	262,666	113,969	169,635
	Redirects	224,268	190,162	167,415	146,841	164,934
Icelandic	Bytes	30,219	26,652	22,839	23,810	17,940
	Categories	24,557	21,667	23,530	20,528	19,311
	Disc. Bytes	23,567	21,940	19,650	22,843	18,519
	Ext. Ref.	34,985	25,640	27,282	23,613	17,896
	Images	24,896	19,535	24,376	14,293	21,767
	Redirects	30,269	21,664	18,675	16,696	20,952

Language	Feature	CIRA KW-GL	CIRA KW	CIRA GL	CIRA REST	WP REST
Italian	Bytes	925,618	887,202	773,619	740,435	570,527
	Categories	395,591	544,987	527,489	842,875	565,216
	Disc. Bytes	990,000	685,346	785,534	659,731	586,433
	Ext. Ref.	981,074	538,524	791,251	573,657	603,272
	Images	983,212	713,633	862,838	601,062	593,749
	Redirects	678,599	682,145	650,058	594,109	604,890
Japanese	Bytes	721,346	587,482	620,582	495,769	468,203
	Categories	661,975	434,795	582,160	566,138	411,658
	Disc. Bytes	626,757	519,264	511,864	489,626	482,261
	Ext. Ref.	797,062	487,930	759,839	464,993	487,935
	Images	706,529	416,219	733,822	454,134	500,689
	Redirects	639,253	491,188	513,509	472,067	498,181
Macedonian	Bytes	63,361	54,849	45,148	36,395	41,742
	Categories	48,568	31,831	60,132	53,286	39,209
	Disc. Bytes	58,965	43,604	51,311	40,373	41,155
	Ext. Ref.	53,151	35,191	40,472	29,338	43,188
	Images	62,089	44,329	66,348	45,170	40,017
	Redirects	66,156	54,033	39,614	42,327	41,053
Romanian	Bytes	197,772	223,391	169,238	180,257	159,434
	Categories	193,091	168,851	146,937	225,275	155,008
	Disc. Bytes	185,705	195,933	170,531	183,146	159,225
	Ext. Ref.	262,646	172,103	209,773	145,609	161,817
	Images	237,537	120,565	217,815	107,440	167,854
	Redirects	227,158	193,359	230,158	183,234	153,744
Russian	Bytes	630,487	838,433	637,266	787,254	562,474
	Categories	508,201	510,806	522,063	868,658	563,056
	Disc. Bytes	598,314	690,061	517,095	641,957	624,374
	Ext. Ref.	969,570	595,162	1,022,591	579,289	560,675
	Images	785,753	595,625	856,783	625,888	574,920
	Redirects	513,942	596,538	475,201	811,799	585,276
Spanish	Bytes	787,955	786,504	692,432	683,283	528,597
	Categories	519,039	555,329	569,025	756,581	522,153
	Disc. Bytes	672,978	650,916	647,033	606,584	557,360
	Ext. Ref.	778,247	609,908	721,241	558,799	566,833
	Images	843,614	642,799	825,015	502,628	575,098
	Redirects	719,789	671,247	702,217	626,927	547,183
Turkish	Bytes	150,316	143,878	149,487	136,038	117,493
	Categories	127,348	102,895	114,674	136,479	120,837
	Disc. Bytes	166,061	122,865	173,297	150,287	110,617
	Ext. Ref.	163,100	125,904	166,482	128,767	119,930
	Images	155,474	116,878	159,276	118,993	124,747
	Redirects	168,516	124,757	144,545	127,694	121,837

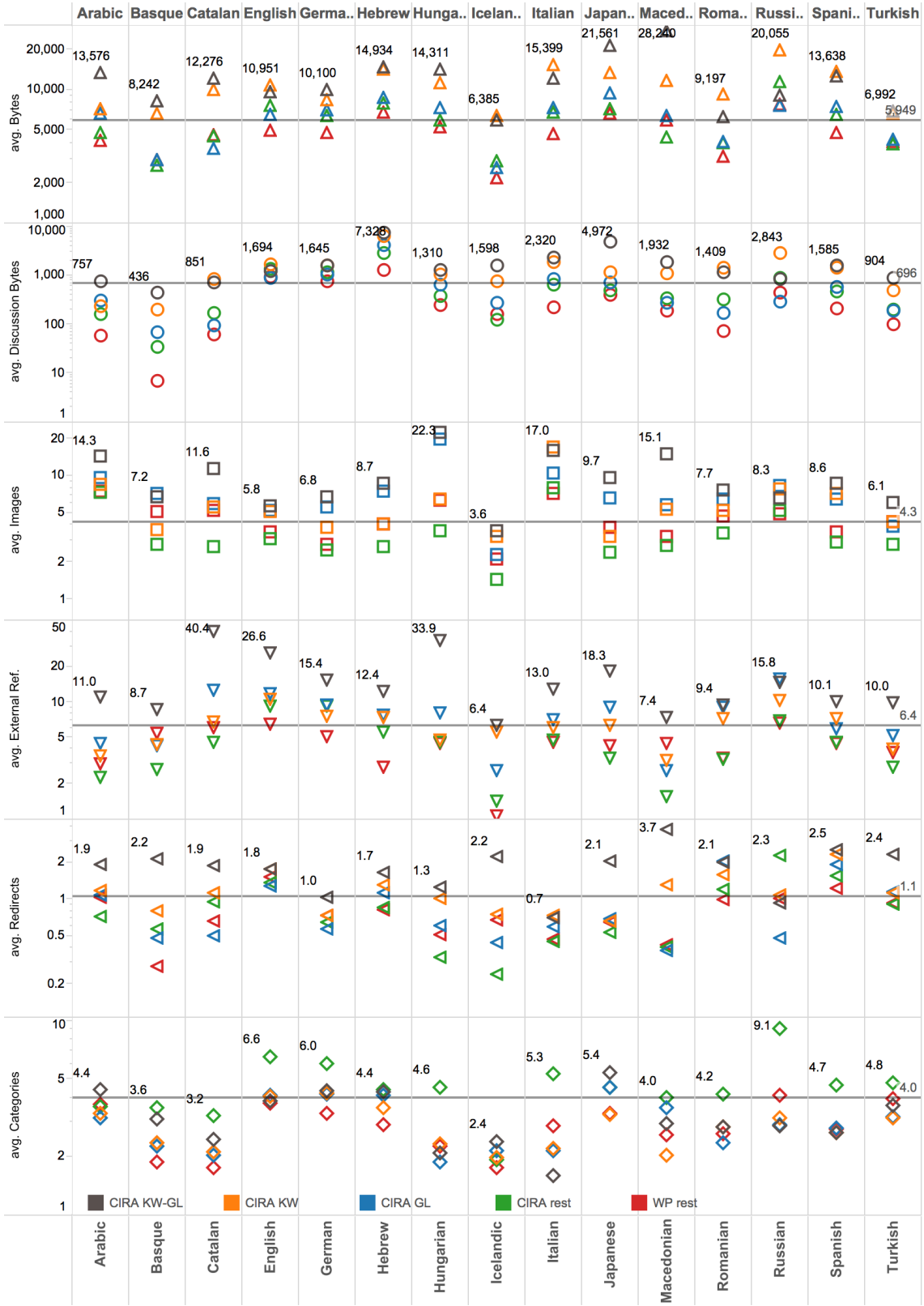


Figure 35. Average value for each article feature in each article type. *The horizontal line is the mean value for all articles, averaged over the 15 Wikipedia language editions.*

There are clear differences in the features between the different article types, and common patterns across languages also appear. For instance, those articles in CIRA which do not contain keywords in the title or geolocation tags (CIRA rest) are the ones with more categories in eleven languages out of fifteen. CIRA Geolocated articles have a higher value for the number of external references and number of images in many languages, only surpassed by CIRA Keywords-Geolocated. CIRA Keywords tend to have more Bytes in the text, Bytes in discussions, and also redirects, and likewise, it is sometimes exceeded by CIRA Keywords-Geolocated. Generally, the small minority of articles coincident in these two groups is always better developed than the other types of article. All in all, the articles in the different segments of CIRA tend to be developed in greater detail than the rest of articles contained in a Wikipedia language edition **(RQ3).**

Differences between languages are visible when comparing absolute values for each feature. For instance, Russian language presents articles in CIRA with more Bytes than the rest, and Hebrew have four times more Bytes in discussion pages than the others. Regarding images, Hungarian and Italian have a higher average than the rest for most of the segments of CIRA, and Spanish have more redirects for the different segments of CIRA. In fact, Redirects and External References have less variation across languages.

Discussion. The analysis of features showed, first and foremost, that besides the different features and topics of CIRA segments, they tend to be developed into greater detail than the rest of Wikipedia. This reinforces the importance of setting CIRA as a priority to translate into other language editions. It would have been interesting to compare the features from the same CIRA articles in different language editions, but it seems reasonable to think that usually the local editors might have better access to the information to create more developed articles – besides the motivation to do it. Therefore, finding CIRA content developed at greater detailed justifies bridging the culture gap in order to obtain higher quality articles.

The results also showed that even though all the CIRA segments tended to show a higher value in the different features than the rest of Wikipedia, each of them showed a trend towards a specific feature. In Section 7.3.2, I presented the CIRA article types, showing specific examples of them. The results from the feature analysis confirm that the article topic or information and its final features are closely related.

On the one hand, CIRA Keywords are most usually summaries of general topics related to the territory or to their inhabitants (e.g. 'History of England', 'American Football Players'), and accordingly, they require to be developed into a considerable extension.

On the other hand, CIRA Geolocated are very descriptive articles about cities or places in general within the territory – then, it is not surprising that they tend to be more referenced and contain more images than other types. The rest of CIRA stands out as having a higher number of categories. These are articles about a multitude of topics but present very specific meanings for the community, and, considering that CIRA is also thematically rich, editors add categories for all its possible relationships. For instance, a notable writer in a small city, could be categorised in several ways: as a writer, his writing genre, notable inhabitant of that city, the period of history when he was notable, political affiliations, among others.

8.1.5 Editor Engagement and Interlanguage Links in CIRA (RQ4)

In the previous sections, I focused on measuring the participation in CIRA and the article characteristics. Now I turn the attention again to the study of the cross-language content availability. In particular, I study the participation in CIRA articles in relation to the number of Interlanguage Links, to see whether engaging articles tend to be more shared.

Results. Figure 36 shows the average value for each article type (the main CIRA segments and the rest of Wikipedia) given the number of Interlanguage. It reveals that for equal number of Interlanguage links, the participation in CIRA is higher **(RQ4)**, i.e. articles receiving more participation tend to be more shared. Regarding the segments of CIRA, there is no common pattern: in nearly half of the languages, for the same number of Interlanguage links, the average number of edits for CIRA Keywords tend to be higher. While in the rest of languages, it is either CIRA Geolocated the one higher or the rest of CIRA. Furthermore, the graph shows that each language presents a different shape for CIRA. German or English language editions show a pattern for the different article types which initially rise, then stabilizes and finally grows constantly. These languages CIRA keep all degrees of cross-language availability and with similar levels of participation.

Other language editions like Basque show a big growth in the beginning with sparse lines and disappear after 100 Interlanguage Links, which means that their CIRA is essentially unknown in almost 200 languages. Instead, Japanese present a longer shape for the rest of CIRA than for the CIRA Geolocated, which means that most of their cities and places are less available than articles which could represent their cultural creations. The shape is an interesting indicator which shows that the amount of participation is related to the degree of availability, namely, the universality of its content. Figure 36 complements the tables, provided in Section 7.3.5, which presented the interlanguage cross-availability.

On the contrary, the shape from the non-CIRA articles is very similar for all languages - the same ups and downs appear. There are gaps at 30, 105 and 185 Interlanguage Links, approximately. These gaps can be explained by how article creation processes take place and involve different groups of languages (some articles are created consistently for different groups of Wikipedia language editions without a real community coordination). These are articles that describe all the events from a particular year, the countries or historical figures, which are copied and adapted. However, not all communities want or are able to establish this copying processes, which may explain the groups of languages.

Additionally, in the same graph, I depicted the cumulative function for the number of articles in relation to the number of Interlanguage Links. When the curve has a steep slope, it means that a majority of the articles tend to have few Interlanguage Links. Also, I drew a horizontal reference line for the first 0-4 Interlanguage Links, in order to highlight the exact percentage of articles in each Wikipedia language edition with these values. In fact, small languages like Basque or Macedonian with a small CIRA percentage of a 10-15% have also a small percentage of articles in this range (18-28%), implying that the rest is more universal, whereas bigger languages like Japanese or English with a high extent percentage of CIRA (40%), also have a high percentage of quite non-shared articles (68-80%), implying that their very shared or universal content does not occupy an important percentage.

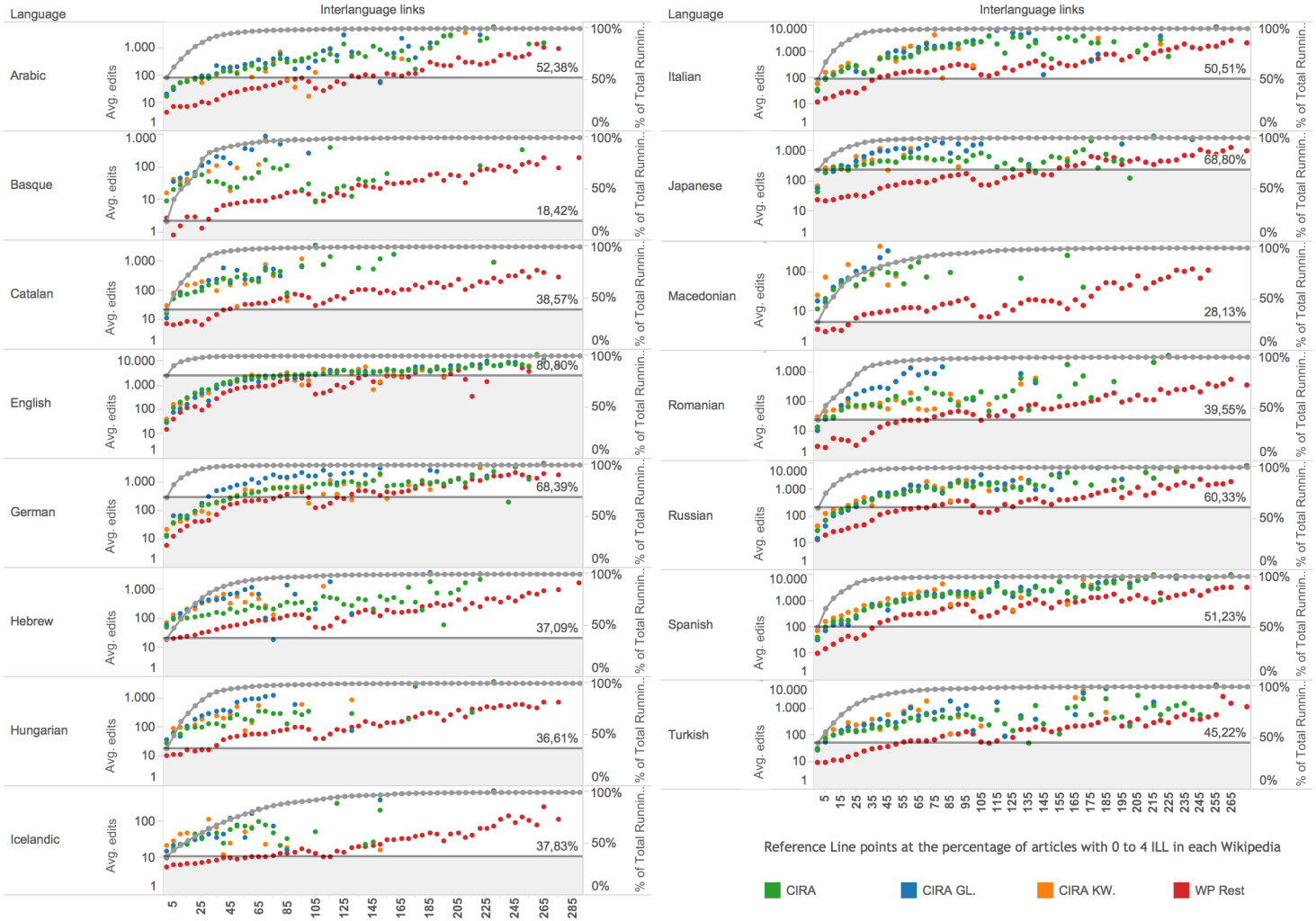


Figure 36. Average number of edits for each article type and by number of Interlanguage Links. *The grey dotted line is the cumulative function (% of Total Running) for the number of articles, and the horizontal reference line points at the percentage of articles with 0 to 4 Interlanguage Links.*

Discussion. There exists a clear relationship between the number of Interlanguage Links and participation. Previous research had shown that geographical articles related to a language edition which received more page views tended to be more shared across languages (Hecht, 2013). This pattern exists in terms of edits for all the different types of articles analysed, in the different segments of CIRA and the rest of Wikipedia. I consider this a positive finding because it means that the more shared are meanings from cultural identities - and participation is a form of sharing in the community -, the more they stand out visible from outside the community. However, the examination of the segments of CIRA in different languages showed that their cross-language availability is very diverse. In some languages, only a small minority of articles were shared for a group of 30, which is roughly the 10% of all the available languages in Wikipedia, and again it reminds of the importance of the problem of the culture gap

8.1.6 Prioritising the Culture Gap (RQ5)

Results. The most outstanding CIRA content in the eyes of the rest of Wikipedia language editions is the one which obtained most participation. This is a good starting point in order to find out which content from each language CIRA should be recommended to editors from other languages in order to import it. The Wikimedia Foundation provided a recommender to bridge the language gap (Wulczyn et al., 2016) which took into account numerous factors, including the content popularity as a main factor, in order to recommend articles (its name is GapFinder). However, I consider that in order to bridge the culture gap, articles should be recommended according to a different criterion. It is true that some concepts of a distant cultural identity may not very appealing to most of readers, but this does not imply that this content cannot be very useful or its learning or impressions may not be long-lasting to those who read it.

In order to evaluate more factors and establish a criterion in order to recommend articles from CIRA, I have correlated the number of Interlanguage Links to participation and article features from Section 8.1.4 using Spearman correlation, with results significant at p-value < 0.001 for all languages. I made the average for the correlation coefficients in order to obtain a result for all languages. I have found a strong correlation in the number of bots involved in an article (0.765) – this was expected since many Interlanguage links between two articles were introduced by bots. None of the article features provided a correlation coefficient higher than 0.3. The previously examined number of edits (0.5), and especially the number of editors (0.595) showed higher coefficients. After evaluating these different factors, I have selected the number of editors as the key variable. Besides being the participation feature which correlates best, it is also in line with the definition of cultural identity. Because if an article is edited by many editors it implies that they all agree at some extent on the importance of the meanings that are described.

Therefore, I propose recommending those articles that are created by the contributions from many editors, but still, have not been created in other language editions. With this aim, I propose building predictive model which takes into account the number of Interlanguage Links and the number of editors **(RQ5)**. For pragmatic purposes and considering different examined models, I choose a simple linear regression, even though the number of editors per article does not follow the assumption of a normal distribution.

As an example, Figure 37 depicts different lines according to linear regression models for CIRA and the rest of articles for the Catalan Wikipedia. This graphs shows the same information from Figure 36 but at an article-level of detail. As a parenthesis, for those who know about the Catalan Culture, many outliers in the upper part of the graph are important concepts (e.g. 'Jaume I el Conqueridor', 'Ramon Llull', etc.) and nonetheless, they are not shared more than in forty languages.

Table 18. Linear regression coefficients. *The linear regression models the relationship between number of editors and number of Interlanguage Links for each CIRA article. All coefficients are significant for a p-value < 0.0001.*

Language	Term	Value	Std. Err	t-value	p-value
Arabic	slope	1.94	0.010	193.19	< 0.0001
	intercept	10.23	0.112	90.55	< 0.0001
Basque	slope	0.62	0.006	101.54	< 0.0001
	intercept	7.2	0.066	108.76	< 0.0001
Catalan	slope	2.34	0.008	288.37	< 0.0001
	intercept	8.58	0.044	193.07	< 0.0001
English	slope	15.71	0.011	1323.31	< 0.0001
	intercept	15.69	0.085	183.83	< 0.0001
Hebrew	slope	1.79	0.015	112.74	< 0.0001
	intercept	24.30	0.207	117.08	< 0.0001
Hungarian	slope	1.89	0.011	162.82	< 0.0001
	intercept	12.81	0.091	140.62	< 0.0001
Icelandic	slope	0.80	0.008	90.83	< 0.0001
	intercept	5.15	0.069	74.47	< 0.0001
Italian	slope	4.39	0.013	333.97	< 0.0001
	intercept	18.69	0.128	145.91	< 0.0001
Japanese	slope	5.53	0.018	302.06	< 0.0001
	intercept	24.30	0.087	278.68	< 0.0001
Macedonian	slope	0.87	0.010	86.33	< 0.0001
	intercept	5.21	0.084	61.41	< 0.0001
Romanian	slope	1.23	0.008	153.25	< 0.0001
	intercept	8.29	0.069	120.01	< 0.0001
Russian	slope	5.04	0.010	496.91	< 0.0001
	intercept	8.80	0.064	137.14	< 0.0001
Spanish	slope	6.17	0.016	378.69	< 0.0001
	intercept	13.24	0.137	96.31	< 0.0001
Turkish	slope	2.79	0.021	137.77	< 0.0001
	intercept	16.89	0.206	81.87	< 0.0001

I have obtained the coefficients (Slope and Intercept) for the equation which models the relationship of the two variables (Number of editors = Slope * Number of Interlanguage links + Intercept). Then, I have calculated them for each language edition and considering CIRA as a whole (Table 18). Each coefficient is verified with a t-test, then the t-value is a statistic that measures the ratio between the coefficient and its standard error, being significant with a final p-value < 0.0001. With these coefficients, I introduce the real

number of editors from an article into the equation and isolate the Interlanguage links, obtaining the expected value according to the model. For example, to date, the article 'Cronologia de la repressió del català' has 1 Interlanguage Links and 154 editors participated in its creation. This article is a very developed piece about the repression the Catalan language suffered during its history until today. Therefore, it is a meaning which can be conveyed as essential for the cultural identity of many editors from the Catalan Wikipedia. Introducing the values into the equation (number of Interlanguage Links expected = (154 – 8.58)/2.34), I obtain the value which points out that the article should exist in at least 62 Wikipedia language editions.

In fact, some articles may already have many Interlanguage Links, but according to their number of editors they are still underrepresented in other languages. Therefore, the value obtained for the expected Interlanguage links is a useful reference which can be obtained for each article from CIRA. In order to assess the degree of priority that each article in CIRA should have, I propose an index by dividing the number of expected Interlanguage Links by the current Interlanguage Links (in case it has no Interlanguage Links, the expected value is divided by 1). In the following figures, I visualize the CIRA articles which should get priority to be translated for the Catalan and the English Wikipedia, according to the number of Interlanguage links they have and the value of the index. For the ease of understanding which articles get prioritised, I depict them in three columns according to the segment of CIRA. Not to clutter the figures, I only show the top 5 with highest value for the index for each Interlanguage Link. The rest of the figures for the 13 remaining Wikipedia language editions are located in Section 2.4 Appendix 2.

As an overview, the articles with a higher priority in each language edition have one commonality: they are edited by many editors of the community. They are about all sorts of topics found in CIRA. To put two extreme examples. In language editions like the Catalan or Basque, I see some articles explaining part of the sociological reality in which the editors live, associations, political parties or cultural creations. In others, like Italian or English, a good number of articles about Mass Media (TV Shows, series, music bands, etc.) are visible. The index provides a referential value, but still, other aspects could be taken into account in order to filter the articles and make more fine-grained detailed lists of articles to export-import (such as the number of references or other article features related to quality).

Discussion. I have provided the number of editors as discriminatory factor in order to prioritise those articles from the cultural identity representations which should be translated into more language editions. The number of editors is a suitable choice, considering its good correlation with the number of Interlanguage links at the same that it reminds of the Cultural Identity definition. I introduced these variables into a simple linear regression, which allowed computing the number of Interlanguage links an article would deserve. While this implementation is an initial approximation to a recommender system, I suggest incorporating this variable and CIRA into the article recommendation tool and the translator developed by the Wikimedia Foundation and tailor them to bridge the culture gap. I will expand this in the Chapter 10 of this thesis, in a section dedicated to propose design recommendations and bridge the culture gap.

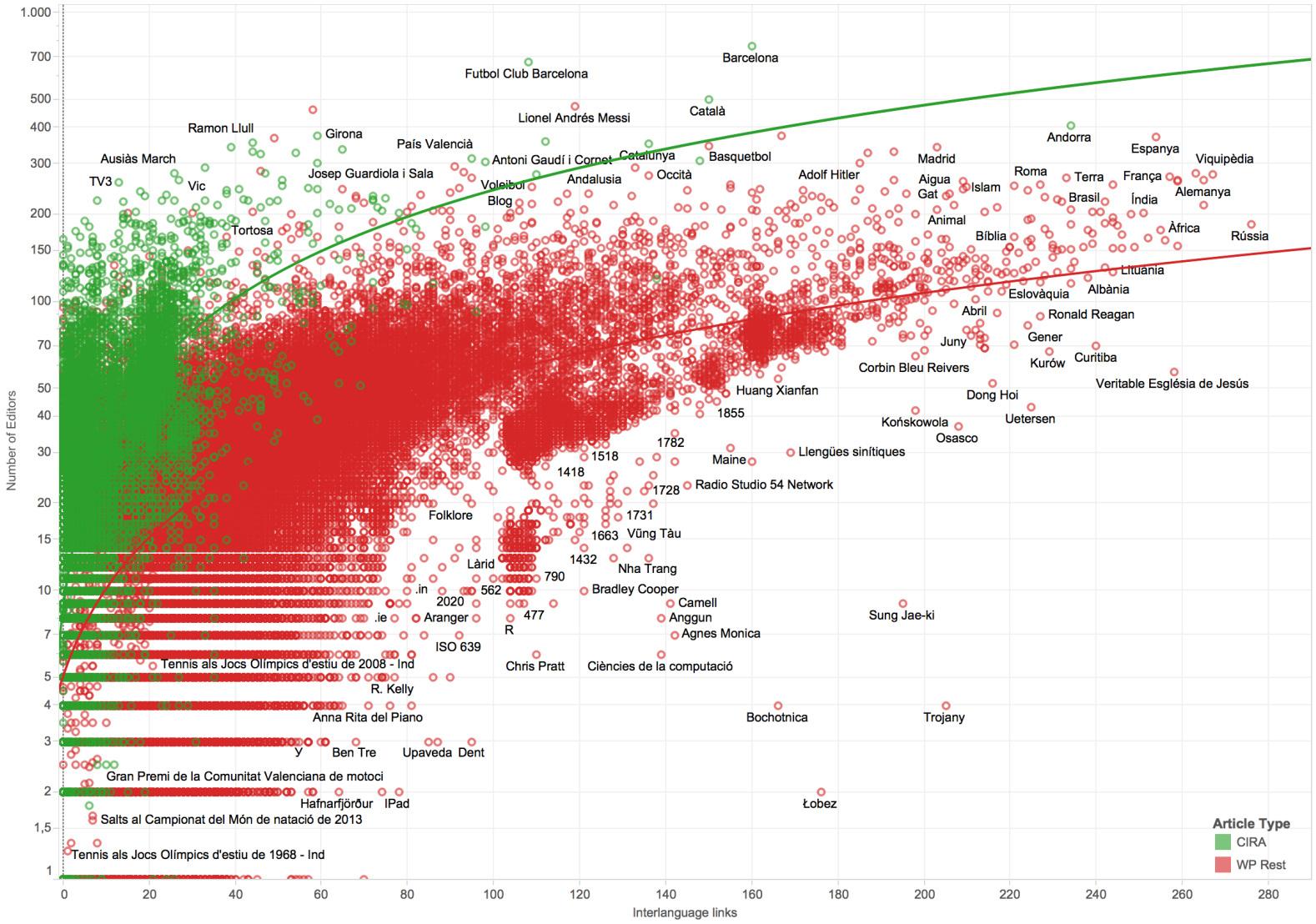


Figure 37. Articles by number of editors, number of interlanguage links, and coloured by article type (top layer is CIRA and then the rest of Wikipedia) in Catalan Wikipedia. *Trend lines created using simple linear regression.*

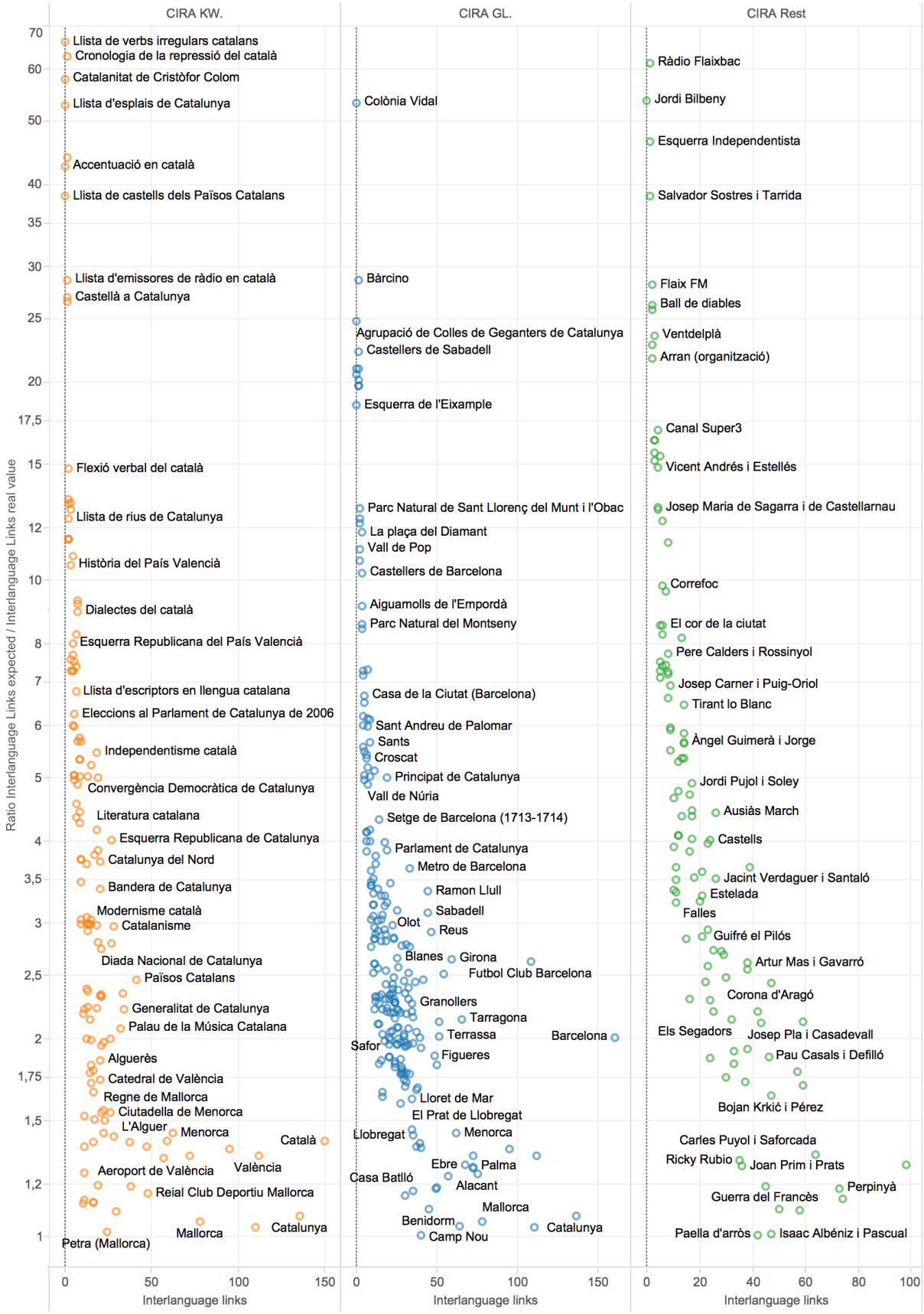


Figure 38. Catalan Wikipedia CIRA articles according to the priority ratio (Interlanguage Links expected / Interlanguage Links current value), by Interlanguage links current value and coloured by CIRA segment. *Only the top 5 articles for each Interlanguage Link are shown.*

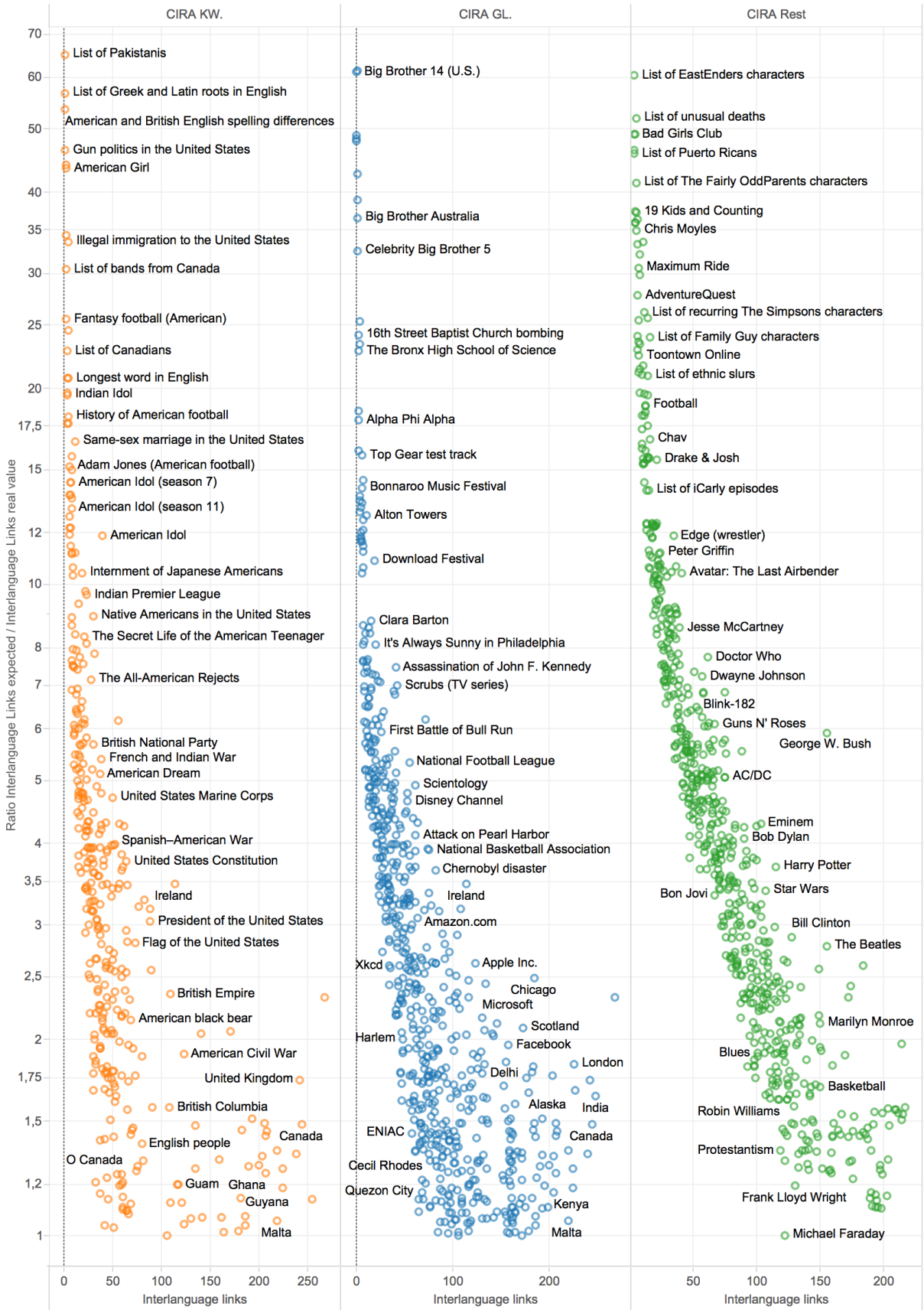


Figure 39. English Wikipedia CIRA articles according to the priority ratio (Interlanguage Links expected / Interlanguage Links current value), by Interlanguage links current value and coloured by CIRA segment. *Only the top 5 articles for each Interlanguage Link are shown.*

8.1.7 Summary of Results

This section provided new insights on editor participation in cultural identity representations. In the previous section I found that cultural identity representations occupy around a quarter of each Wikipedia language edition. The results from this section confirm that editors engage in this content beyond the first edit of article creation: the percentage in number of edits that occurred in CIRA is equivalent or even much higher than the percentage in number of articles. To be more confident, I analysed the different segments of CIRA at article level; the segment with both Keywords on title and Geolocation tags had the highest participation. This was followed by the groups of articles with these two features separately, the rest of CIRA and then the rest of Wikipedia. Hence, it can be stated that cultural identity representations reflect a higher participation than the rest of Wikipedia language edition content **(RQ1)**.

The fact that the participation accumulated in these articles is higher than in the rest of Wikipedia may indicate that editors are motivated to edit them. Yet, other factors like the demand by readers could also influence these content choices. In this sense, I studied the number of page views in the same segments of CIRA and I found that they present a higher value for the different segments of CIRA than for the rest of Wikipedia. However, editor participation in terms of edits present higher values than the readers' page views, which means that although a demand from readers makes more likely that cultural identities become relevant, editors already engage in editing this content with higher participation in all cases **(RQ2)**.

After knowing that cultural identity representations occupy a considerable extent and attract a high participation, it was interesting to investigate more about its articles' features. Generally, CIRA articles were developed into a greater detail than the articles from the rest of Wikipedia **(RQ3)**. CIRA is thematically diverse, and the different segments also show patterns in article characteristics. Generally, CIRA Geolocated articles stand for its number of images and external references, CIRA Keywords tended to be longer in its text and discussions, and the rest of CIRA had many more category affiliations than any other content type.

In order to understand which content from CIRA is mostly available across language editions, I studied the relationship between Interlanguage Links and Participation. Results showed that for all segments of CIRA and the Wikipedia language edition, the more work devoted to an article, the more possible it exists across languages **(RQ4)**. In other words, any CIRA segment has a higher participation in average at equal number of Interlanguage Links. However, some languages CIRA had a wide range of cross-language availability (e.g. English and German), while others were more restricted to lower number of language editions (e.g. Basque and Macedonian).

Further on, I examined the relationship between Interlanguage Links and article and participation features to discover that the best correlation is with the number of editors (0.59). Then, by taking the number of editors from CIRA, I created a linear regression model in order to find which articles are outstanding and should be a priority to translate into other languages **(RQ5)**. The output from the model showed that, for different CIRA, the articles to recommend deal with sociological aspects from their societies, popular places, mass media cultural products, among others. It is a good reflection of some topics which may be well-known but for some reason have not been created in other languages. All things considered, I can see there is an opportunity to work towards a cross-language exchanges and bridge the culture gap. Once measured the extent of CIRA, its creation over time and the high participation it attracts, an important challenge development is to provide mechanisms in order to make this content available in as many languages as possible.

8.2 Cultural Identities in Editors' Participation

In this section, I present the results for the analysis on the influence of cultural identity on editor participation. Therefore, I study aspects such as the editor types who contribute most to this content, their proportion of participation in it, and their contributions across language editions as a way of exporting the own cultural identity.

8.2.1 Research Questions

Content representing cultural identity attracts a higher participation than the rest of Wikipedia. In good measure, this is explained because editors take those actions which present opportunities to be congruent with their identities. In Wikipedia, there exist different types of editors depending on their position in the community – or outside of it, remaining anonymous. Their community identity features are completely different from each other, from the development of their User Page to contributing with files, along with the level of participation. I wonder if this crucial and central content like cultural identity representations, which also demonstrated to be popular in terms of readers' demand, is created by all types of editors.

RQ6. *Which kind of editors engage more in cultural identity representations?*

One aspect found in previous research is that editors are known to present different behaviours when they reach higher levels of participation. They adopt different and more specific activities that, at the same time, widen their focus into more different tasks than in their initial period after registering (Bryant et al., 2005). In other words, the motivation behind editors' actions evolves towards a different composition of factors. Even though contributing to cultural identity representations seems totally compatible with these other activities, its rellevance could differ depending on the level of participation. Therefore, I ask:

RQ7. *Which is the relationship between creating cultural identity representations and participation level?*

Highly participative editors and those with a functional role in a community are more likely to become multilingual editors. Since in other language editions editors can also choose the topics they want to contribute with, I assume that they may prefer identity-congruent imbued topics. Then, part of their contributions to other languages are likely to be from their primary language cultural identity representations. In other words, they may become exporters of their cultures. This leads to the following question:

RQ8. *Which editors do export their cultural identity representations in non-primary languages?*

Small language editions communities tend to have a larger percentage of multilingual editors. Exporting could be seen as an opportunity to mitigate the culture gap between Wikipedia language editions. Previous research found that multilingual editors tend to edit the same article across languages (Hale, 2014). Hence, multilingual editors could engage consistently in exporting their cultural identity representations in a similar way they edit them in their primary language. On the contrary, it could also be argued that a non-primary language is a very

different environment and that a foreign cultural identity may not be relevant to motivate contributions along time. Therefore, I ask:

RQ9. *Which is the relationship between exporting cultural identity representations and participation level in non-primary languages?*

In Section 7.3.6, I found that the most cross-language available topics in Cultural identity representations are geography and people. This means that from an outside perspective, any language edition finds most useful to cover some figures and places than traditions or history. However, this does not guarantee that these concepts are developed using the same perspective from their origin's language edition. Assuming that editors export their cultural identity representations, the particular content they choose might indicate what they consider a priority to show to the world or what they want to complete according to their points of view. The last question is:

RQ10. *What content from cultural identity representations is most exported to other Wikipedia language editions?*

8.2.2 Editor Types and Participation in CIRA (RQ6)

In order to know which kind of editors engage more in creating content representing their cultural identities, I first calculate the percentage in number of edits made by each editor type in CIRA and in Wikipedia. I only consider editors who have each language edition as their primary language of activity, as I consider these as the real community (as explained in Chapter 6). In addition, I separate the bots for a specific analysis because I am only interested in studying human behaviour.

Results. In Figure 40, the top subfigure reveals the percentage of edits made by bots to CIRA, which is in average between two and three times smaller than that in Wikipedia (in average in Wikipedia they represent the 34.9% of edits and in CIRA the 17.4%). This means that human editors account for a larger share than in the rest of Wikipedia. In Figure 40, the bottom subfigure compares the impact of each human editor type contribution in CIRA and in the whole Wikipedia language edition. It reveals that the only human editor types whose impact is bigger in CIRA than in the entire Wikipedia in all language editions is anonymous – administrators' impact is stable in many cases, and plain registered editors' impact is lower **(RQ6)**. Generally, anonymous editors have also a higher impact as a group than administrators (a part for smaller language editions like Basque, Icelandic or Macedonian).

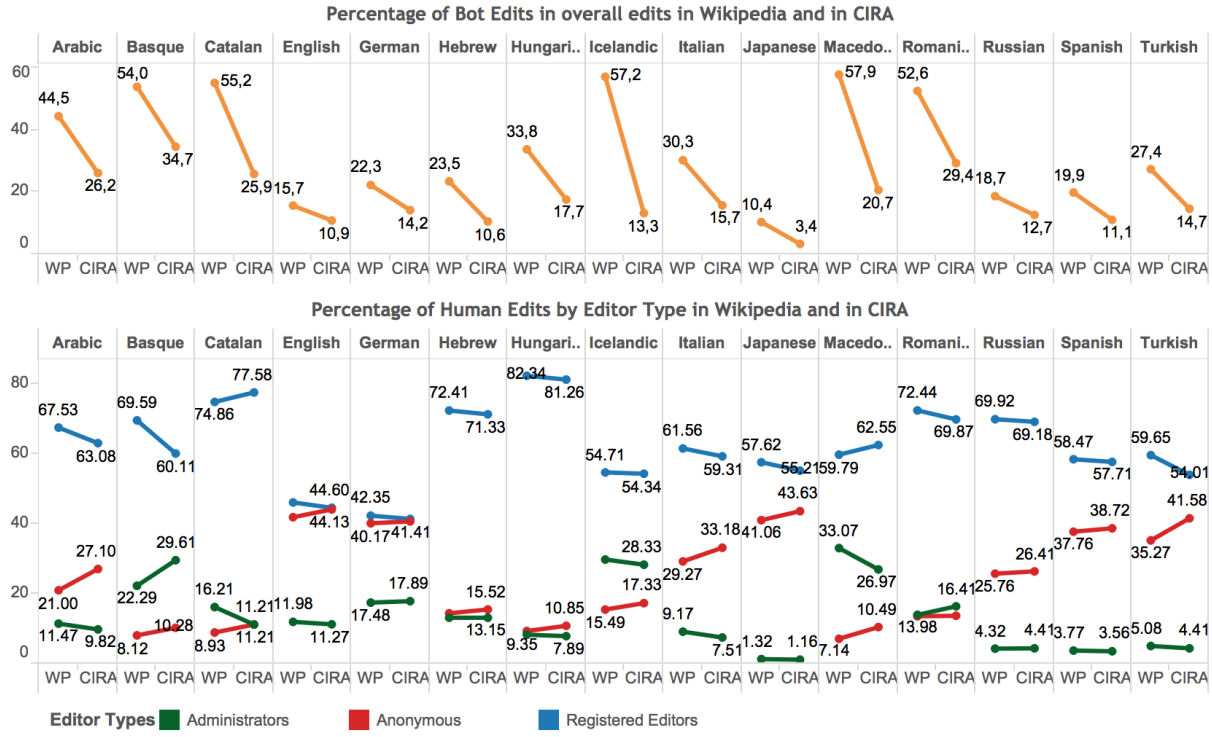


Figure 40. Percentage of contributions by editor type in the entire Wikipedia and in CIRA. *Top: Percentage of edits made by bots in Wikipedia and in CIRA. Bottom: Percentage of edits by editor type in Wikipedia and in CIRA, taking only into account edits made by humans.*

Discussion. The fact that bots' contribution represents a much smaller percentage in CIRA than in the entire Wikipedia could indicate that editors prefer editing in this content

themselves. However, in order to provide a more accurate explanation it would be necessary to analyse the tasks bots do and see if there exist other relationships with this type of content (some of them are: importing content into Wikipedia, spell checkers, revert changes to fight anti-vandalism, introducing Interlanguage Links, among others).

Regarding human editors, the different types have a similar impact in CIRA than in the entire Wikipedia. Besides the percentually higher impact of anonymous editors' contributions, the rest of editors do not indicate any important change. This means that, in communities like Wikipedia, whose role division and structure depend on different factors, cultural identity can be relevant to the different types of editors, and a further analyse at editor level is required.

8.2.3 Proportion of Participation in CIRA (RQ7)

To assess whether the differences observed between the impact of different user groups at the aggregated level are consistent also when considering individual editors separately (and not just produced by outliers), I compute the percentage of edits made in CIRA by each editor, and compare the distribution of this variable for different groups. This metric allows to explore whether they prioritise editing in CIRA over other type of content, hence it allows to assess the influence of cultural identity on participation at an editor level.

To focus on editors which are more likely to be local to a language edition, I furthermore only take into account for each language edition the primary editors (those registered who have more edits in that language edition than in the other language editions). Regarding the anonymous editors, it is not possible to study them individually since they are only identified by their IP numbers, and these values change. Some anonymous users may maintain their IP and other have a different IP at every connection, while sometimes a big institution has multiple computers using the same single IP, in a way that would make data unreliable. Then, the only partial solution is to calculate the percentage of anonymous edits in CIRA in relation to their edits in the whole Wikipedia as if they were a single editor.

a) Anonymous editors proportion of participation in CIRA

Results. In Figure 41, I have compared the percentage of anonymous edits in CIRA to the percentage of registered editors edits in CIRA, and to the percentage of CIRA articles. This rapidly shows, first, that the anonymous editors as a group have made a higher number of edits in CIRA than in other article types, second, that their focus towards CIRA is even more evident than with the whole group of registered editors. This confirms anonymous devoted work for CIRA and explains the bigger impact detected in the previous figure.

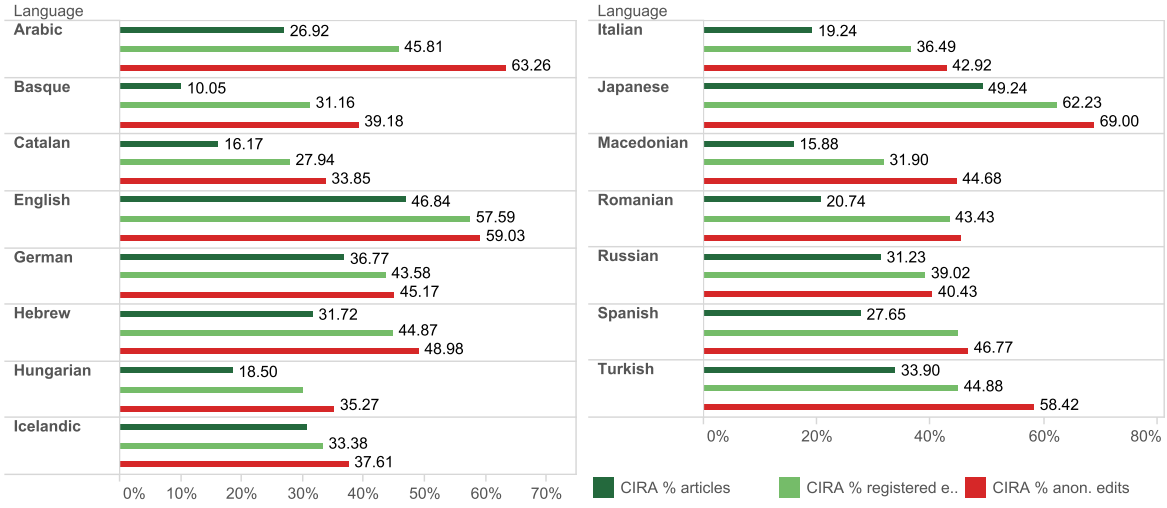


Figure 41. Percentage of anonymous and registered editors' edits in CIRA compared to the percentage of CIRA articles.

a) Administrators proportion of participation in CIRA

From now on, in order to focus on editors which are more likely to be primary or local to a language edition, I only take into account for each language edition the editors who have more edits in that language edition than in the other language editions. The distribution of the proportion of edits to CIRA is not normal, so I perform a non-parametric test (Mann-Whitney U test) to check whether there are significant differences between administrators and non-administrators. Namely, to see whether administrators have a higher proportion of edits in CIRA than the rest of registered editors.

Results. Table 19 shows the results which confirm that administrators devote a higher proportion of their contributions to CIRA for most of the languages, while in the Japanese, and although much less markedly in the English, we find the opposite result. Differences are not significant for the German, Hebrew, Romanian and Turkish Wikipedia. Administrators result may be explained in light of their tasks; they tend to interact preferentially with inexperienced users (Laniado & Tasso, 2011; Laniado, Tasso, Volkovich, & Kaltenbrunner, 2011) and are responsible of ensuring content quality (Suh et al., 2009). Therefore, their proportionally higher engagement in creating such a central and demanded content like cultural identity representations is consistent with their position and role in the Wikipedia project.

Table 19. Proportion of edits in CIRA: admins vs. non-admins. *The values are the Mann-Whitney U test results and mean ranks. Darker colours present higher mean ranks, indicating higher proportion of edits in CIRA. Significant results (p-value<0.05) are marked with a star.*

Proportion of Edits in CIRA

Languages	Admins Mean Rank	Non-Admins Mean Rank	Number of Admins	Number of Non-Admin	Z	p-value	
Arabic	98,579	79,031	32	158,036	-2.743	0.006	*
Basque	2,052	1,502	8	2,998	-2.045	0.041	*
Catalan	23,065	17,928	24	35,861	-2.747	0.006	*
English	3,478,475	3,627,281	1,307	7,253,201	-2.722	0.006	*
German	358,752	340,695	241	681,160	-1.499	0.134	
Hebrew	41,080	36,841	38	73,648	-1.307	0.191	
Hungarian	43,268	33,070	35	66,114	-3.604	0.000	*
Icelandic	2,037	1,527	20	3,039	-2.905	0.004	*
Italian	111,171	98,895	100	197,702	-2.311	0.021	*
Japanese	96,961	146,441	47	292,819	-4.225	0.000	*
Macedonian	3,289	2,339	13	4,670	-2.997	0.003	*
Romanian	24,479	21,555	18	43,094	-1.092	0.275	
Russian	204,136	158,606	86	317,150	-5.143	0.000	*
Spanish	402,207	329,177	71	658,297	-3.610	0.000	*
Turkish	79,883	70,886	26	141,749	-1.227	0.220	

b) Participation in CIRA within edit buckets

I further examine each editor's proportion of edits in CIRA. The identity-based motivation framework here would predict that all editors might present a percentage of their edits in CIRA, because depending on the situation they experience these meanings might appear as a salient choice. However, considering that an editor develops a community identity in Wikipedia and varies the activities and tasks with experience (Bryant et al., 2005), it is plausible to think that motivation associated with cultural identity could be less important when an editor becomes a very engaged member of the community, since the motivation composition is known to vary along time.

For this reason, I propose investigating the relationship between each editor's proportion of participation in CIRA and his overall participation in the community, by comparing them by the edit buckets defined in Section 5.3 and used in Chapter 6. Considering this, I depict the histograms with the percentage of edits in CIRA for each bucket, as they provide a descriptive picture of each part of the community. At the same time, a Kruskal-Wallis test has been performed to determine whether there appear significant differences in the editors' percentage of edits in CIRA by edit bucket, along with a pairwise comparison of groups through Dunn's test procedure with a Bonferroni adjustment for the p-value. Extended test results with all the values for the pairwise comparisons can be found in Table 42 in Appendix 3.

Results. Figure 42 and Figure 43 show the histograms and the Kruskal-Wallis generated mean ranks for all languages and edit buckets. Histograms show the distribution of editors in each edit bucket according to their proportion of edits in CIRA. The first bucket (1-100

edits) represents the most extensive group in the community: editors that register, make very few edits and in most cases, leave. All languages present for this bucket a bimodal distribution: a proportion of editors varying from 20% to 45% for different languages dedicate over 90% of their edits to CIRA, while the vast majority of the remaining editors perform less than 10% of their edits on them. Such bimodal distribution is due to the fact that most of these users edit just one or few articles. In the second bucket (101-1000 edits), editors who perform less than 10% of edits on CIRA are 40-30% (instead of the 60% from the first bucket). Bear in mind that an editor who has made it to this second bucket had overcome a learning process, and is growing an identity in the community. The third and fourth bucket (1001-5000 and 5001-10,000) shows a skewed distribution, where a minority of a 25% of editors engage in CIRA below the 10% of edits. In the fifth bucket (10,000+) there are only those very participative editors, with a distribution similar to a Gaussian. This means that the group of editors with this level of participation is more homogenous, and most of them have a proportion of edits in CIRA of 20-30%, depending on the language. For most languages, the last bucket shows a Gaussian distribution. This suggests that in the core of the community most editors cannot either focus only on CIRA or ignore it. One could suspect that the community dynamics affect this group of editors: for instance, by requiring some editors to revert some vandalic changes or by the need to update specific facts from the information related to their environment (the death of a celebrity or an event). Nonetheless, in this bucket there still exist few outliers with a high participation and a high percentage of edits in CIRA, which may indicate that the relationship with CIRA is a personal aspect.

All in all, according to Figure 42 and Figure 43, it is possible to see a relationship between editing more in Wikipedia and having a higher proportion of edits in CIRA exists mainly after the first two buckets. This seems consistent across all the languages and it can also be seen in the mean ranks obtained through the Kruskal-Wallis test results. The test mean ranks showed two different trends. For large languages like German, English, Arabic, Italian and Turkish, the proportion of edits in CIRA grows until the second or third bucket, and then it slightly decreases. While for small and medium languages like Macedonian, Basque, Icelandic, Catalan, Hungarian and Romanian, the proportion of edits in CIRA stabilizes in the fourth bucket and slightly decreases – namely, the core of the community engages proportionally more into editing CIRA in smaller languages than in bigger languages compared to the less active layers of the community.
However, the results from the pairwise comparison provided by the Dunn's test indicate that there only exist significant differences between the first bucket (and occasionally the second) and the other buckets. Hence, it is only possible to obtain two conclusions; first, that editors who reach more than 100 and 1000 edits reach a higher proportion of edits in CIRA, and second, that the core of the community tends to converge towards a Gaussian distribution, and therefore, they are generally involved in the creation and edition of this content (**RQ7**). This second conclusion is consistent with the higher proportion of edits in CIRA found for administrators, who are the most involved users in the community. In fact, it is important to mention that, again editors from Japanese language edition present a singular case which should be studied apart: it is the only one which shows a very strong proportion of edits in CIRA for the periphery editors and then it only decays (notice also the first bucket, CIRA is the gateway of entrance for a 46% of those with 1-100 edits). In this language edition, editors from the core of the community tend to be less participative in CIRA than the newcomers.

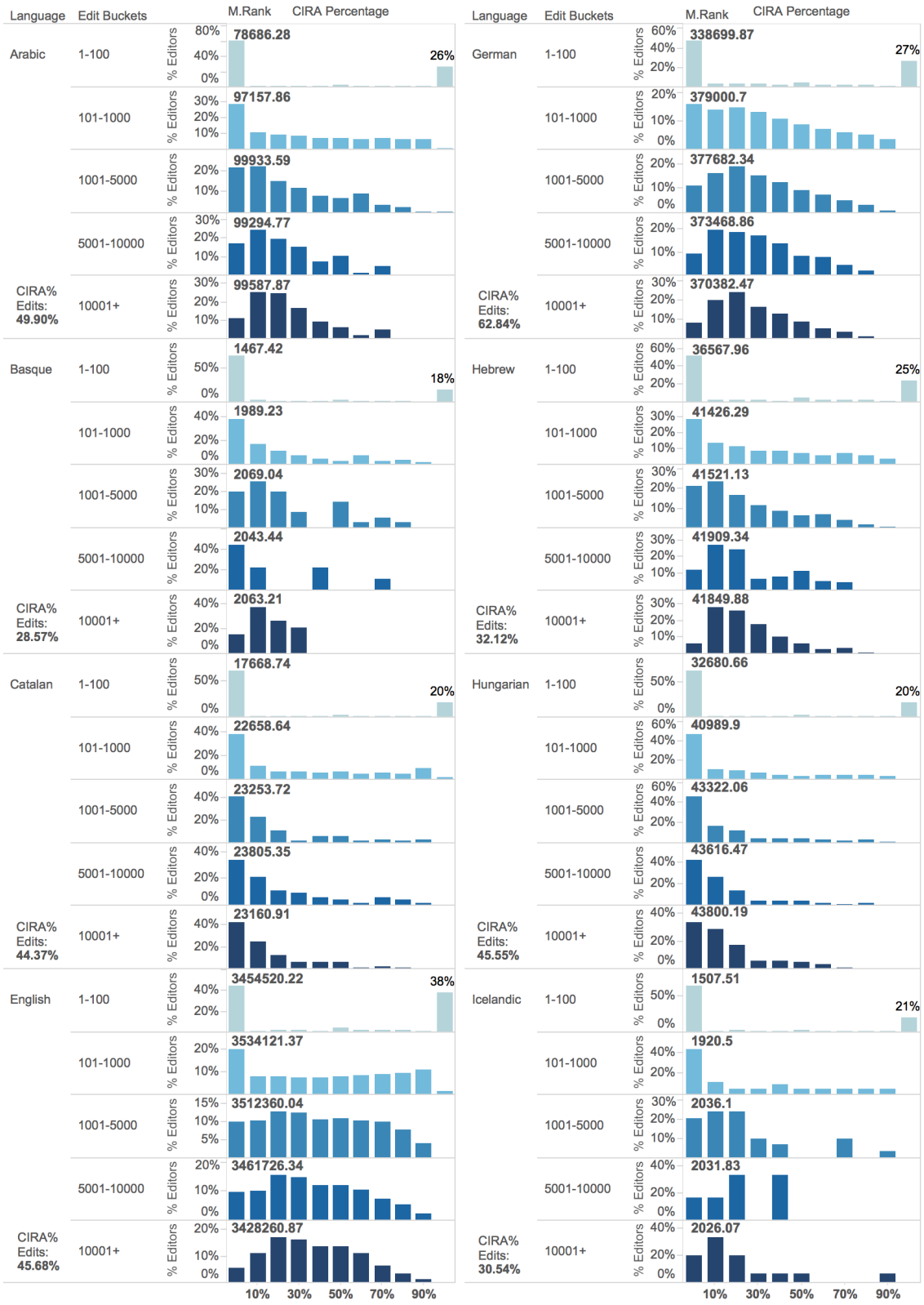


Figure 42. Histogram for each edit bucket showing the distribution of the proportion of edits made in CIRA. *Mean ranks for the proportion of editor edits in CIRA by Edit Bucket. Results of a Kruskal-Wallis test are statistically significant for all languages with p-values always lower than 0.001. Results for Arabic, Basque, Catalan, English, German, Hebrew, Hungarian, Icelandic.*

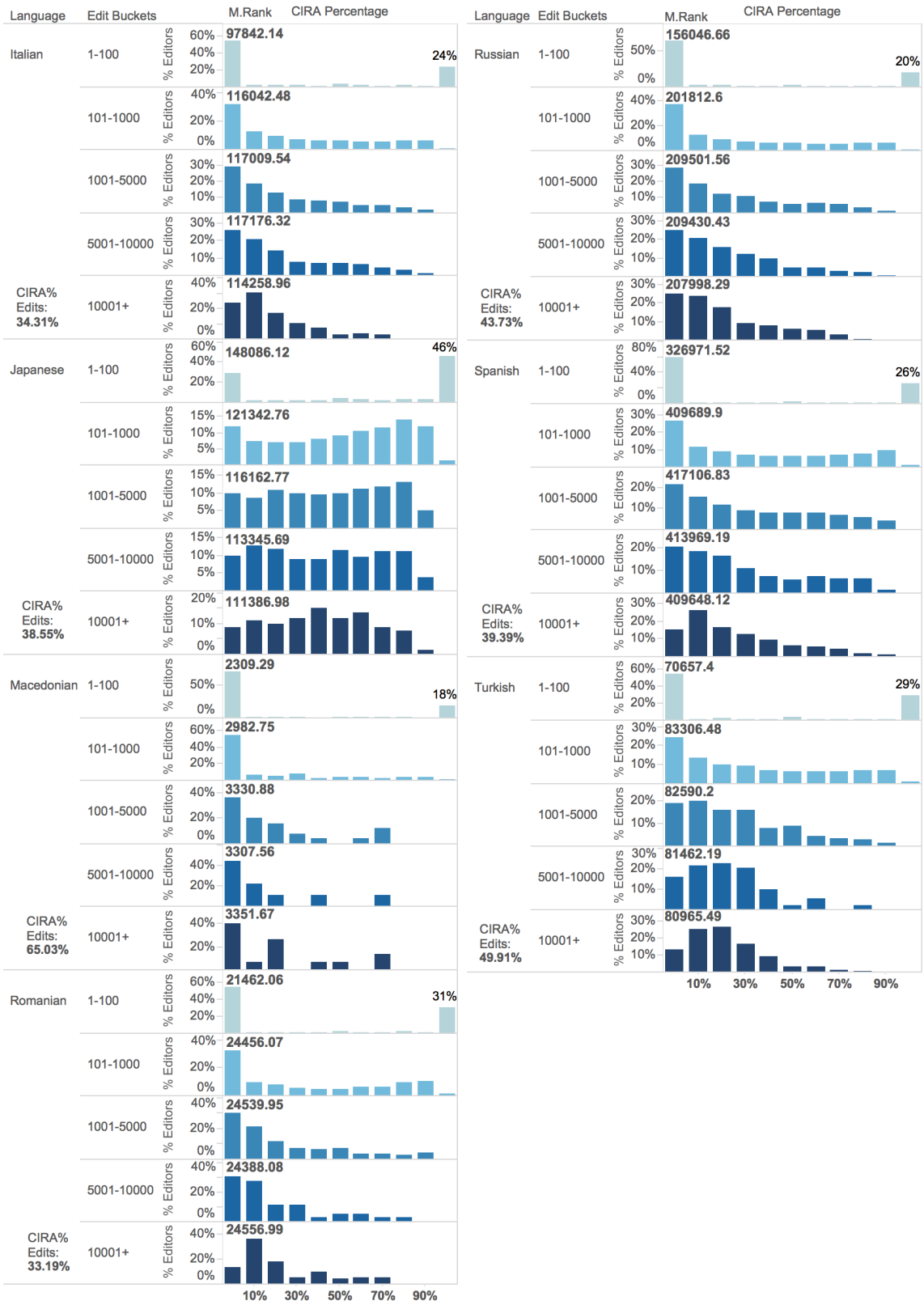


Figure 43. Histogram for each edit bucket showing the distribution of the proportion of edits made in CIRA. *Mean ranks for the proportion of editor edits in CIRA by Edit Bucket. Results of a Kruskal-Wallis test are statistically significant for all languages with p-values always lower than 0.001. Results for Italian, Japanese, Macedonian, Romanian, Russian, Spanish, Turkish.*

c) Proportion of participation in CIRA during the first 7 days

The very first week of activity has been considered especially meaningful to detect those editors who would become highly participative (Panciera et al., 2009), since in few days they already exhibit different characteristics than the rest of editors. Furthermore, in this period of the editor life, content choices may be free from other conditionings and community dynamics, and may provide clues on the motivations that attracted a user to participate in the project. Therefore, I expect that by studying this particular period of time it is possible to understand the influence of cultural identity in editors who later might become part of the core of the community. In this case, I use the Mann-Whitney test in order to assess whether the proportion of edits in CIRA during the first 7 days differs for administrators with respect to the other registered editors.

Results. While Table 19 showed that administrators tend to have a higher proportion of participation in CIRA than the rest of registered editors, Table 20 reveals that editors who will become administrators, in the beginning had already a higher degree of participation in CIRA. This suggests that editors that are more prone to get involved in the project and eventually develop a sense of belonging to the community (community identity) may be also especially sensitive to cultural identity in the first phase.

Table 20. Proportion of edits in CIRA during the first 7 days by administrator functional role. *The values are the Mann-Whitney U test results and mean ranks. Darker colours represent higher mean ranks, indicating higher proportion of edits to CIRA. Statistically significant results (p-value<0.05) are marked with a star.*

Proportion of Edits in CIRA in the first 7 days

Languages	Admins Mean Rank	Non-Admins Mean Rank	Number of Admins	Number of Non-Admi..	Z	p-value	
Arabic	95,603	79,031	32	158,036	-2.345	0.019	*
Basque	1,957	1,502	8	2,998	-1.729	0.084	
Catalan	23,051	17,928	24	35,837	-2.794	0.005	*
English	3,408,433	3,456,423	1,244	6,911,584	-0.907	0.364	
German	376,956	340,688	241	681,160	-3.057	0.002	*
Hebrew	38,706	36,843	38	73,648	-0.582	0.560	
Hungarian	41,465	33,071	35	66,114	-3.045	0.002	*
Icelandic	1,863	1,528	20	3,039	-1.941	0.052	
Italian	110,670	98,896	100	197,702	-2.262	0.024	*
Japanese	110,404	146,439	47	292,819	-3.122	0.002	*
Macedonian	3,170	2,340	13	4,670	-2.676	0.007	*
Romanian	26,196	21,555	18	43,094	-1.752	0.080	
Russian	196,782	158,608	86	317,150	-4.425	0.000	*
Spanish	387,798	329,178	71	658,297	-2.937	0.003	*
Turkish	79,624	70,886	26	141,749	-1.204	0.229	

d) Proportion of participation in CIRA over time

I now inspect how the influence of cultural identity varies over time for editors, focusing again for each user on the first 7 days of activity as compared to her/his overall activity. I perform a Sign Test, a statistical test employed to compare two related samples – in this case the same variable, i.e. a user's percentage of edits in CIRA, in two different moments of time. I consider separately editors having different participation levels, and we only consider editors who have been active for at least 6 months, as for shorter periods of activity the difference may typically be irrelevant. This time interval has been commonly used as threshold for considering that an editor has overcome the survival period (Ciampaglia & Taraborelli, 2015; Halfaker, Geiger, Morgan, & Riedl, 2013a).

Table 21. **Proportion of edits in CIRA during the first 7 days' vs final by edit bucket.** Number of editors whose overall percentage of edits in CIRA with respect to the first seven days is higher (CIRA % increases), lower (CIRA % decreases), equal (Ties). Darker colours represent higher values. Significant Sign Test results (p-value < 0.05) are marked with a star.

		Edit Buckets: # Editors				
Language	Variable	1-100	101-1000	1001-5000	5001-10000	10001+
Arabic	CIRA % Increases	3,114	842	167	50	66
	CIRA % Decreases	2,905	772	158	43	77
	Ties	5,583	125	2	0	0
	p-value	∗0.00733	0.08588	0.65721	0.53382	0.40301
Basque	CIRA % Increases	99	57	18	6	5
	CIRA % Decreases	73	46	13	3	14
	Ties	134	2	1	0	0
	p-value	0.05662	0.32446	0.47249	0.508	0.064
Catalan	CIRA % Increases	1,164	526	129	39	64
	CIRA % Decreases	1,052	413	108	26	58
	Ties	1,808	95	3	0	0
	p-value	∗0.01837	∗0.00025	0.19389	0.13664	0.65078
English	CIRA % Increases	280,268	59,741	10,965	2,230	2,638
	CIRA % Decreases	261,328	58,451	11,724	2,614	3,621
	Ties	340,893	3,290	49	5	1
	p-value	∗4.7e-146	∗0.00017	∗4.8e-07	∗3.7e-08	∗2.2e-35
German	CIRA % Increases	45,938	12,034	2,898	667	852
	CIRA % Decreases	41,433	10,190	2,595	702	996
	Ties	47,512	413	8	2	1
	p-value	∗1.9e-52	∗4.1e-35	∗4e-05	0.35813	∗0.00088
Hebrew	CIRA % Increases	3,362	1,312	291	71	138
	CIRA % Decreases	2,839	994	289	76	111
	Ties	3,106	101	3	0	0
	p-value	∗3e-11	∗4e-11	0.96687	0.74146	0.09941
Hungarian	CIRA % Increases	2,605	992	257	75	127
	CIRA % Decreases	2,091	588	133	56	72
	Ties	4,717	239	11	0	0
	p-value	∗7e-14	∗3.7e-24	∗4.7e-10	0.11579	∗0.00012
Icelandic	CIRA % Increases	100	38	18	1	10
	CIRA % Decreases	75	32	11	5	4
	Ties	176	8	0	0	0
	p-value	0.06964	0.55009	0.26521	0.219	0.18
Italian	CIRA % Increases	10,962	3,765	927	245	322
	CIRA % Decreases	9,237	3,023	729	195	314
	Ties	14,409	442	14	2	0
	p-value	∗7.2e-34	∗2.3e-19	∗1.2e-06	∗0.01949	0.78134
Japanese	CIRA % Increases	14,761	5,182	1,133	193	184
	CIRA % Decreases	15,666	5,840	1,354	282	259
	Ties	20,459	345	1	0	0
	p-value	∗2.1e-07	∗3.8e-10	∗1e-05	∗5e-05	∗0.00043
Macedonian	CIRA % Increases	106	55	16	4	8
	CIRA % Decreases	74	41	9	5	7
	Ties	228	18	0	0	0
	p-value	∗0.02085	0.18457	0.23	1	1
Romanian	CIRA % Increases	1,188	383	99	15	29
	CIRA % Decreases	1,211	356	77	20	42
	Ties	2,296	64	2	0	0
	p-value	0.65331	0.33885	0.11343	0.49896	0.1544
Russian	CIRA % Increases	14,340	5,582	1,400	368	491
	CIRA % Decreases	11,455	4,220	1,063	301	388
	Ties	20,871	870	17	0	0
	p-value	∗4.2e-72	∗5.3e-43	∗1.2e-11	∗0.01071	∗0.00058
Spanish	CIRA % Increases	20,405	5,512	1,199	247	375
	CIRA % Decreases	16,888	4,961	1,081	233	360
	Ties	31,715	619	13	0	0
	p-value	∗4.5e-74	∗7.6e-08	∗0.01427	0.55293	0.60557
Turkish	CIRA % Increases	4,502	783	200	47	64
	CIRA % Decreases	3,877	718	152	38	62
	Ties	7,270	72	0	0	0
	p-value	∗9.3e-12	0.09855	∗0.01224	0.38554	0.92901

Results. Table 21 shows that editors in the first bucket tend to keep the same proportion, probably also due to the bimodal distribution seen in Figure 42 and Figure 43: they typically edit just a few articles, corresponding in many cases to just a 0% or 100% percentage. For higher buckets, when a significant trend exists, the relative importance of CIRA tends to increase from the first week, suggesting that the influence of editors' cultural identity does not decrease as they get more involved into the project, and on the contrary some community effect seems to foster activity on this kind of content. The Japanese and the English Wikipedia, which showed also in the previous sections to follow an inverse pattern, are again an exception: these two language editions present a very large proportion of CIRA (almost half of the articles), and editors who get more and more involved in the community tend to decrease their relative participation on these articles.

Discussion. This section has shown that the entire community participated at some point in editing CIRA, from the core to the periphery, with special relevance of the groups of administrators and anonymous editors. As a group of editors, anonymous editors show a high proportion of edits in CIRA (an average of 46.6%), which is higher than CIRA percentage in terms of articles and of overall edits. This suggests that this editor type engages in CIRA for being more motivated by the content itself rather than by an interest for the entire Wikipedia project. At the same time, administrators tend to have a higher proportion of edits in CIRA than the rest of registered editors. This could be explained due to the responsibilities in the community and in the light of their maintenance tasks.

An examination of the editing community divided by levels of participation has provided some explanations on the influence of cultural identity on editors. Having an overall higher participation increases the chances to have a higher proportion of edits in CIRA. Editors with few edits are distributed as a bimodal, completely focused on CIRA or ignoring such content; editors with more than 100 edits tend to spread over different percentages, according to their personal preferences and cultural background; and the most participative tend to be distributed as a Gaussian, implying that they are a more homogenous group of editors who at some point edit CIRA, and find it hard to ignore or to dedicate solely to this type of content.

The homogeneity shown by the group of editors with the highest level of participation could indicate that editors tend to converge. In other words, that editing CIRA could also be encouraged or discouraged by other factors that can relate to community effects. In this sense, the analysis for the proportion of edits in CIRA during the first 7 days provided a scenario in which editors' edits may only reflect editors' content choices. Results have shown that editors who have become administrators, in the beginning also showed a higher proportion of edits in CIRA. This could mean that users that will develop a community identity and assume a special role in the community, at the beginning of their experience in Wikipedia were already more prone than others to participate by taking identity-congruent actions driven by their cultural identity.

In further analysis, I have compared the proportion of edits in CIRA during the first 7 days and in the end. It showed that a majority of editors tend to increase this proportion, also at higher levels of participation. These results complement the previous and are consistent with the examination of the edit buckets, providing a longitudinal perspective. As said, the main explanation for why editors tend to increase the proportion of edits in

CIRA could be found in community effects, considering that being involved in it implies editing some content to fulfil some role duties, surveilling new content (which could be created by anonymous, who have a predisposition towards CIRA), among others. These community effects probably accompany the influence of identity-based motivation, which does not seem to disappear coherently with the idea that cultural identities are stable constructs related to the self-concept. These are not likely to change over time for the majority of people, neither for editors, especially during the time they are engaged in Wikipedia.

Regarding the different languages, they present the same pattern in all the experiments - with exception of the Japanese, which showed to always follow an opposite trend. To understand differences between cultures it would be necessary to approach the editors with a different methodology to understand if their cultural values transmit the notion of expansion or at least proselytism. Previous research (Pfeil et al., 2006) has found that cultural values underlie and repercute in some behaviours (power distance, uncertainty avoidance, among others). The identity-based motivation framework explains that when any identity appears relevant to a particular context, its underlying values influence how to make sense of the world and some particular behaviours are more easily triggered - which Oyserman & Destin (2010) refers to as procedural-readiness.

8.2.4 CIRA Exporters Among the Editor Types (RQ8)

a) Community composition (exporters)

In the preceding analysis, I analysed editor participation in content representing their identities in their primary language. Since the majority of the core of the community (both with functional roles and very participative) are multilingual editors, I conjecture that part of their contributions is dedicated to represent their cultural identities in non-primary languages (in other words, exporting their local knowledge). Therefore, editors who edit their primary language CIRA in non-primary languages as 'exporters', because they are contributing to create their cultural identity representations in another language.

In fact, I consider that the activity of exporting CIRA exemplifies how the influence of cultural identity can drive editors to engage in identity-congruent acts in Wikipedia, in particular in a scenario where some of the effects previously assumed (both from the community dynamics or the readers' demand) probably do not exert the same influence. Exporting CIRA can be considered an activity only performed by editors who feel identity-congruent with the meanings of their cultural identity in order to overcome the language barriers, among others. Therefore, by understanding who are the exporters it is possible to understand better the influence of cultural identity on editor participation.

Results. In order to understand who engages in exporting, I calculated the percentage of exporters among primary multilingual editors by editor type (functional roles) (Table 22). It is necessary to bear in mind that almost the totally of functional roles are multilingual editors: starting from the Security Force and Administrators, followed by Quality Patrol, and Production Force. Hence, the high proportion of exporters among these functional roles imply that almost the totality of the core of the community engages in this activity.

Interestingly, the more central the functional role the higher the proportion of exporters - as if exporting would also indicate the centrality in the community like the community identity features did. Regarding registered editors with no functional role, around a 39.41% of primary multilinguals are exporters (median 40.02%, standard deviation 4.89%), and only a 3% (median 2.41%, standard deviation 1.85%) when considering all editors from a language edition (primary and non-primary).

Table 22. Percentage of CIRA exporters among primary multilingual editors by editor types.

Languages	Security Force	Admin.	Quality Patrol	Product. Force	Registered Editors
Arabic	100	82.14	79.09	64.76	38.83
Basque	0	100.00	0	0	40.53
Catalan	100	95.00	75	87.34	38.13
English	100	100.00	90.23	86.56	41.53
German	100	98.25	0	68.71	36.25
Hebrew	100	96.97	92.96	72.32	40.83
Hungarian	100	93.33	0	69.19	32.23
Icelandic	0	100.00	0	0	37.22
Italian	100	98.91	100	85.38	40.02
Japanese	100	94.59	100	0	51.2
Macedonian	0	100.00	81.82	61.54	31.27
Romanian	0	82.35	0	81.06	42.56
Russian	100	93.42	72.56	65.94	34.95
Spanish	100	100.00	87.18	89.56	42.52
Turkish	100	100.00	87.97	65.77	43.19

b) Comparing exporters to the rest of the community

In order to verify that the exporting activity is based on identity-congruent actions, I propose comparing the exporters to the rest of editors from each community in terms of their proportion of edits in CIRA in their primary language community. This way, if exporters show a higher proportion of edits in CIRA would confirm that these editors are more predisposed to engage in creating their cultural identities. In order to test this, I compare the proportion of edits in CIRA of different groups of editors: exporters, and also those primary multilinguals who are not exporters, primary non-multilingual and non-primary, using a Kruskal-Wallis test. Notice that I included the 'non-primary', a user type who is registered in a community but is most participative in another, as a control group. This group of editors is expected not to share the elements of the cultural identity and engage in a lower proportion in their representation.

Results. Table 23 shows there exist differences between the different editors according to the language affiliation and the exporter condition. The most common pattern in eleven out of fifteen languages show that exporters have the highest proportion of edits in CIRA in their primary language, followed by primary non-multilingual, primary multilingual and non-primary. In other three languages (Arabic, Hungarian and Russian) the primary multilingual was in second position after exporter. While the Japanese shows the primary as the one with the highest proportion of participation in CIRA, which shows again that this language follows a different pattern.

Table 23. Proportion of edits in CIRA by language affiliation (exporter, primary multilingual, primary non-multilingual and non-primary) to a Wikipedia language edition. *The values are the Kruskal-Wallis test results and mean ranks. Darker colours represent higher mean ranks, indicating higher proportion of edits to CIRA. All results are significant with a p-value < 0.001.*

Proportion of Edits in CIRA by Language Affiliation

Language	Exporter	Primary Multilingual	Primary	Non-Primary
Arabic	118,120	87,031	86,914	83,588
Basque	7,419	4,643	5,097	4,314
Catalan	41,911	27,687	28,954	24,696
English	4,128,765	2,789,884	3,664,346	2,984,976
German	472,594	364,700	392,640	304,870
Hebrew	54,331	41,735	42,801	35,967
Hungarian	56,371	41,599	40,547	33,733
Icelandic	5,333	3,865	3,866	3,235
Italian	156,760	111,838	119,629	94,841
Japanese	142,516	101,135	163,713	93,349
Macedonian	7,226	4,764	5,015	4,352
Romanian	39,303	27,974	29,644	23,984
Russian	247,777	178,558	177,118	160,732
Spanish	492,749	348,292	358,986	336,505
Turkish	103,905	77,199	80,241	68,743

In order to verify whether the differences between each specific group are significant, I conducted a Dunn's test with a Bonferroni adjustment. Results are not shown as the pairwise comparisons results are significant with a p-value < 0.001 for all cases. These results confirm that exporters are indeed more engaged in participating in their cultural identities, and that this may either occur in their primary language or in non-primary language. In a way, they also reinforce the idea that the core of the community, who are multilingual editors and exporters in majority, have a higher proportion of edits in CIRA in their primary language. The lower results for non-primary editors are not surprising, although the extent of the difference from exporters varies a lot depending on the language.

c) Participation in CIRA in non-primary languages by editor types

Finally, once known the editors who involve engage exporting CIRA, it would be interesting to find out their impact. In the analysis of multilingualism from Section 6.3.4, it has become evident that the percentage of edits multilingual editors dedicated to their non-primary language was rather small compared to the total of their edits (1-3% for editors higher than 1001 edits and 1% in average for functional roles). Hence, it would be reasonable to expect exporters to have a rather small impact. To check it, for each language CIRA, I have calculated the sum of the number of edits by each editor type in all the other languages in which there were equivalent articles.

Results. Figure 44 shows the percentage of edits made by each editor type averaged for all languages, in Wikipedia, in CIRA and in CIRA in non-primary languages. This graph is similar to Figure 40 from Section 8.2.2 but it adds the last column with CIRA in non-primary languages. The figure shows again that CIRA tends to be more edited by human

editors (anonymous and registered) than bot editors than the entire Wikipedia. However, when CIRA is edited in a different language than its associated language, the percentage of human edits tends to decrease and bots overall impact is higher but not comparable to the whole Wikipedia.

Regarding the impact in terms of percentage of edits made by exporters is small: a 0.87% of the overall edits required to create a CIRA in their non-associated languages. Some of the languages whose exporters have the most impact in creating CIRA in non-primary languages are: Russian (2.87%), Catalan (1.87%), Spanish (1.30%) and Italian (1.03%), while those with least are Basque (0.29%), Icelandic (0.37%) and Hebrew (0.37%). In a CIRA in a non-primary language, the percentage of the contributions made by anonymous and primary editors is very similar to the rest of Wikipedia.

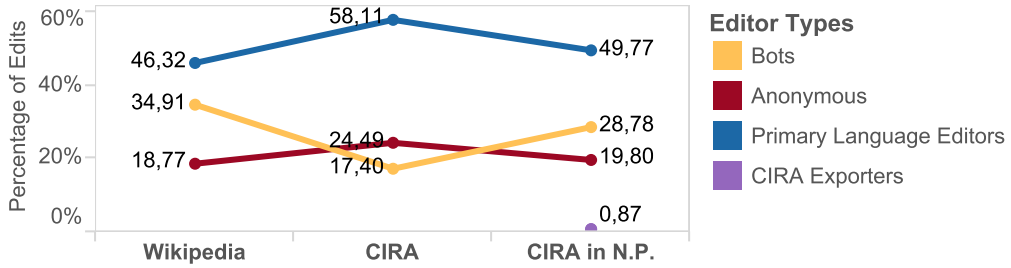


Figure 44. 15 Percentage of edits by editor type in the entire Wikipedia, CIRA and own language CIRA in the other languages (average over the 15 Wikipedia editions). *Multilingual CIRA Exporters account only for a 0.87% of the edits of their primary language CIRA in the other languages.*

Discussion. These results have shown that being exporter is a common characteristic to those editors at the core of the community with a functional role. Interestingly, the more central in terms of a role, the higher the percentage of exporters **(RQ8)**. This could suggest that editors very involved in the project, who had built a community identity, also engage in identity-congruent actions. In fact, these results are consistent with those from the previous section which showed that editors who become very participative have higher proportion of edits in CIRA, either during the first seven days after registering and in the end. Considering the rest of the community, around a third of registered multilingual editors with no functional role also performed edits to export their CIRA. The CIRA exporting could be seen as the opportunity for editors to bridge their culture gap. By now, the impact of editor exportation in non-primary languages is very small compared to the total number of edits these articles receive. This could be tackled by stablishing cross-language editing routines and coordination between different communities' members. These are aspects to consider when designing new interfaces and tools.

8.2.5 Proportion of Participation in CIRA (in Other Languages) (RQ9)

a) Edit buckets in non-primary languages

Once characterised the editors who engage in exporting CIRA, a further analysis of their participation in non-primary languages can explain how sporadic or regular the exporting activity is. In fact, the study of the proportion of participation in cultural identity representations in non-primary languages would confirm if, in a similar way to their primary language, this other context makes their identities salient and foster their contributions into the representation of their cultural identities.

This said, I propose exploring the relationship between exporting and participating in non-primary languages (multilingual editing) to see if this activity decreases along with increasing the overall multilingual participation. In a similar way to Section 8.2.3, I have compared the proportion of editors' edits in CIRA in non-primary languages of the different edit buckets (this time in non-primary languages) running a Mann-Whitney U test, which is employed to determine if there are statistical differences between two independent samples in a dependent variable. In addition, I also provided the histogram for each language edition in Figure 45. Instead of using the five buckets like in previous analyses, I reduced the buckets to two: 1-100, and more than 100 edits, given that editors are less and there is not a distribution variation such as in the primary language distributions seen before.

Results. Looking at the mean ranks for each of the two edit buckets, in all languages the first bucket (1-100 edits) has a value almost twice lower than the second and final bucket (101+ edits). According to a Mann-Whitney U test such differences are significant with p-values < 0.001. This can be explained due to the low percentage of exporters among the primary multilinguals, since an important number of multilingual editors in their first edits do not export (between half and two thirds of the primary multilingual editors).

The histograms shown in Figure 45 elaborate this insight. The first bucket reveals that an average of 14.73% of primary multilingual editors in their first edits dedicate fully to export their CIRA, while a 68.66% in average have not exported or present less than the 10% of their edits in non-primary languages. Previous section showed that the percentage of exporters among simple registered editors is a 39% of primary multilingual editors, which is consistent with these results.

The second and final aggregated bucket depicts a skewed distribution. This presents as the highest value an average percentage of 37% of the editors dedicating a 0-10% of their edits to CIRA in non-primary languages. Considering these results and that in Section 6.3.4 analyses showed that some editors reach a high number of edits in their non-primary languages, it implies that editing more in non-primary languages does not have an important over reducing the percentage of edits in CIRA, in other words, exporting less CIRA **(RQ9).**

Figure 45. Histogram showing the distribution of the proportion of edits made in CIRA in non-primary languages. *Edit buckets: 1-100 edits and 101+ edits in non-primary languages. Mean ranks for the proportion of edits in CIRA in non-primary languages by edit bucket. Results of a Kruskal-Wallis test are statistically significant for all languages (p-values < 0.001).*

b) Comparing primary and non-primary language CIRA editing

Previous analyses confirmed that editors show a proportion of edits in CIRA in their non-primary languages similar to the one in their primary languages – even at high levels of participation in non-primary languages. Hence, I assume that non-primary languages present a context where editors edit content representing their cultural identities as identity-congruent choices. Yet, it is not known in which measure this second scenario makes cultural identities as relevant as in their primary language, where the rest of editors share a good portion of same cultural identity meanings. I propose comparing editors' proportion of edits in CIRA in their primary language and in their non-primary language.

In previous sections, I compared the proportion of edits in CIRA considering all edits in the editor primary language edition (namely, including also those employed to modify a guideline or to communicate with another editor). In this case, I want to compare exclusively content choices in the two different scenarios. For this, I have taken the group of exporters from each language edition, and I have computed the percentage of edits in CIRA taking only into account the edits in articles. To compare the two scenarios, I have used non-parametric Sign Test, which is often employed to determine whether there is a statistical difference between two paired observations – e.g. the same practice in two scenarios by each person.

Results. According to the test results, editors usually have a smaller percentage of edits devoted to CIRA in non-primary languages than in their primary language (Table 24). Results are significant for 11 out 15 languages – three of the non significant are smaller languages which imply a small sample of exporter editors.

Table 24. Number of editors by proportion of edits in CIRA: primary vs non-primary languages. *Sign Test results shows the number of times either CIRA % Non-Primary or CIRA % Primary Language are higher, and the number of Ties (CIRA % Primary = CIRA % Non-Primary). The differences are statistically significant for 11 languages.*

Editors by Proportion of Edits in CIRA: Primary vs Non-Primary Languages

Language	CIRA % Non-Primary	CIRA % Primary Lang.	Ties	p-value
Arabic	1,323	1,498	649	0.001 *
Basque	111	105	38	0.734
Catalan	795	878	415	0.045 *
English	26,044	49,110	8,864	0.000 *
German	6,395	12,796	2,190	0.000 *
Hebrew	806	1,446	241	0.000 *
Hungarian	671	691	142	0.607
Icelandic	63	69	17	0.663
Italian	2,725	3,590	756	0.000 *
Japanese	1,816	2,252	452	0.000 *
Macedonian	85	85	33	1.000
Romanian	551	764	395	0.000 *
Russian	4,813	5,479	910	0.000 *
Spanish	6,103	7,953	3,110	0.000 *
Turkish	1,299	1,604	584	0.000 *

Discussion. In terms of exportation, multilingual editors who have only made few edits in non-primary languages show a different behaviour than those who have achieved more than 100 edits. The first are sporadic exporters, while those with a higher number of edits in non-primary languages have incorporated multilingualism as an activity – not that they spread their edits over multiple languages, but they have a considerable quantity of edits which implies a certain regularity. In other words, they export their primary language CIRA consistently, even though they are in a very different scenario.

For instance, factors related to the community identity such as the functional roles or article patrolling do not affect in non-primary languages in the same level of intensity when they do. In other words, the community effects (such as patrolling articles to revert vandalic changes or coordination) do not encourage or discourage participating in the own cultural identity representations in non-primary languages. In addition, the percentage of articles from the primary language CIRA in a non-primary language is also smaller (as seen in Section 7.3.6). Perhaps because of this, the distribution of exporters with high levels of participation do not tend into the Gaussian in which converged editors in the primary language, and instead, it shows a skewed shape.

When comparing the proportion of edits in CIRA in the primary language and in the primary language, the first tends to be significantly higher than the second. This can be considered expected, since as said, it is a very different scenario. One of the reasons can be found in the editors' lack of skills for editing about complex topics in that language (Kim et al., 2016). In addition, it must be reminded that many articles related to the editors' CIRA are not created yet, the existing CIRA occupies a much smaller percentage, and creating new articles implies a greater effort than expanding existing ones. It is important to notice that the rest of the editors in the scenario do not share the same cultural identity, which could also influence the interpretation of new articles as not notable enough to remain in that Wikipedia language edition.

8.2.6 CIRA Exported Articles (RQ10)

By definition, exported articles from a language are CIRA articles edited in other languages by primary editors from that language. These articles can be either created by primary editors from those languages or by exporters. Considering this, I propose measuring two metrics to understand the interactions with exported articles: the number of times an exported article has been created in any language edition by exporters, and the number of exporters who edited an exported article.

In Section 7.4.6, I considered the number of editors as a valuable factor in order to determine which articles from CIRA deserved being translated into other languages – in fact, this metric showed a higher correlation with the number of Interlanguage links than any other variable. Since exported articles have been specifically chosen by exporters, I propose using the Spearman correlation to find a relationship between the engagement metrics (number of edits, editors, and page views) of the article in its associated language edition and the two new and above-mentioned metrics metrics for the article as exported to other language editions.

Results. The number of exporters who edited an exported article is the variable which correlates best with the rest of engagement metrics, in particular, with the number of editors who edited the same article in the primary language (0.591), and the number of edits (0.562) and page views (0.446) in the same article (results significant at p-value<0.001). It is interesting to remark that the best correlation happens between number of editors in both primary and non-primary language, since it validates the previous election of number of editors as a good indicator to prioritise articles to bridge the culture gap in Section 8.1.6. In fact, it means that CIRA articles attracting the attention of editors in their primary language are more likely to be exported by more editors than those with many edits or page views.

This being said, Table 25 shows a sample of the exported concepts for each Wikipedia language edition; they are the ten exported articles in which intervened the largest number of exporters. In addition, in order to visualize a bigger sample, I propose using the tag cloud visualization for each of the fifteen language editions both the articles which most editors exported and the articles which have been most created by exporters. Figure 46 and Figure 47 present the CIRA exported articles both with most exporters (top) and most times created by exported (bottom) from the Catalan and the English Wikipedia. The rest of the figures for the 13 remaining Wikipedia language editions are located in Section 2.5 Appendix 2.

In general terms, the exported articles in which most exporters intervened tend to be important concepts such as populated cities, historical figures, the language name, etcetera. Instead, the articles which are most created by exporters are sometimes anecdotic choices, with the simple aim of advertising a concept in other languages (it can be a brand, a company, a celebrity or a little city) **(RQ10)**.

Discussion. Every language exported concepts provide a varied range of concepts which seems to introduce their community to the world. Editors have selected them in a disorganized way, and in the end, the capital and the country names are usually among the most popular concepts chosen by exporters. Languages like Hebrew and Arabic include priests and the sacred books like Koran and Mishneh Torah. Other languages like Catalan and Basque reflect their desire for a different political status, reflected in the names of the presidents and the pro-independence movements parties, among others.

In short, the exported concepts confirm that when editors are in a non-primary language they tend to choose those key and central concepts to their cultural identities. Perhaps with a fine-tuned recommender which provides less key concepts and eases the translation process, editors would be able to settle new cross-language participation as a fundamental part of their engagement in Wikipedia. I view this as the way more communities can benefit from content about other cultures and points of views they would hardly read otherwise.

Table 25. Top ten exported concepts according to the number of exporters who edited them.

Language	Top 10 Articles with Most Exporters
Arabic	Egypt, Libya, Muhammad, The Arabian Gulf, League of Arab States, Algeria, Bayda Libya, Morocco, Saudi Arabia, Arabic Language
Basque	Basque Language, Basque Country, Donostia, Basque Autonomous Community, Txillardegi, Euskaldun, Bilbao, Real Sociedad, Althetic Bilbao, Basque Wikipedia
Catalan	Catalonia, Barcelona, Catalan, Artur Mas i Gavarró, Carles Puigdemont i Casamajó, Girona, Generalitat de Catalunya, País Valencià, Catalan Independence, Lleida
English	United States, India, Ryan Higa, Australia, United Kingdom, Canada, Barack Obama, South Africa, 20th Century Fox, Michael Jackson
German	Germany, Berlin, Mecklenburg-Vorpommern, Frankfurt am Main, Düsseldorf, Hans-Dietrich Genscher, German language, Switzerland, Vienna
Hebrew	Israel, Tel Aviv Jaffa, Jerusalem, Mishneh Torah, Haifa, Haim Drukman, Hasamba, Holocaust, Acre, Hebrew University of Jerusalem
Hungarian	Hungary, Budapest, Hungarian language, Zichyújfalu, Esztergom, Pécs, Debrecen, Miskolc, Szeged, Kingdom of Hungary
Icelandic	Icelandic, Reykjavik, Icelandic, Akureyri, Ísafjarðardjúp, Alþingi, Olafur Ragnar Grimsson, Icelandic flag, Davíð Oddsson, Eyjafjallajökull
Italian	Italy, Sicily, Turin, Rome, Società Sportiva Lazio, Vatican City, Radio Studio 54 Network, Lingua Italiana, Trieste
Japanese	Japan, Tokyo, Japanese, Shinkansen, Kobe, Japanese Betula Pendola, Sapporo, Ujizane Imagawa, Toshiro Mifune, Osaka
Macedonian	Macedonia, Skopje, Macedonian Language, Kocani, Negotino, Macedonian People, Gotse Delchev, Kumanovo, Bitola, Macedonian Orthodox Church
Romanian	Romania, Timisoara, Cluj Napoca, Bucarest, Braşov, Moldavian Republic, Iași, Suceava, Județul Brașov, Traian Băsescu
Russian	Russia, Moscow, St. Petersburg, Russian language, Putin Vladimir Vladimirovich, Nizhny Novgorod, Russian empire, Kazan, Kazakhstan, Sebastopol
Spanish	Chile, Spain, Argentina, Mexico, Buenos Aires, Spanish Language, Madrid, Colombia, Costa Rica, Bogotá
Turkish	Turkey, Ankara, Mustafa Kemal Atatürk, Hayko, Istanbul, Galatasaray, Recep Tayyip Erdogan, Turkish, Besiktas, Ottoman Empire

Figure 46. Top 50 Catalan Wikipedia CIRA exported articles by number of exporters in non-primary languages (top) and by times created in non-primary language by exporters (bottom).

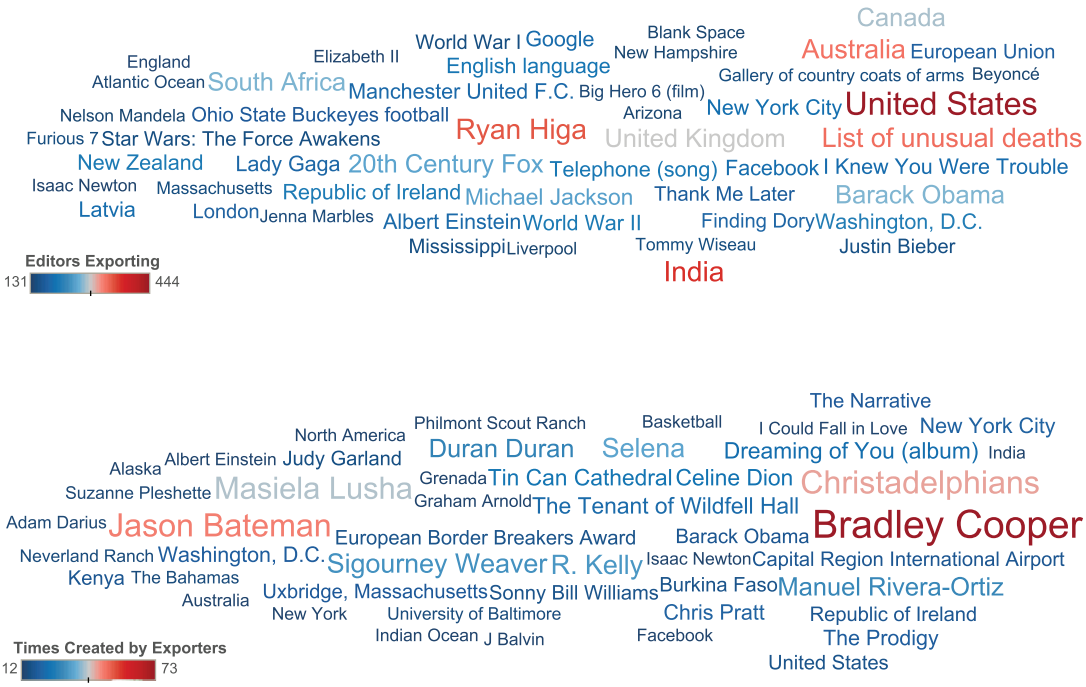


Figure 47. Top 50 English Wikipedia CIRA exported articles by number of exporters in non-primary languages (top) and by times created in non-primary language by exporters (bottom).

8.2.7 Summary of Results

The first part of this section showed that both the community core and anonymous editors are mostly responsible for creating content representing cultural identity **(RQ6)**. The contribution of anonymous editors has more weight proportionally in CIRA than in the rest of Wikipedia. These editors, whom I assume to be generally less involved in the community, and therefore more often motivated by the content itself rather than by the project, appear to be more influenced by their cultural context.

Regarding registered editors, increasing the participation level tends to increase the proportion of participation in cultural identity representations. After the first 100 edits, there is a strong diversity in editors' proportion of edits in CIRA but in general higher than for those who remain in the first edits. This could suggest that editors who exceed the first 100 edits are more influenced by cultural identity **(RQ7)**. By analysing higher participation levels, I found that the core group of the community is more homogenous - the highest level of participation - and big differences in the proportion of edits in CIRA disappear. This could suggest that editors may also edit CIRA if it is required by their functional role duties or other community dynamics.

In order to examine a scenario without these community-based effects, I explored the proportion of edits in CIRA during the first 7 days. Editors who eventually become administrators are significantly related to the degree of participation in CIRA in this initial period. This could indicate that editors who reach the core of the community have also developed a cultural identity, which ties them to their collective group beyond Wikipedia. Hence, in their first edits their content preferences are very identity-congruent choices.

The second part of this section analysed the activity of editors across languages, given that a considerable part of each Wikipedia community, and especially those most participative and the functional roles, contributes to multiple language editions. I examined how many of those also edit their language CIRA in their non-primary languages, and categorised them as exporters. Exporting CIRA is a much more specific activity, in which editors need to purposely change their usual language in order to engage in representing their cultural identities in the context of another community. Therefore, participation in CIRA in non-primary languages is not likely to be positively influenced by other factors such as the community dynamics but be driven by cultural identity and identity-based motivation.

In fact, almost the totality of the functional roles and a third of the multilingual editors contribute to export their primary language cultural identity representations into other language editions **(RQ8)**. This reaffirms the conclusion which states that the community core tends to be more engaged into editing content representing their cultural identities, and therefore, identity-based motivation influences them in a higher degree. In fact, exporters also showed a higher proportion of edits in CIRA in their primary language than primary multilingual editors and primary editors. Regarding exporters impact in their CIRA in non-primary languages, it is rather small: the percentage of edits they account for is around the 1% of the total edits.

I analysed the relationship between multilingual editors' participation in non-primary languages and their degree of participation in CIRA in non-primary languages. In the first 100 edits, exporters are a minority. After that, they spread over different degrees of participation. In general, achieving more edits in non-primary languages slightly affects the proportion of edits in CIRA in non-primary languages **(RQ9)**, as it remains stable or decreases a bit. The distribution presents a skewed shape instead of the Gaussian, which suggests that in non-primary languages there are not community effects which also encourage editors to edit in CIRA. Regular multilingual editors incorporate exporting as part of their content choices - even though they are in a scenario in which other editors do not share the same meanings. Consequently, when measuring for each editor the percentage of edits in CIRA in primary language and in non-primary languages, the first is usually bigger than the second.

Regarding the specific exported concepts, I found there exists a correlation between the number of editors in the article in the primary language and the number of exporters in the article in non-primary languages. This confirms the criterion established in section 7.4.6 to propose CIRA articles with a higher priority to be translated across languages editions. By examining the exported articles with a highest number of exporter editing them, it was possible to see that editors often edit the most central concepts in their cultural identities, such as the country names, capital, language name, and political figures in their past and current - in some particular cases, religious books or cultural mass celebrities also appeared **(RQ10)**. This points out that even in their non-primary languages, editors are moved by editing identity-congruent content, and that they prioritise those central concepts of their cultural identity.

"Every new beginning comes from some other beginning's end" Seneca

CONCLUSIONS, FUTURE RESEARCH AND DISSEMINATION

Chapter 9. Thesis Conclusions and Future Research

The goal of this thesis is to understand the influence of identity-based motivation in digital engagement. To this end, in Part 1 I initially defined and modelled the concept of digital engagement, explaining its aspects and its manifestations, such as participation. In Part 2, I employed the digital engagement model to understand Wikipedia Editor Engagement, by reviewing the current research studies on each of the aspects and manifestations of engagement. Finally, in Part 3, throughout an empirical analysis, I explored how identities become salient in the engagement with an object such as Wikipedia, and their influence on editor participation. For this purpose, I chose the community identity and the cultural identity, although other social identities could be cued, emerge and be relevant in Wikipedia, and likewise and more extensively, they could be relevant in other online platforms with a social component.

In this thesis, I used the identity-based motivation framework (IBM) (Oyserman, 2009; Oyserman & Destin, 2010) from the field of Social Psychology, in order to have a theoretical basis to explain the influence of specific identities on participation. The model mainly postulates that some identities are more relevant in any scenario, and therefore, people are motivated to act in identity-congruent ways. This means that interactions with a digital object can be analysed in terms of the possible identity-based meanings imbued in them. The IBM framework has been applied to the study of academic performance and engagement, consumption choices and health behaviours. Using it in the study of engagement with digital objects is a novel contribution to the field, and in particular the case studies from this research are contributions to the online communities' literature.

Once applied the model to Wikipedia, I believe it is possible to extract some general conclusions and lessons for the design of digital objects. According to IBM, any social identity can become relevant to foster participation, although wider identities can be more easily cued than narrow ones (e.g. being a fan of a football team is an identity easily more relevant in a context than being a professor). Most generally in digital objects, identity-congruent actions can either take place in a user self-representation dedicated area, or go unnoticed through the specific choices given by the object's main designed activities. For an identity to become salient, the object characteristics should be designed to provide available options which can be meaningful to editors. Even though many digital objects could provide contexts cuing identities, this is more the case of the rich variety of objects involving a social component.

In some objects, users develop an identity revolving around a purpose with a social scope (e.g. an online community) or even integrated in the narrative (e.g. a video game). Once users internalise these identities with their associated values and activities, future interactions may trigger new and more participative behaviours. Users tend to adopt a more complex view according to an in-object identity. Because when identities are cued in a scenario, they involve a procedural-readiness to make sense of the world according to their mind-set, and their associated actions may not necessarily serve a personal goal but an object-based goal (Oyserman, 2009). For instance, some augmented reality based mobile games require the players to internalise some goals and create a new in-game

identity, which sets them to reinterpret the environment accordingly. Then, in apparently off-game scenarios like a street walk, in the middle of the night, the new in-object identity could become relevant and motivate the users to play the game.

In some other objects, there may be no associated values or original purpose embedded, but still their wide scope or function may allow them to take actions clearly imbued by meaning. A good example of these are social network sites: they can be oriented to a specific area of life (e.g. LinkedIn with professional identities) or they can remain general (e.g. Twitter with messages about all sort of topics). The object provides the opportunity to choose between several actions, and those which are pragmatic options with identity-based meaning are more likely to be preferred over others. Social Networks allow each user to create specific groups based on social identities and display them as preferred. They even include some sophisticated algorithms to give precise recommendations and provide identity-congruent actions to perform - from meeting someone to sharing a picture.

In fact, enabling users to find a way to act in consistence with their identities may improve success of any specific digital object based on participation - whether there is a general purpose, a narrative or none of them. Sometimes the most salient identity in the context ends up being the in-game identity, which can only be created with materials from the object (e.g. a wizard identity in a fantasy game) and there is no place for external identities to relate to further actions. But generally, many digital objects can include some degrees of freedom in order to allow their users to explore their identity representations, either in a specific user dedicated space or through some actions.

I chose Wikipedia, where both scenarios occur: on the one side, there is a community identity, which is based on the project's goal and has very specific associated values, rules and activities, and on the other side, there are other social identities which are not encouraged to be represented in user pages, but nonetheless, they cannot be prevented from contributing to the topics they choose to. Empirical results suggest that identity-based motivation plays an important role in participative editors, both through the development of a community identity and in the particular articles imbued of cultural identity meanings they contribute to.

In the following sections of this chapter, I review the specific outcomes and conclusions of the three parts of the thesis. I propose reviewing them backwards: this will allow me to go from particular to general, from empirical results to the inherent problems of the concept. In Section 9.1, I review the main conclusions and findings from the case studies on community identity and cultural identity in Wikipedia, in Section 9.2 I summarise the main conclusions of the characteristics of Wikipedia Editor Engagement. In Section 9.3, I explain how I defined engagement in relation to different aspects of the user and the object. Finally, in Section 9.4 I present some of the limitations of this work, and possible future lines of research.

9.1 Identities in Wikipedia

In Part 3 of this thesis, in order to fulfil the **Thesis Objective 3**, I investigated the influence of identity-based motivation on Wikipedia editor engagement, by taking into account identities such as cultural identity and Wikipedia Community Identity. These two identities were especially suitable to Wikipedia, in accordance with its informational purpose, rules and collaborative management. The editor's community identity emerges in Wikipedia revolving around its purpose in a similar way that it does in other online communities (Danescu-Niculescu-Mizil & West, 2013; Ren et al., 2012). The editors' cultural identity becomes relevant in Wikipedia as it is an object divided by languages, where editors-speakers share a social space and aim at sharing information with readers'.

Community Identity Case Study. I selected the concept of community identity as a particular group identity created within the Wikipedia project. Editors develop it as they internalise the project's values and adopt a collective mind-set, which presupposes developing specific characteristics and participating in some activities. According to the identity-based motivation framework, the more an editor develops this identity, the more identity-congruent will be their future participation in the project will be. By mining and analysing 15 Wikipedia language editions, I could both provide an engagement characterisation of the Wikipedia communities, and generalise that when editors increase their participation, they also develop the community identity features.

Firstly, participative editors generally tend to develop a longer User Page in order to present themselves to the community. However, still a big part of the active and most active editors does not develop their User Page, which suggests that, in line with previous research (Bryant et al., 2005; Krupa et al., 2009), editors are mostly recognised by their participation level and type of contributions.

Secondly, participative editors often acquire a functional role in the community, although this is a product of a mutual decision between them and those already holding a functional role. There is not complete coincidence between the group with a high participation and the group with a functional role, as the first is larger. However, those who acquire a role vary in their behaviour patterns. For instance, depending on the type of their role, they show different editing session characteristics such as duration, number of contributions and time between sessions.

Thirdly, participative editors tend to become multilingual, at the same time they focus more and more their participation in their primary language. Multilingualism is an initiative aimed to contribute to Wikipedia as a global project, but it does not imply equally distributing participation among several language editions.

Fourthly, participative editors tend to dedicate a proportion of their participation to community oriented activities, such as uploading images, editing categories, among others. These are wikitasks which are out of the scope of personal preferences, and which imply a collective mind-set. Increasing the editor participation is also related to increasing the proportion of edits dedicated to communicating with other editors, although the increase is less evident than in community oriented activities. This is especially consistent

with a study on MovieLens online community (Ren et al., 2012), which revealed that the final engagement was more influenced by the attachment to the community values rather than by the attachment between users.

To my knowledge, this is the first study to broadly characterise engagement at the same time of providing an explanation linked to editors' community identity. Similar patterns were found in all the employed language editions, from different sizes to sociological contexts, confirming the wide-validity of these findings. Nonetheless, languages showed interesting differences in the distribution of functional roles in the community (some of them like the German Wikipedia relied on a wider group of Production Force, while the Japanese Wikipedia focused on Administrators). This role distribution could be analysed in terms of cultural values, as some cross-cultural study (Pfeil et al., 2006) inferred them in the interactions between editors in specific articles. Generally, functional roles are granted after few years of contributing to Wikipedia, and there is a relationship between the level of rights and the active years in the project.

In the analysed Wikipedia language editions, there is a visible decrease in user retention; along the years, lower proportions of new editors survive an initial period of six months. However, I believe that having identified the community identity features can be useful to design mechanisms to help new editors develop such features. In other words, the fact that participative editors develop the community identity features may suggest that in order to renew the core of the community, i.e. the most participative, it is necessary to help new editors to develop them (from self-presenting through a User Page to a functional role renewal plan). In this sense, often there are no 'best solutions' in design, since they all imply a trade-off between other aspects. Considering the decay in number of active editors it may be worth introducing changes. In Section 9.2, I give some recommendations in this direction.

Cultural Identity Case Study. I selected cultural identity as collective identity based on the shared cultural codes by people in a particular historical context (Hall, 1990). This type of identity is based on aspects such as language, traditions, history, and therefore, their meanings could be easily mapped to Wikipedia content. Therefore, any speaker of a language finds personally meaningful part of the language based cultural identities symbols and meanings and shares them, at least to some extent, with the rest of speakers. Therefore, building upon previous research which applied an identity-based motivation framework to environments such as school classrooms (Oyserman et al., 2003; Oyserman & Destin, 2010), I assumed that this same identity-based motivation could foster participation congruent with cultural identity. Hence, by studying the articles imbued with cultural identity meanings, their creation and the history of editors' edits, I explored the influence of this motivation type.

To summarise, this case study made two contributions:

1) *I explored the representation of cultural identities in Wikipedia articles*. I took into account 40 language editions to have sociocultural diversity and extend the validity of the results. I first mapped the meanings of the cultural identities to Wikipedia articles obtaining CIRA (Cultural Identity Related Articles) by use of semantic heuristics and computing techniques. While this content is spread in all sorts of topics (from history, to

science and sports), an analysis showed it is mainly unique and not shared across languages, creating a culture gap. The fact that there is a content imbalance in the Wikipedia's various language editions, imbalance mainly due to the editors' different cultural identities, implies that there is an opportunity to work on intercultural enrichment and make Wikipedia more multicultural.

In this sense, I found that CIRA articles about Geography and People tended to be more shared across languages. Language editions tended to share more articles from their CIRA with other languages depending on their size, along with their geographically and linguistic proximity. However, the culture gap is a problem in the measure that CIRA corresponds to an important part of the language gap (the overall number of non-shared articles), and their articles are more developed in their features (they are longer, have more images, etc.). In order to find which articles from CIRA should be recommended for translation, I pointed out a solution based on the number of editors. In order to explore the solution and encourage future research, the resulting datasets of Cultural Identity Related Articles for 40 language editions are made available. This aspect will be later developed in Section 10.3 and 10.4.

2) *I explored the influence of identity-based motivation and cultural identity by measuring the characteristics of CIRA and editors' interactions while building it*. The extent of CIRA in Wikipedia is far from negligible. I found that, on average, around one quarter of each language edition is made of content related to the corresponding cultural identities, and the proportion gets higher if we consider the attention in terms of edits. Later, I analysed how the different types of editors (in terms of functional role, and number of edits) contributed to the creation of CIRA.

Results showed that the editors who are more engaged with CIRA than with the rest of Wikipedia are anonymous editors. In fact, I assume them to be generally less involved in the community and possibly more attracted by content topics than by community values. Editors with higher levels of participation and administrators also tend to show a higher proportion of participation in CIRA. This holds also when only taking into account the first days after registering, which suggests that identity-based motivation may play an important role for engaging editors that will eventually get high levels of involvement in the project.

To complete the picture, I also detected multilingualism practices such as editors exporting the content representing their cultural identity to other Wikipedia language editions. The most exported articles were about specific political or geographical themes, which may involve very central and valuable meanings for a cultural identity. This practice emerged as a common activity especially in the core of the community, represented by the most participative editors. In fact, editors who export CIRA tend to have also a higher proportion of edits in CIRA in their primary language. These findings reveal that cultural identities can be relevant even while editors act in other language editions than their primary language edition. This suggests that with a proper channelling and tools, collaboration between different languages could be established in order to mutually enrich their content with articles related to their respective cultural identities and bridge the culture gap.

Taken all together, in the past Wikipedia literature, differences across language editions were referred as cultural contextualisation (Hecht, 2013). Some of its detected causing factors, such as geography, language and community, were analysed always independently from the editing process. This case study contributes to this research stream, and provides a more explicative stand. I believe that the theoretical insights supported by empirical evidences are helpful in understanding cultural contextualisation and embracing it as an opportunity for improving each Wikipedia language edition and cross-language collaborations, rather than seeing them as undesired contextual side-effects.

9.2 Wikipedia Editor Engagement

Numerous studies have been dedicated to Wikipedia editors' motivation, and agreed that some of the most important reasons are: the ideology of the project, acquiring writing skills and building a reputation in a community. However, up to now no study has been dedicated to identity-based motivation in Wikipedia. In Part 2 of this thesis, in order to fulfil the **Thesis Objective 2**, I reviewed the main studies on each of the aspects that influence Wikipedia editor engagement, according to the model of digital engagement devised in Part 1. Almost all aspects of this model have been studied in Wikipedia.

Wikipedia is an object in which the most important engagement manifestations are higher interaction (known as editor participation) and frequent return (known as editor retention). Nonetheless, these are mainly directed by the editors, since not many algorithms are employed to suggest new actions - like articles or editors to contact - the watchlist being the only tool that enables article following and alerts when changes take place. Hence, continuity in Wikipedia is mainly user-directed. When editors are presented with some new interactions triggered by Wikipedia, they are a consequence of other editors' messages and changes in the content. These interactions can have contradictory effects depending on the editor accumulated participation; for instance, several studies found that newcomers tend to be discouraged by deletions to the content they had just created, instead of seeing an opportunity to continue their activity (Halfaker, Kittur, & Riedl, 2011a).

In general terms, editors need to undertake several learnings to properly act in Wikipedia, ranging from the interface to the content and behaviour rules. Otherwise, they may break their fluent dialogue and disengage at any step. These learnings are related to different types of literacies, which received attention from several studies. Some of them highlight the number of guidelines, rules and policies (Butler et al., 2008), while others pay more attention to the usability issues (Cowan, 2011). While these literacies can be relevant during the interaction, they could also block the entrance to new editors. More precisely, some studies detected that there are potential editors who learn about the system before registering and enter the system with the first edit. Other studies suggested that the reduced number of women in Wikipedia may be due to women's lack of technical skills necessary to contribute (Hargittai & Shaw, 2015).

Since the very beginning of Wikipedia, most of the design changes (from interface to rules) have appeared due to the need to manage the production of content - including those involving the social aspects of the project (functional roles flags). Other steps involving communication or other social aspects (such as the implementation of user pages and talk spaces) have been gradually implemented after the launching. Instead, other possible social spaces such as a synchronous chat space for users have not yet been developed. This implies that editors need to access other external tools for the purpose. Nonetheless, it is good to bear in mind that Wikipedia succeeded while the parallel project Nupedia failed. One of the reasons for Wikipedia's success is because it offered a new interface 'wiki', which allowed visitors to become contributors with just a few clicks.

For this reason, the current decline in the number of active editors, along with the inability to retain new editors, could be seen as a consequence of the lack of changes in the software aimed at improving the design. As seen in Section 3.6 dedicated to the technological infrastructure governance, the Wikimedia Foundation prioritises the changes or tools that are most needed by the most active editors. Even though their teams are aware of the state of the community and the decline in number of editors, they also have other priorities, such as developing strategies to improve content readership, among other objectives. It is possible to say that to date, there has not been much pressure on implementing disruptive changes in favour of simplifying design and norm structure.

In fact, the decision-making process of accepting a new implementation is decided by consensus among the community - where the most participative editors tend to be more involved in these discussions. Therefore, in order to propose and accept changes which would favour newcomers' retention, there should be previously a wide awareness of the need for a bigger and more diversity community. Likewise, there should be awareness of the difficulties in the editing learning process, and such difficulties should be looked at as opportunities to re-design and simplify the process. Current editors have invested a lot of energy, and in fact, they have evolved along with the creation of the system. This is why they might be less aware of the need for change. Conversely, for the past few years, Wikimedia Foundation has been involved in several experiments on editor retention and is totally aware of the situation. Therefore, as a side-goal, Part 2 of the thesis lays the foundations for a discussion on how communities and Wikimedia Foundation should tackle Wikipedia editor engagement.

9.3 Digital Engagement

The empirical part of this thesis has been oriented towards measuring participation in Wikipedia (high interaction between user and object, and in multiple periods of time). Before the advent of Social Media and other massive online platforms, engagement was only referred to as participation in some fields of Social Sciences (e.g. civic engagement or brand engagement). In the most common use of the term, i.e. with technology in the field of Human-Computer Interaction, engagement refers to some qualities of the user experience such as attention and it is manifested through a sustained use of an object through time.

These multiple uses of the term have created a dislocation in research studies, and an agreement on which meaning should prevail was not yet reached. In Part 1, in order to fulfil **Thesis Objective 1**, I defined digital engagement and created a conceptual model to encompass participation, based on the current literature. I defined digital engagement as the quality which guarantees that the connection between a user and a digital object remains active. Therefore, for a connection between a user and an object to continue, there are always several causes. The challenge in studying engagement lies in prioritising the aspects which can best explain the engagement with a digital object, and which manifestations or facets can also be more explicative or what metrics are more valuable for understanding if it meets object design.

On the user side, I considered that motivation is the central aspect which explains why a user finds a reason to a continue a connection with an object. Attention instead could be explicative of how the user interacts with the object, the prior object used, or how the interaction would unfold in the future. The object's capacities to provide new interaction are not less important. Certain objects include strategies based on the content meaning, challenges or novelty which provide an appealing continuity to the user. Depending on how important the object's characteristics or the user's motivation are, the interaction may be object-directed or user-direct. However, if the object's design provides no affordance for the user to hint the next possible interaction, it means that the object has bad usability or, alternatively, the user does not have the literacy or the cognitive capacities to understand it. Either way, it is most likely that their fluent dialogue will end.

The use of digital objects has proliferated in all sorts of contexts. Therefore, the way we interact with them and our different states of attention may vary. In Section 2.3.5 I explained how different types of objects may be limited to very specific sorts of interaction, but still, they can all be engaging. Engaging is synonymous of alluring and desirable. This is perhaps what attracts researchers, designers and technology lovers in general, who all try to understand it in order to be ready to design the innovations and be an active part of the future.

9.4 Limitations and Future Lines of Research

a) Limitations

In this thesis, identity-based motivation was taken as the framework to help in understanding editors' actions in Wikipedia. In the Cultural Identity Case Study, I explored the influence of this motivation type by taking the articles imbued of cultural identities meanings that editors interacted with as a proxy. The advantages of using a quantitative approach are several, considering that in Wikipedia we computed the results for the entire population. The quantitative approach allows obtaining robust results in multiple sociocultural backgrounds, which is especially useful to obtain a wide validity of the conclusions.

However, as previously discussed, there are other factors influencing participation to this kind of content. Therefore, I must acknowledge that the methodology employed presents

some limitations when it comes to discovering the specific influence exerted by motivation. In fact, many researchers have studied Wikipedia editors' motivation using surveys or questionnaires (Xu & Li, 2015; X. Zhang & Zhu, 2006), instead of focusing on the behavioural outcome (i.e. the manifestations of engagement). These studies scrutinize how editors self-reflect on the reasons which motivate their actions (see Section 3.4.3).

Nonetheless, studies focusing on both motivation and participation are rare. One such study, for instance, is the study of how racial-ethnic identities influenced academic engagement in a high school class (Oyserman, 2009; Oyserman et al., 2003). In this study, conducted by the author of the identity-based motivation framework, both questionnaires and basic measurements were employed in a controlled environment, consisting in a task during a short period of time.

Using mixed methods in the present thesis would have been preferable, since such methods would have enriched the Cultural Identity Case Study under analysis. In particular, complementary self-reported methods would have been useful to assess the editors' degree of consciousness of their cultural identity-based choices in contributing to Wikipedia content. The languages analysed in the empirical research present important differences regarding some specific results, like the extent of content representing Cultural Identities - ranging from a quarter in average for all language editions to half of Wikipedia for English and Japanese language editions.

The results obtained for the English and Japanese Wikipedia could be further explained by applying complementary methods. The Japanese community, which presented totally different results than the rest, would deserve special attention. While in other languages, editors tended to gradually increase their proportion of participation into their cultural identities, in the Japanese language edition, editors tended to start strongly engaged with their cultural identities choices and then slowly decrease. This is why the case of Japanese would deserve further examination.

Nonetheless, it is good mentioning that self-reported methods such as questionnaires also present some flaws and disadvantages such as researcher bias and validity issues, or low response rates and lack of good sampling frames (Fricker & Schonlau, 2002). This last disadvantage has been reported as particularly sensitive in online community, considering that participation tends to be unequal (K. B. Wright, 2006). In Wikipedia, it would be difficult to access groups of editors such as anonymous editors, and nearly impossible to control the environment. In addition, I must argue that considering the large number of languages examined in the present study, it would have been difficult to obtain equivalent complementary results in the context of this thesis.

b) Future Lines of Research

The above-mentioned limitations pave the way to continue exploring the topic of identity-based motivation in digital engagement.

Since this work has been divided in different parts, each of them encourages new lines of research. Below I discuss five of these possible developments.

- **Cultural Identity Related Articles (CIRA) for all Wikipedias.**

The selection of articles related to cultural identities has been developed for 40 languages, but it could be expanded to the 293 Wikipedia language editions. The addition of Esperanto, besides other languages, may be of particular interest in order to study a cultural identity associated to a language, but no territories.

In fact, I plan to automatise this selection process so that the included articles, and their extent in Wikipedia is calculated and updated in a website that can be consulted by Wikipedia communities. In this future version of CIRA, I plan to use WikiData, a parallel database to Wikipedia language editions that allows categorising articles as entities. For instance, CIRA articles could include a new property named 'language-based article'.

These datasets are invaluable corpora that could be retrieved by a software which requires some contextualisation. Wikipedia is often used as a resource for other applications based on Natural Language Processing and having a more fine-grained selection of content nuclear to a culture can improve the performance of translators, searchers, among other tools.

- **Analysing editors' first interactions in CIRA.**

The analysis of the representation of cultural identities in Wikipedia content showed that some articles are more central than others, they receive more links coming from other languages, and are developed in greater detail (see Section 8.1.4) than others. Similarly, the CIRA articles which were exported by more editors tend to be key concepts to cultural identities (see Section 8.2.6). On the other hand, previous research (Hecht & Gergle, 2010a) showed that the articles concerning the territories where the language of the Wikipedia language edition is spoken tend to be edited by anonymous editors who are closer to those coordinates.

Therefore, I consider that either these more central CIRA articles or the segment of CIRA geolocated articles could attract different types of editors in different moments. I propose several experiments in this regard. First, it would be interesting to analyse these articles in relation to their editors and their position in the community in terms of edit count and functional role. Second, this could also serve to find out why bots have edited less CIRA articles. Third and finally, an analysis of the centrality of the articles that editors edit first during their initial period, would also be interesting. Right after having registered is a crucial moment for an editor, and possibly they need to edit prominent articles.

- **Studying the effects of presenting identity incongruent choices.**

Part of the empirical research of this thesis has shown that the articles whose meanings are in congruence with editors' identities, end up receiving a higher participation. While

this happens spontaneously, some techniques and algorithms used in searchers and social networks sites are aimed to foster user's comfort and drive continuity by presenting personalised information based on prior user searchers and choices.
In order to understand the importance of freedom in choosing topics in Wikipedia, a similar recommender system could present distant and perhaps uncomfortable meanings in order to see editors' reactions (for instance if they engage in and if they introduce new points of view). It would be interesting to measure both the effectiveness in attracting editors and the extent editors change this content and make it as congruent as possible. This algorithm could be tested to the purpose of completing points of view in articles.

- **Measuring engagement in other Wikimedia projects.**

No research has been dedicated to understand which kinds of editors develop other Wikimedia Foundation projects such as Wikibooks, Wikidictionary or Wikivoyages. These projects were created as a way of directing content that did not fit Wikipedia policies. It would be interesting to investigate whether there are editors engaged in the above-mentioned projects who are not involved in Wikipedia – since it would mean that they are relevant enough to draw newcomers in.

However, I conject that the same very engaged editors from each Wikipedia language edition community make incursions into these other projects. If so, what would be the consequences of their engagement with these other wiki-projects over their engagement with their primary Wikipedia language editions? I detected that the higher is the participation in a primary language, the more likely will an editor become multilingual (but from this it does not ensue a more equal participation in the multiple languages). It would be interesting to study if these patterns manifest themselves when editors divide their participation between Wikipedia and other Wikimedia projects.

- **Studying identity in education and games.**

The influence of identity on engagement has been assessed in the context of the Wikipedia editing community. However, other social objects such as educative massive online courses or massive multiplayer online games could be studied with a similar objective, namely understanding the intersection between the object purpose and the possibilities of representing the users' identities.

First, it would be necessary to single out which kind of identities can become salient in these social objects. For instance, in an educative platform, the alumni may provide personal information about their profession or career. Second, these identities should be mapped to the object's possible purpose and their activities. Extending the same example, certain professions would fit best certain courses (e.g. mathematics could be related to those who work as economists). Third, it would be possible to measure the engagement in the different activities in order to assess the influence of identities in their participation, as well as other aspects such as academic performance.

Chapter 10. Societal Impact and Dissemination

In this chapter, I first reflect on the relationship between research and society (10.1). I pay special attention to the ethical considerations derived from the research and Wikipedia. In this sense, I make several design recommendations to improve Wikipedia editor engagement (10.2), and to bridge the content culture gap in Wikipedia (10.3). Finally, I explain the dissemination of this research to the Wikipedia communities (10.4).

10.1 Ethical Considerations

The enormous amount of research studies appearing every year on Wikipedia shows that the interest for this object is far from decreasing. This is in part due to Wikipedia's transparency and content characteristics, but also to its popularity and the overall relationship with society. While I am writing these lines, Wikipedia is the fifth most visited website in the Internet[75], and the most accessed general reference work. Therefore, the problems affecting Wikipedia's Editor Engagement or the final Wikipedia content, from biases to incompleteness, have direct implications on people's education. Indeed, from my point of view, I believe that having investigated these issues and having assembled this study can be considered as my contribution to the project.

Therefore, this research cannot be isolated from society. In this sense, the European Commission has provided concrete normative orientations in the form of six policy keys that 'Responsible Research and Innovation'[76] should follow. They work as an umbrella of values, connecting practices and actors, and ultimately help as a reference point to discuss their implementation. They are: Ethics, Science Education, Governance, Public Engagement, Gender Equality and Open Access. I propose using them to briefly examine each of the ethical considerations and implications derived from both Wikipedia as a project, and this research work. Additionally, I propose several improvements to Wikipedia based on the results of this thesis and linked to these policies.

Ethics and Science Education. As already mentioned, Wikipedia has the objective of providing the 'sum of all human knowledge' for free, to everyone, in all those languages whose speakers want to work on it. In addition, it is the encyclopaedia that anyone can edit. Therefore, the project makes emphasis on an egalitarian view of education and decision-making. Ethics pervades all the aspects from the project: from its governance to the content management.
On the one hand, the content main policy, that of a Neutral Point of View, consists in representing fairly the different points of view in an article text. As a consequence, an engaged editor cannot freely impose her ideas, instead, he is encouraged to continue developing on the existing points of view. On the other hand, sources should be cited in every article and every information added has to be referenced. Wikipedia cannot be used as a final reference for any work, but a place where readers can find a good summary - a

[75] http://www.alexa.com/topsites
[76] http://www.rri-tools.eu/about-rri

starting point for further investigation. Because of this, Wikipedia can be considered an ally of science, since it provides accessible and selected knowledge from the vast amount of scientific publications.

Governance and Public Engagement. Wikipedia communities establish relationships with all sort of society actors in order to find contributors, and at the same time, to propagate its use. In addition, the blogs and social networks of the Wikimedia Foundation publish news regarding the new advances in software as well as research studies on Wikipedia. Wikimedia local organisations (also called chapters) allow their volunteers to work directly with the different societal actors, such as schools and museums from their nearby environment. There are outreaching programs aiming to expand the network of collaborators.

This flow of information among the wide variety of actors in the Wikimedia movement happens in total transparency and encourages each of them to get involved with their different functions. I believe transparency stimulates the different actors' engagement in the project. The empirical work from this thesis has focused on measuring the editors as the true core of the project. However, as seen in Part 2, engagement ultimately ends up depending on the technology governance and how the different actors involved can decide and implement good design changes. *In Section 10.2 I will provide three design recommendations in order to improve engagement.*

Gender Equality. Wikipedia's founder, Jimmy Wales, has stated in several occasions that a bigger and more diverse community is needed - and this necessity has also been pointed by researchers (Morgan et al., 2013). Ideally, there should not be a big difference between society and Wikipedia, and the communities should include all sort of sociological profiles. However, there are important gaps such as the lack of women in the editing community. Several studies (Lam, Uduwage, Dong, & Sen, 2011) reported the causes of such gap (lack of time, harassment, conflictive environment, and lack of technological skills, among others) and thanks to these studies there are some plans to raise awareness on the problem.

A direct consequence of the lack of diversity is the lack of topics that could interest the absent editors. More precisely, while content related to women is generally lacking, there are also content gaps concerning specific parts of the world where there are no (or very few) Wikipedia editors. In the work of this thesis I extracted the cultural identities from several language editions and found the extent of the culture gap between them. I adapted my research to propose a solution and established a criterion to prioritise those unique articles from each content representing cultural identity that should be translated first. *In Section 10.3, I stress the importance of bridging the culture gap, and come up with some suggestions on how to effectively do it.*

Open Access. Wikipedia's open and free culture makes available all the content and editors' interactions. However, not all research on Wikipedia has the opportunity of being published in such a license, as this depends on the conferences and journals' publishing houses. Furthermore, the datasets produced are often not shared, affecting replicability of research. For the purpose of this research, I uploaded the CIRA datasets, along with the published papers and conference presentations on this topic, in a website named

wikiidentities.org. *In Section 10.4, I develop a plan on how to disseminate this research through a website and in Wikipedia community events.*

10.2 Design Recommendations for Engagement

Both Part 2 (containing a literature review based on the digital engagement model) and Part 3 (containing the empirical research) confirmed the difficulties of retaining new editors. Some editors say that the current size of the communities is optimal, and justify such statement by the growth in the number of articles. However, the need for higher quality content and topic diversity, implies also a higher diversity in the composition of the community. In addition, the non-stop massive use of Wikipedia by readers suggests that there is a huge potential for turning a part of them into contributors. In the following four subsections, I provide several design recommendations based on this research in order to improve Wikipedia Editor Engagement.

10.2.1 New Community Identity features: Task Labelling, Editing Profiles and Recommendation System

The first design recommendation consists in building a system (a complex algorithm) **'Tasks-Profiles'** to identify wiki-work available in articles and in the history of editors' edits. This would allow, first of all, to discover new tasks and see the editors' profiles more easily, and secondly, to recommend both collaborations between editors with similar affinities and tasks, hence making editors more prone to collaborate and to accomplish the recommended tasks.

In fact, in the light of this thesis, to leverage community identity as motivation, it would be important to help new and all kind of editors to share more easily their latent skills and preferences regarding wiki-tasks. This is based on the idea that editors identify themselves mostly through the work they carry out (either topics or types of tasks) (Bryant et al., 2005). Likewise, quantitative research approaches have found that editors tend to fall in different profiles according to the types of work they mostly do (fact checker, copy editor, fact updater, vandal fighter, wiki gnome, etc.) (Welser et al., 2011; D. Yang et al., 2016). Having stated this, defining the types of wiki-tasks and topic interests would allow new editors to identify with and integrate faster in the community.

By making these tasks visible in the User Page, newcomers would be able to demonstrate their value more easily, by doing something more than merely checking the number of edits, which does not allow them to compete in recognition with more experienced editors. Also, presenting the work which still needs to be done in each article would possibly improve task discoverability. From previous research (Bryant et al., 2005), it is known that experienced editors are able to detect different sorts of tasks that newcomers cannot.

Tasks-Profiles would be two sides of the same coin; for instance, an article would require 'fact checking', and an editor who performs this operation repeatedly would be identified as 'fact checker'. By task, I mean those already detected by the current literature, but also specific topics (medicine, literature, etc.).

In fact, the benefits of a better definition and division of labour and of making such division visible are several. First, it would help new editors to present their role to the community (editors who tend to stay longer have a longer User Page). Second, it would stimulate core-periphery community dynamics based on new editors' potential value, and therefore it would set more precise expectations. And third, it would make 'advanced work' more accessible for the entire community. Perhaps this would raise privacy issues, but editors should also be able to administer what information from their profile is visible (if for instance they do not want to show the number of edits on specific topics).

The system would be based on identification and labelling processes. This could be run by automated bots, which would find tasks in articles and store meta-data on them, or, through user identification (e.g. marking that an article needs fact checking, to be longer, more pictures, more points of view, etc.). In addition, editors could also self-identify in a specific task they wish to be recognised by. After that, this information would be visible in articles and in user profiles. Editors would be able to access it initially, accessing the other editors' User Pages and the articles' meta-data.

Once a whole system of labelling would be set, more advanced discoverability techniques could be employed; for instance, a searcher could help editors find other editors specialized in a particular topic, task, or type of articles. Likewise, a personalised recommender system could propose new articles according to editors' preferred tasks, or possible collaborators according to their skills. This whole new way of labelling wiki-work would have an impact on all sort of communities, especially thinking about the topics where there is a limited number of experienced editors.

10.2.2 Community Self-Awareness: Redefining 'Wikipedian' and Community Engagement Monitoring

The second design recommendation includes three complementary strategies aimed at improving community self-awareness and empowering every editor to be concerned with the engagement of the whole community.

In the first place, if we want engagement to improve, I believe it needs to be looked at as an aspect concerning the present community. In this sense, I propose changing the definition of Wikipedian[77] from "volunteer who writes and edits Wikipedia's articles" into "member of the Wikipedia community, who writes and edits Wikipedia". While such change might seem insignificant, it actually adds a social dimension and recognises the value of belonging to the social group. Each editor would focus not only on producing content, but also on taking care of her peers and enjoying the learning process.

[77] https://en.wikipedia.org/wiki/Wikipedia:Wikipedians

To date, aspects regarding Editor Engagement have been mainly raised by the Wikimedia Foundation; currently there are some initiatives promoted by Wikimedia Foundation dedicated to fight harassment in Wikipedia, and individual proposals in order to improve the treatment newcomers receive. However, by properly defining the community with a dual focus (content and engagement), it may be possible to start many other initiatives, and ultimately, a positive cycle to make the community grow.

In the second place, another strategy to raise awareness on engagement is to simply make it visible. Editors usually take the number of articles of the Wikipedia language edition as the current position, in a sort of competition with other Wikipedia language editions. I propose that the state of each Wikipedia language edition community engagement is made available in real-time, so that it can be consulted by all the editors. It is true that there are some dashboards (VitalSigns) in some Wikimedia Tool websites, but they clearly lack the intentionality of being a common tool for all editors, from the newest to the oldest.

I believe that by providing a set of metrics, encompassing from general participation to new editor retention or survival, and a new definition including the dual-focus stated earlier, editors would be more aware that they have the capacity to improve the community engagement - and consequently, this would trigger behaviour changes. If the dashboard visualizes engagement in a usable way, I believe it can become a common place and bring positive effects. Seeing the participation of the community as a value in itself can raise awareness and help editors act in a more cohesive way.

In the third place, I propose the community makes one more step towards assuming a dual focus of action: content and engagement. At the very beginning of Wikipedia, some specific rights were created such as the Administrator flag (whose mission was to perform some protective actions and give stability to the project). Other functional roles were created with complementary purposes but always for content management purposes. The new step would be to create roles dedicated to community engagement, which would act as declared 'leaders' in this second scope (similarly to the way there already are content production leaders).

These editors would follow the engagement statistics at greater detail, and would be responsible for communicating the new implementations or software changes made available by the Wikimedia Foundation. By setting in each community some roles dedicated to focus on engagement, they would act as a counterpart to the hardcore producers, raise awareness on the current needs, and ease the implemention of changes aimed at improving editor retention. In fact, this new role of community engagement would be a way of bringing closer the current production-oriented community and the Wikimedia Foundation. Even without changing the governance culture based on consensus, it would still be possible to set new voices standing for a different interest than content production. Ultimately, it would be easier to address the long-time detected problems by research studies.

10.2.3 Design Continuity: Automated-Bots and Extensions for New Editor Assistance

The third design recommendation is the use of bots and extensions for editor assistance. To date, bots have been mainly used by experienced editors who wanted to leverage more work than they could manually edit. There are several types of bots, dedicated to introducing links between language editions, to update specific facts or send welcome messages to new editors. Besides the watchlist and a few tools, bots are among the few automatised algorithms within Wikipedia - and they are all designed for experienced editors.

I propose to start using bots and new extensions in order to assist the editing process. This means, in the first place, helping newcomers by showing them all the different sorts of rules which may apply in a situation or the most usual aspects editors need to learn. In the second place, it implies detecting and suggesting possible extra features for the articles (e.g. 'this article would require a picture). Artificial Intelligence may be far from creating content, nonetheless it can suggest possible improvements.

10.2.4 Identity-Congruent Campaigns: Attracting New Editors

The fourth and last design recommendation proposes communication campaigns to attract new editors and create new content based on specific results of the case study on the influence of cultural identity in Wikipedia editor engagement (Chapter 8).

In fact, one of the most interesting results from the editors' analyses revealed that acting in cultural identities related articles was fundamental to editors both in the core of the community and outside of it. Anonymous editors showed a higher proportion of participation in Cultural identity representations than to the rest of Wikipedia content, while administrators and very participative editors showed a higher proportion of participation in Cultural identity representations (even if we only take into account their very first days after registering in the site). Hence, based on this evidence, I propose campaigns to attract new editors using the cultural identity theme.

By creating campaigns dedicated to engage new editors and encourage them to grow cultural identity representations (e.g. including pictures in the article about their hometown), it may be possible to increase the content in Wikipedia and engage new editors – who are possible future participative members of the community. As Oyserman (2010, p. 1030) puts it: "identity-based motivation processes can be beneficial (goal supporting) or detrimental (goal undermining) depending on how identity is constructed in a specific context and on the behavioral and procedural options available in that context".

Campaigns and contests like Wiki Loves Monuments[78] and Wiki Loves Earth[79] are successful probably because they provide the opportunity to contribute with identity-congruent content. These contests only aim at growing content and are disseminated among members from the current community. Wiki Loves Monuments started in 2011 with a few European countries participating, while up to now it has involved more than 40 countries all over the world. Considering the imbalances between cultural content across languages, it is especially important to involve editors from the so-called Third World or Global South (see the white gaps in Figure 48). Nonetheless, I suggest that calls to new editors concerning identity-congruent actions (such as building the cultural heritage) may be a way to help them discover editing and make them grow Wikipedia content at the same time.

Figure 48. Wiki Loves Monuments 2016 is running in more than 40 countries.

10.3 Bridging the Wikipedia Content Culture Gap

An important contribution of this thesis is the analysis of Cultural Identity Related Articles in 40 Wikipedia language editions. To recapitulate, these articles occupy around a quarter of each language edition, and around 60% of them exist only in their language. This is what we call the culture gap, it is a gap in knowledge which could exist in other languages. It is due to the fact that some language editions lack the capacity of importing cultural content created by other language editions.

In this sense, this thesis proved that a large pool of editors partly under the influence of identity-based motivation had created a high number of articles related to their cultural identities. Therefore, the lack of editors in a language edition is the first cause for not being able to reproduce or import articles from another culture. As a consequence, this lack of editors can be due to a scale in demography, or to factors discouraging editing in the native language in Wikipedia. According to Van Dijk (2009) some of the causes of

[78] http://www.wikilovesmonuments.org
[79] http://wikilovesearth.org/

this phenomenon are the digital divide, the lack of popularity of the Wikipedia project or, more generally, the lack of literacy and skills to edit in Wikipedia.

In anthropology, the term of culture gap refers to a *misunderstanding of the other* or the difficulties in inter-cultural understanding that were common during the Colonial period in the XIX century. Today, the world is very connected through the use of Internet. Yet, it still remains very plural and diverse, and having access to information can be the key for living together in a context of tolerance and mutual understanding (as the UNESCO cultural diversity declaration explains[80]). Wikipedia can make available the concepts from each culture, and help understand their most important meanings, which are the base for intercultural communication (Lustig & Koester, 2010).
In fact, if the Wikipedia communities - and the Wikimedia movement in general - want to gather the sum of human knowledge for each language edition, they need to address the culture gap. Only in this way can they make each Wikipedia language edition much more multicultural than they are now. In order to do this, I propose employing the following strategies:

- The first strategy consists in the use of Wikiprojects and Challenges in order to coordinate efforts in the exchange of content between language editions. For instance, Wikipedians from Central and Eastern Europe organized an event where they edited content in their native language editions about their respective countries. More than 20 language editions were involved in the 2016 edition, and thousands of articles have been created since the first edition in 2012[81]. Similarly, the Catalan Community had created a Challenge, an online event which consisted in coordinating efforts between volunteers from the Catalan language community and from other language communities in order to create content from the Catalan culture in as many language editions as possible.

- The second strategy, already in use only in the Catalan Wikipedia, consists in creating a list of the 100 or 1000 fundamental articles from each Wikipedia language edition. This could be a successful strategy to export and import articles. This proposition is inspired by the list of 1000 articles that every Wikipedia should have, and would certainly simplify and set clear what each language edition considers important for their culture. While this is a finite list, in Section 8.1.6, I proposed a more automatized approach which relies on the use of certain article features (number of editors) in order to find articles that have higher priority for each current number of Interlanguage Links. This means that, for instance, one article could exist in 5 languages, but according to its relevance in the local Wikipedia, it should exist in at least 30 languages.

- The third strategy regards the use of recommending tools. The Wikimedia Foundation had developed a recommendation tool to help editors find articles in other languages that are unavailable in their language. I propose that the translator and this article recommendation tool[82] could include the CIRA or subparts of it

[80] http://unesdoc.unesco.org/images/0012/001246/124687e.pdf#page=67
[81] https://meta.wikimedia.org/wiki/Wikimedia_Central_and_Eastern_Europe
[82] http://recommend.wmflabs.org/

(e.g. articles including cultural identity related keywords in their title) as preferential content to translate and export across languages. It would retrieve the updated CIRA articles either from a separated database or from WikiData. In this sense, perhaps it would be interesting to create a property in WikiData database in order to label all the CIRA articles as language-based (e.g. 'Barcelona' article would have the property language-based associated to the Catalan Wikipedia). This way either editors and the recommending systems could retrieve articles from CIRA, the first to consult about the particular meanings related to a cultural identity and the second to recommend editors to import them to their language.

A further step would be to integrate these tools in the Wikipedia to make them more usable. Results from the analyses on editors' participation in cultural identity representations suggest that the core of the community edit in multiple language editions, and they even export the most important concepts in their cultural identities (see Section 8.2.6). Since editors are often encouraged by the marker of number of articles in their language edition and compare them to others, I believe that setting visible indicators of the current coverage of the culture gap for each language edition and individual users' efforts in this direction, and adjusting the mentioned recommendation and translation tools could help in stimulating these import-export tasks. This would ultimately improve the multicultural content coverage of the Wikipedia language editions.

10.4 Dissemination: Wikiidentities.org and Community Events

Finally, I describe the actions dedicated to the dissemination of this research and of the several working ideas presented in this chapter. One of the main assets of this research is its applicability and transformative nature. The actions I have undertaken consist in the creation of a website and in the dissemination of my research in Wikipedia community events and online spaces.

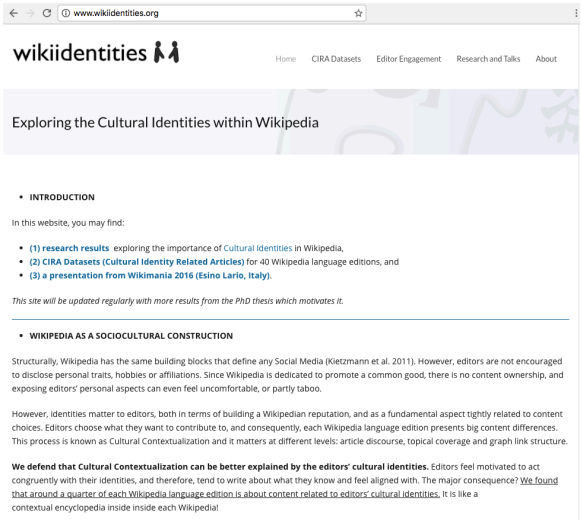


Figure 49. First version of the website wikiidentities.org released especially for the event Wikimania 2016.

Wikiidentities.org is a website I created to disseminate my research. In there, the reader may find the papers I have published, along with other parts of the thesis like interactive graphs. CIRA Datasets, code and other information are also provided[83] as a way of showing transparency in the research process, along with the idea of finding collaborations for future research.

I believe this can motivate and encourage new research on cultural identities, their structure, relationships and meanings. Furthermore, this website has the objective of raising awareness on the current situation of community engagement in the Wikimedia communities; some data visualizations of different language editions engagement not included in this thesis are provided.

In parallel to this, I had the chance to attend several events organized by the Wikipedia communities. On a local venue, in March 2016 I attended the fifteenth anniversary of the Catalan Wikipedia held in Barcelona, which coincided with the celebration of the 500,000 articles. There I had the chance to explain the state of the engagement, and several research studies on the factors causing the decrease of new editor retention, which surprised many of the present Wikipedians.

Several months later, on the 19th-20th of November 2016 I attended the Catalan Wikipedia meeting held in Valencia. I held a presentation entitled "Towards a User-Centered Wikipedia[84]" where I presented part of this thesis' results and discussed about the possible solutions to improve the Wikipedia Editor Engagement (solutions I mention in Section 10.2).

Figure 50. The author of this thesis discussing strategies to improve the Wikipedia Editor Engagement in Valencia.

[83] https://github.com/wikiidentities/cira_datasets_190715
[84] http://www.slideshare.net/MarcMiquel/usercentered-wikipedia-viquitrobada2016

On a more global venue, I held a presentation at the 2016 Wikimania conference in Esino Lario, Italy (June 21st-28th, 2016). I discussed the research results from CIRA and the strategies to bridge the culture gap[85].

Last but not least, I also used the online channels (Twitter, mailing lists, among others) from the Wikimedia Foundation to disseminate the different results of my thesis.

In a similar way to the future work proposed, I plan to undertake these actions to bring my individual contribution to Wikipedia. I truly believe the results and design recommendations can be the starting point for both improving Wikipedia editor engagement and bridging the culture gap.

[85]https://wikimania2016.wikimedia.org/wiki/Critical_issues_presentations/Identity-based_motivation_in_Wikipedia_as_a_key_to_collaboration_and_content_spreading_between_language_editions

Publications

Part of the research presented in this thesis has already been published in peer-reviewed articles (shown below), while other publications are under review.

- Miquel-Ribé, Marc, & Rodríguez Hontoria, Horacio (2011). Cultural configuration of Wikipedia: measuring autoreferentiality in different languages[86]. In *Proceedings of recent advances in natural language processing: Hissar, Bulgaria, 12-14 September 2011* (pp. 316-322). http://nlp.lsi.upc.edu/papers/ranlp11-mr.pdf

- Miquel-Ribé, Marc & Morera, David (2012). Extensive Survey to Readers and Writers of Catalan Wikipedia: Use, Promotion, Perception and Motivation[87]. Wikiacademy Berlin. 29 June 2012. https://ca.wikipedia.org/wiki/Fitxer:IV_sondeig_WP_ca_-_Resultats_preliminars.pdf

- Miquel-Ribé, Marc (2015). User Engagement on Wikipedia, A Review of Studies of Readers and Editors[88]. In *International AAAI Conference on Web and Social Media (ICWSM Workshop: Wikipedia, a Social Pedia: Research Challenges and Opportunities)*. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10645

- Miquel-Ribé, Marc & David Laniado (2016). Cultural Identities in Wikipedias[89]. In *Proceedings of the 7th 2016 International Conference on Social Media & Society* (SMSociety '16). ACM, New York, NY, USA, Article 24, 10 pages. DOI: http://dx.doi.org/10.1145/2930971.2930996

[86] This paper contains material or has been used to write Chapter 7.
[87] This survey has inspired the main ideas of this thesis.
[88] This paper contains material or has been used to write Chapter 3.
[89] This paper contains material or has been used to write Chapter 7.

Bibliography

Aaltonen, A., & Seiler, S. (2015). Cumulative Growth in User-Generated Content Production: Evidence from Wikipedia. *Management Science*, 2253–2217. http://doi.org/10.1287/mnsc.2015.2253

Ahmed, S. M., & Palermo, A.-G. S. (2010). Community engagement in research: frameworks for education and peer review. *American Journal of Public Health*, *100*(8), 1380–1387. http://doi.org/10.2105/AJPH.2009.178137

Ainley, M. (2006). Connecting with learning: Motivation, affect and cognition in interest processes. *Educational Psychology Review*, *18*(4), 391–405. http://doi.org/10.1007/s10648-006-9033-0

Altheide, D. L. (2000). Identity and the Definition of the Situation in a Mass-Mediated Context. *Symbolic Interaction*, *23*(1), 1–27. http://doi.org/10.1525/si.2000.23.1.1

Alvarez, M. S., Balaguer, I., Castillo, I., & Duda, J. L. (2009). Coach autonomy support and quality of sport engagement in young soccer players. *The Spanish Journal of Psychology*, *12*(1), 138–148.

Anderson, J. R. (2009). *Cognitive Psychology and its Implications*. New York, NY: Worth Publishers Ed.

Antin, J., & Cheshire, C. (2010). Readers are not free-riders: reading as a form of participation on wikipedia. (pp. 127–130). WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration.

Apic, G., Betts, M. J., & Russell, R. B. (2011). Content disputes in Wikipedia reflect geopolitical instability. *PloS One*, *6*(6). http://doi.org/10.1371/journal.pone.0020902.g001

Appleton, J. J., Christenson, S. L., Kim, D., & Reschly, A. L. (2006). Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument. *Journal of School Psychology*, *44*(5), 427–445. http://doi.org/10.1016/j.jsp.2006.04.002

Aragón, P., Laniado, D., Kaltenbrunner, A., & Volkovich, Y. (2012). Biographical social networks on Wikipedia: a cross-cultural study of links that made history. (p. 19). WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration.

Arapakis, I., Lalmas, M., Cambazoglu, B. B., Marcos, M.-C., & Jose, J. M. (2014). User engagement in online News: Under the scope of sentiment, interest, affect, and gaze. *Journal of the Association for Information Science and Technology*, *65*(10), 1988–2005. http://doi.org/10.1002/asi.23096

Arazy, O., Ortega, F., Nov, O., Yeo, L., & Balila, A. (2015). Functional roles and career paths in Wikipedia (pp. 1092–1105). CSCW '15: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing.

Arcas, L. B. (2014, March 28). *Los procesos de co-creación y el engagement del cliente: un análisis empírico en medios interactivos*. Doctoral dissertation. Universidad de Zaragoza. Spain.

Attfield, S., Kazai, G., & Lalmas, M. (2011). Towards a science of user engagement (position paper). In *WSDM Workshop on User Modelling for Web Applications*.

Ball, W. J. (2005). From Community Engagement to Political Engagement. *Political Science and Politics*, *38*(2), 287–291. http://www.jstor.org/stable/30044291

Banhawi, F., Ali, N. M., & Judi, H. M. (2012). User engagement attributes and levels in facebook. *Journal of Theoretical and Applied Information*, *41*(1), 11–19. http://www.jatit.org

Benbunan-Fich, R., Adler, R. F., & Mavlanova, T. (2011). Measuring multitasking behavior with activity-based metrics. *ACM Transactions on Computer-Human Interaction, 18(2), 1–22*. http://tochi.acm.org/

Biuk-Aghai, R. P., & Lei, K. H. (2010). Chatting in the Wiki: synchronous-asynchronous integration. WikiSym '10: Proceedings of the 6th International Symposium on Wikis and Open Collaboration. ACM.

Bouvier, P., Lavoue, E., & Sehaba, K. (2015). Defining Engagement and Characterizing Engaged-Behaviors in Digital Gaming. *Simulation & Gaming*, *45*(4-5), 491–507. http://doi.org/10.1177/1046878114553571

Boyle, E. A., Connolly, T. M., Hainey, T., & Boyle, J. M. (2012). Engagement in digital entertainment games: A systematic review. *Computers in Human Behavior*, *28*(3), 771–780. http://doi.org/10.1016/j.chb.2011.11.020

Brodie, R. J., Hollebeek, L. D., Juric, B., & Ilic, A. (2011). Customer Engagement: Conceptual Domain, Fundamental Propositions, and Implications for Research. *Journal of Service Research*, *14*(3), 252–271. http://doi.org/10.1177/1094670511411703

Brown, E., & Cairns, P. A. (2004). A grounded investigation of game immersion. (pp. 1297–1300). In *CHI'04 extended abstracts on Human Factors in Computing Systems* (pp. 1297-1300). ACM.

Bryant, S. L., Forte, A., & Bruckman, A. (2005). Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. (pp. 1–10). WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration, New York, New York, USA: ACM Press.

Burke, M., & Kraut, R. (2008). Mopping up: modeling wikipedia promotion decisions (pp. 27–36). CSCW '08: Proceedings of the 2008 ACM conference on Computer Supported Cooperative Work.

Butler, B., Joyce, E., & Pike, J. (2008). Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia (pp. 1101–1110). CHI '08: Proceedings of the SIGCHI conference on Human Factors in Computing Systems, ACM.

Callahan, E. S., & Herring, S. C. (2011). Cultural Bias in Wikipedia Content on Famous Persons. *Journal of the American Society for Information Science and Technology*, *62*(10), 1899–1915. http://doi.org/10.1002/asi.21577

Cassidy, D., Breakwell, N., & Bailey, J. (2014). Keeping them clicking: promoting student engagement in MOOC design. *The All Ireland Journal of Teaching and Learning in Higher Education*, *6*(2), 1-15.

Chapman, P., Selvarajah, S., & Webster, J. (1999, January). Engagement in multimedia training systems. In Systems Sciences, 1999. HICSS-32. Proceedings of the 32nd Annual Hawaii International Conference on (pp. 9-pp). IEEE.

Cheung, G., Zimmermann, T., & Nagappan, N. (2014). The first hour experience: how the initial play can engage (or lose) new players (pp. 57–66). CHI PLAY '14: Proceedings of the first ACM SIGCHI annual symposium on Computer-Human Interaction in play. ACM.

Chung, J., & Tan, F. B. (2004). Antecedents of perceived playfulness: an exploratory study on user acceptance of general information-searching websites. *Information &*

Management, *41*(7), 869–881. http://doi.org/10.1016/j.im.2003.08.016

Ciampaglia, G. L., & Taraborelli, D. (2015). MoodBar: Increasing New User Retention in Wikipedia through Lightweight Socialization (pp. 734–742). CSCW '15: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. ACM.

Ciampaglia, G. L., Flammini, A., & Menczer, F. (2015). The production of information in the attention economy. *Scientific Reports*, *5*, 9452–6. http://doi.org/10.1038/srep09452

Cohen, B. H., & Lea, R. B. (2004). *Essentials of Statistics for the Social and Behavioral Sciences (Essentials of Behavioral Science)*, 1–307. New York: NY. John Wiley and Sons.

Cowan, B. R. (2011). *Causal Effects of Wiki Site Design on Anxiety and Usability*.

Csikszentmihalyi, M. (1991). *Flow*. New York: Harper Collins.

Danescu-Niculescu-Mizil, C., & West, R. (2013). No country for old members: User lifecycle and linguistic change in online communities. www '13: Proceedings of the 22nd International conference on Wolrd Wide Web, ACM.

de Vreede, T., Nguyen, C., & de Vreede, G. J. (2013). A Theoretical Model of User Engagement in Crowdsourcing (Vol. 8224, pp. 94–109). WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration, Berlin, Heidelberg: Springer Berlin Heidelberg.

Deaton, A. (1997). *The Analysis of Household Surveys*. Washington, D.C.: The World Bank Publications.

Deci, E. L., & Ryan, R. M. (2012). *Overview of self-determination theory. In the Oxford handbook of human motivation* (pp. 85–107). New York: Oxford University Press.

Dobrian, F., Sekar, V., Awan, A., Stoica, I., Joseph, D. A., Ganjam, A., et al. (2011). Understanding the impact of video quality on user engagement. (Vol. 41, pp. 362–373). SIGCOMM: Computer Communication Review. ACM.

Dunn, O. J. (1964). Multiple Comparisons Using Rank Sums. *Technometrics*, *6*(3), 241. http://doi.org/10.2307/1266041

Dupret, G., & Lalmas, M. (2013). Absence time and user engagement: evaluating ranking functions. (pp. 173–182). WSDM '13: Proceedings of the sixth ACM international conference on Web search and data mining. ACM.

Ebner, M., Kickmeier-Rust, M., & Holzinger, A. (2008). Utilizing Wiki-Systems in higher education classes: a chance for universal access? *Universal Access in the Information Society*, *7*(4), 199–207. http://doi.org/10.1007/s10209-008-0115-2

Ensslin, A. (2011). "What an un-wiki way of doing things": Wikipedia's multilingual policy and metalinguistic practice. *Journal of Language and Politics*, *10*(4), 535–561. http://doi.org/10.1075/jlp.10.4.04ens

Eom, Y.-H., Aragón, P., Laniado, D., Kaltenbrunner, A., Vigna, S., & Shepelyansky, D. L. (2015). Interactions of cultures and top people of Wikipedia from ranking of 24 language editions. *PloS One*, *10*(3), e0114825. http://doi.org/10.1371/journal.pone.0114825

Farina, J., Tasso, R., & Laniado, D. (2011). Automatically assigning Wikipedia articles to macrocategories. WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration.

Fogg, B. J. (2002). *Persuasive technology: using computers to change what we think and do*. Massachussets: Morgan Kaufmann.

Forlizzi, J., & Battarbee, K. (2004). Understanding experience in interactive systems. (pp.

261–268). DIS '04: Proceedings of the 5th conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques. ACM.

Forte, A., & Bruckman, A. (2008). Why do people write for Wikipedia? Incentives to contribute to open–content publishing (pp. 1–11). HICSS '08: Proceedings of 41st Annual Hawaii International Conference on System Sciences.

Freyne, J., Jacovi, M., Guy, I., & Geyer, W. (2009). Increasing engagement through early recommender intervention (pp. 85–92). RecSys '09: Proceedings of the third ACM conference on Recommender systems. ACM.

Fricker, R. D., & Schonlau, M. (2002). Advantages and Disadvantages of Internet Research Surveys: Evidence from the Literature. *Field Methods*, *14*(4), 347–367. http://doi.org/10.1177/152582202237725

Gehring, V. V. (2004). *The Internet in Public Life*. Lanham, Maryland: Rowman & Littlefield.

Geiger, R. S., & Halfaker, A. (2013). Using edit sessions to measure participation in wikipedia. (pp. 861–870). WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration.

Giang, W. C. W., Hoekstra-Atwood, L., & Donmez, B. (2014). Driver Engagement in Notifications a Comparison of Visual-Manual Interaction between Smartwatches and Smartphones. In S. PUblications (Ed.), (Vol. 58, pp. 1–5). HFES '14: Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Gil, Q. (2016). Analyzing conflict and possible solutions around WMF software development (pp. 1–19). Proceedings of Wikimania '16.

Glott, R., Schmidt, P., & Ghosh, R. (2010). Wikipedia Survey – Overview of Results, 1–11.

Hahn, G. J., & Shapiro, S. S. (1994). *Statistical Models in Engineering*. New York, NY: Wiley-Interscience.

Hale, S. A. (2014). Multilinguals and Wikipedia editing (pp. 99–108). WS '14: Proceedings of the 2014 ACM conference on Web science.

Halfaker, A., Geiger, R. S., Morgan, J., & Riedl, J. (2013a). The Rise and Decline of an Open Collaboration Community: How Wikipedia's reaction to sudden popularity is causing its decline. *American Behavioral Scientist*, *57*(5), 664–688. http://doi.org/10.1177/0002764212469365.

Halfaker, A., Keyes, O., & Taraborelli, D. (2013b). Making peripheral participation legitimate: reader engagement experiments in Wikipedia (pp. 849–860). WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration.

Halfaker, A., Kittur, A., & Riedl, J. (2011a). Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work (pp. 163–172). WikiSym '11: Proceedings of the 7th international symposium on wikis and open collaboration. ACM.

Halfaker, A., Kittur, A., Kraut, R., & Riedl, J. (2009). A jury of your peers: quality, experience and ownership in Wikipedia. (p. 15). WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration. ACM.

Halfaker, A., Song, B., Stuart, D. A., Kittur, A., & Riedl, J. (2011b). NICE: social translucence through UI intervention. (pp. 101–104). WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration.

Hall, S. (1990). *Cultural identity and diaspora*. London: Editorial Jonathan Rutherford.

Hall, S. (1997). *Representation: Cultural representations and signifying practices (Vol. 2)*. New York, NY: SAGE.

Hargittai, E., & Shaw, A. (2015). Mind the Skills Gap: The Role of Internet Know-How and Gender in Contributions to Wikipedia. *Information, Communication & Society*, *18*(4), 424–442. http://doi.org/10.1080/1369118X.2014.957711

Hassenzahl, M., & Tractinsky, N. (2006). User experience - a research agenda. *Behaviour & Information Technology*, *25*(2), 91–97. http://doi.org/10.1080/01449290500330331

Heaberlin, B., & DeDeo, S. (2016). The Evolution of Wikipedia's Norm Network. *Future Internet*, *8*(2), 14–21. http://doi.org/10.3390/fi8020014

Hecht, B. J. (2013). *The Mining and Application of Diverse Cultural Perspectives in User-Generated Content*. Doctoral Dissertation. Northwestern University. United States.

Hecht, B. J., & Gergle, D. (2010a). On the localness of user-generated content (pp. 229–232). CSCW '10: Proceedings of the 2010 Conference on Computer Supported Cooperative Work. ACM.

Hecht, B., & Gergle, D. (2009). Measuring self-focus bias in community-maintained knowledge repositories (pp. 11–20). C&T '09: Proceedings of the Fourth International Conference on Communities and Technologies.

Hecht, B., & Gergle, D. (2010b). The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context (pp. 291–300). CHI '10: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM.

Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). *Cultures and Organizations: Software of the Mind, Third Edition*. New York, NY: McGraw Hill Professional.

Hwang, M. I., & Thorn, R. G. (1998). The effect of user engagement on system success: A meta-analytical integration of research findings. *Information & Management*, *35*(4), 229–236. http://doi.org/10.1016/S0378-7206(98)00092-5

Iosub, D., Laniado, D., Castillo, C., Fuster Morell, M., & Kaltenbrunner, A. (2014). Emotions under discussion: gender, status and communication in online collaboration. *PloS One*, *9*(8), e104880–. http://doi.org/10.1371/journal.pone.0104880

Jacques, R. (1995). Engagement as a Design Concept for Multimedia. *Canadian Journal of Educational Communication*, *24*(1), 49–59. https://www.learntechlib.org/?fuseaction=Reader.ViewIssues&source_code=ISSN-0710-4340

Kane, G. C. (2009). It's a Network, Not an Encyclopedia: A Social Network Perspective on Wikipedia Collaboration. (Vol. 2009, pp. 1–6). Academy of Management Proceedings, Academy of Management.

Kappelman, L. A., & McLean, E. R. (1992). Promoting information system success: the respective roles of user participation and user involvement. *Journal of Information Technology Management*, *3*(1), 2–12. https://jitm.ubalt.edu/

Karat, C. M., Karat, J., Vergo, J., & Pinhanez, C. (2002). That's entertainment! Designing streaming, multimedia web experiences. *International Journal of Human-Computer Interaction*, *14*(3-4), 369–384. http://doi.org/10.1080/10447318.2002.9669125

Karimi, F., Bohlin, L., Samoilenko, A., Rosvall, M., & Lancichinetti, A. (2015). *Quantifying national information interests using the activity of Wikipedia editors*. *CoRR Abs/1312.0976*, *1503*, 5522.

Keegan, B., Gergle, D., & Contractor, N. (2013). Hot Off the Wiki: Structures and Dynamics of Wikipedia's Coverage of Breaking News Events. *American Behavioral*

Scientist, *57*(5), 595–622. http://doi.org/10.1177/0002764212469367

Keegan, B., Gergle, D., & Contractor, N. S. (2012). Do editors or articles drive collaboration?: multilevel statistical network analysis of wikipedia coauthorship. (pp. 427–436). WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration.

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, *54*(3), 241–251. http://doi.org/10.1016/j.bushor.2011.01.005

Kim, S., Park, S., Hale, S. A., Kim, S., Byun, J., & Oh, A. H. (2016). Understanding Editing Behaviors in Multilingual Wikipedia. *PloS One*, *11*(5), e0155305. http://doi.org/10.1371/journal.pone.0155305

Kittur, A., Chi, E. H., & Suh, B. (2009a). What's in Wikipedia?: mapping topics and conflict using socially annotated category structure (pp. 1509–1512). CHI '08: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Kittur, A., Chi, E., Pendleton, B. A., & Suh, B. (2007). Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie (Vol. 1, p. 19). WWW '07: Proceedings of the Sixteenth International Conference on the World Wide Web.

Kittur, A., Pendleton, B. A., & Kraut, R. E. (2009b). Herding the cats - the influence of groups in coordinating peer production. (p. 7). WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration. ACM.

Klein, L. (2015). *Design for Voice Interfaces*. (pp. 1–36) Sebastopol, California: O'Reilly Media.

Krupa, Y., Vercouter, L., Hübner, J. F., & Herzig, A. (2009). Trust Based Evaluation of Wikipedia's Contributors. *Engineering Societies in the Agents World X*, *5881* (13), 148–161. http://doi.org/10.1007/978-3-642-10203-5_13

Ksiazek, T. B., Peer, L., & Lessard, K. (2014). User engagement with online news: Conceptualizing interactivity and exploring the relationship between online news videos and user comments. *New Media & Society*, 1–19. http://doi.org/10.1177/1461444814545073

Kuznetsov, S. (2006). Motivations of contributors to Wikipedia. *ACM SIGCAS Computers and Society*, *36*(2), 1. http://dx.doi.org/10.1145/1215942.1215943

Lalmas, M., O'Brien, H., & Yom-Tov, E. (2014). *Measuring User Engagement*. (G. Marchionini, Ed.) (4 ed.). San Rafael, California: Morgan & Claypool Publishers.

Lam, S., Uduwage, A., Dong, Z., & Sen, S. (2011). WP: Clubhouse?: an exploration of Wikipedia's gender imbalance. Wikisym '11: Proceedings of the 7th international symposium on Wikis and open collaboration.

Laniado, D., & Tasso, R. (2011, June). Co-authorship 2.0: Patterns of collaboration in Wikipedia. In Proceedings of the 22nd ACM conference on Hypertext and hypermedia (pp. 201-210). ACM.

Laniado, D., Kaltenbrunner, A., Castillo, C., & Morell, M. F. (2012, August). Emotions and dialogue in a peer-production community: the case of Wikipedia. Wikisym '12: Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration (p. 9). ACM.

Laniado, D., Tasso, R., Volkovich, Y., & Kaltenbrunner, A. (2011). When the wikipedians talk: Network and Tree Structure of Wikipedia discussion pages. ICWSM 2011: In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.

Laurel, B. (1991). *Computers as theatre*. New York, NY: A.-W. P. Company, Ed.

Lavie, T., & Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. *Journal of Human Computer Studies*, *60*(3), 269–298. http://doi.org/10.1016/j.ijhcs.2003.09.002

Lehmann, J., Lalmas, M., Dupret, G., & Baeza-Yates, R. (2013). Online multitasking and user engagement (pp. 519–528). I&KM '13: Proceedings of the 22nd International Conference on Information & Knowledge Management. ACM.

Lehmann, J., Lalmas, M., Yom-Tov, E., & Dupret, G. (2012). Models of user engagement (pp. 164–175). UMAP'12: Proceedings of the 20th international conference on User Modeling, Adaptation, and Personalization. Springer-Verlag.

Lehmann, J., Müller-Birn, C., Laniado, D., Lalmas, M., & Kaltenbrunner, A. (2014). Reader preferences and behavior on Wikipedia (pp. 88–97). HT '14: Proceedings of the 25th ACM conference on Hypertext and Social Media. ACM.

Levy, S. (1984). *Hackers: Heroes of the Computer Revolution*. New York, NY: Bantam Doubleday Dell Publishing Group.

Lieberman, M. D., & Lin, J. (2009). You Are Where You Edit: Locating Wikipedia Contributors Through Edit Histories, 1–8. ICWSM '09: Proceedings of the Third International Conference on Weblogs and Social Media.

Lih, A. (2009). *The Wikipedia Revolution*. London: Aurum Press.

Liikkanen, L. A., & Salovaara, A. (2015). Music on YouTube: User engagement with traditional, user-appropriated and derivative videos. *Computers in Human Behavior*, *50*(C), 1–17. http://doi.org/10.1016/j.chb.2015.01.067

Lovink, G., Tkacz, N., Reagle, J. M., O'Sullivan, D., & Liang, L. (2012). *Critical point of view: a Wikipedia reader*. Amsterdam: Institute of Network Cultures.

Lustig, M. W., & Koester, J. (2010). *Intercultural Competence: Interpersonal Communication Across Cultures* (6 ed.). Upper Saddle River, New Jersey: Pearson.

Mark, G., Iqbal, S., Czerwinski, M., & Johns, P. (2015). Focused, Aroused, but so Distractible: Temporal Perspectives on Multitasking and Communications (pp. 903–916). CSCW '15: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. ACM.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*(2), 224–253. http://doi.org/10.1037/0033-295X.98.2.224

Marsh, T., & Nardi, B. A. (2014). Spheres and Lenses: Activity-Based Scenario / Narrative Approach for Design and Evaluation of Entertainment through Engagement. *Icec*, *8770*(6), 42–51. http://doi.org/10.1007/978-3-662-45212-7_6

Massa, P., & Scrinzi, F. (2011). Exploring linguistic points of view of Wikipedia. (pp. 213–214). WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration. ACM.

Mauri, M., Cipresso, P., Balgera, A., Villamira, M., & Riva, G. (2011). Why Is Facebook So Successful? Psychophysiological Measures Describe a Core Flow State While Using Facebook. *Cyberpsychology, Behavior, and Social Networking*, *14*(12), 723–731. http://doi.org/10.1089/cyber.2010.0377

McCarthy, J., & Wright, P. (2004). *Technology As Experience*, 1–225. Cambridge, Massachussetts: The MIT Press.

McCay-Peet, L., Lalmas, M., & Navalpakkam, V. (2012). On saliency, affect and focused attention. (pp. 541–550). CHI '13: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM.

McWilliams, W. (2013). *Brand Engagemen*t. Mulgrave: WMC Public Relations Pty,

Limited.
Menking, A., & Erickson, I. (2015). The Heart Work of Wikipedia (pp. 207–210). CHI '15: the 33rd Annual ACM Conference. ACM.
Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F. Å., & Lanamäki, A. (2015). "The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia. *Journal of the Association for Information Science and Technology*, *66*(2), 219–245. http://doi.org/10.1002/asi.23172
Miljesic, L., & Ricchiuti, F. (2016). Wikipedia Clicks: Exploring Trends, 1–7.
Morell, M. F. (2010). *Governance of Online Creation Communities: Provision of infrastructure for the building of digital commons*. Doctoral Dissertation. European University Institute.
Morgan, J. T., Bouterse, S., Walls, H., & Stierch, S. (2013). Tea and sympathy: crafting positive new user experiences on wikipedia. (pp. 839–848). WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration.
Nakamura, J., & Csikszentmihalyi, M. (2009). *Flow Theory and Research* (pp. 195–206). Oxford University Press. http://doi.org/10.1093/oxfordhb/9780195187243.013.0018
Neff, J. J., Laniado, D., Kappler, K. E., Volkovich, Y., Aragón, P., & Kaltenbrunner, A. (2013). Jointly They Edit: Examining the Impact of Community Identification on Political Interaction in Wikipedia. *PloS One*, *8*(4), 60584. http://doi.org/10.1371/journal.pone.0060584
Nielsen, J. (1999). *Designing Web Usability*. Indianapolis, IN: New Riders.
Norris, S. (2004). *Analyzing Multimodal Interaction*. Routledge.
Nov, O. (2007). What motivates Wikipedians? *Communications of the ACM*, *50*(11), 60–64. http://doi.org/10.1145/1297797.1297798
O'Brien, H. L. (2008). *Defining and Measuring Engagement in User Experiences with Technology*. Doctoral Dissertation. University of British Columbia. ProQuest.
O'Brien, H. L. (2011). Exploring user engagement in online news interactions (Vol. 48, pp. 1–10). *Journal of the American Society for Information Science and Technology*, *59*(6), 938–955. http://doi.org/10.1002/asi.20801
O'Brien, H. L., & Toms, E. G. (2008). What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, *59*(6), 938–955. http://doi.org/10.1002/asi.20801
Okoli, C. (2009). A Brief Review of Studies of Wikipedia in Peer-Reviewed Journals. *Audio, Transactions of the IRE Professional Group on*, 155–160. http://doi.org/10.1109/ICDS.2009.28
Okoli, C., Mehdi, M., Mesgari, M., Nielsen, F. Å., & Lanamäki, A. (2014). Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership. *Journal of the Association for Information Science and Technology*, *65*(12), 2381-2403.
Okoli, C., Mehdi, M., Mesgari, M., Nielsen, F. Å., & Lanamäki, A. (2012). The People's Encyclopedia Under the Gaze of the Sages: A Systematic Review of Scholarly Research on Wikipedia. *SSRN Electronic Journal*. http://doi.org/10.2139/ssrn.2021326
Ortega, F., & Barahona, J. G. (2007). Quantitative analysis of the Wikipedia community of users (pp. 75–86). Wikisym 07': Proceedings of the 2007 international symposium on Wikis.
Ortega, F., González-Barahona, J. M., & Robles, G. (2008). On the Inequality of

Contributions to Wikipedia. (pp. 304–304). WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration, IEEE.

Oyserman, D. (2009). Identity-based motivation and consumer behavior. *Journal of Consumer Psychology*, *19*(3), 250–260. http://doi.org/10.1016/j.jcps.2009.05.008

Oyserman, D., & Destin, M. (2010). Identity-based motivation: Implications for intervention. *The Counseling Psychologist*, *38*(7), 1001–1043. http://doi.org/10.1177/0011000010374775

Oyserman, D., Kemmelmeier, M., Fryberg, S., Brosh, H., & Hart-Johnson, T. (2003). Racial-Ethnic Self-Schemas. *Social Psychology Quarterly*, *66*(4), 333–347. http://doi.org/10.2307/1519833

Panciera, K., Halfaker, A., & Terveen, L. (2009). Wikipedians are born, not made: a study of power editors on Wikipedia. Presented at the GROUP '09: Proceedings of the ACM 2009 international conference on Supporting group work. http://doi.org/10.1145/1531674.1531682

Pariser, E. (2011). *The Filter Bubble*. New York, NY: Penguin Press.

Peterson, E. T., & Carrabis, J. (2008). Measuring the immeasurable: Visitor engagement. *Web Analytics Demystified*, *14*, 16.

Pfeil, U., Zaphiris, P., & Ang, C. S. (2006). Cultural Differences in Collaborative Authoring of Wikipedia. *Journal Computer-Mediated Communication*, *12*(1), 88–113. http://doi.org/10.1111/j.1083-6101.2006.00316.x

Porter, C. E. (2006). A Typology of Virtual Communities: A Multi-Disciplinary Foundation for Future Research. *Journal of Computer-Mediated Communication*, *10*(1), 00–00. http://doi.org/10.1111/j.1083-6101.2004.tb00228.x

Preece, J. (2000). *Online communities*. New York, NY: John Wiley Sons.

Preece, J., & Shneiderman, B. (2009). The reader-to-leader framework: Motivating technology-mediated social participation. *AIS Transactions on Human-Computer Interaction*, *1*(1), 13–32. http://aisel.aisnet.org/thci/

Przybylski, A. K., Rigby, C. S., & Ryan, R. M. (2010). A motivational model of video game engagement. *Review of General Psychology*, *14*(2), 154–166. http://doi.org/10.1037/a0019440

Raitman, R. S., Augar, N., & Zhou, W. (2005). Employing Wikis for Online Collaboration in the E-Learning Environment: Case Study. (Vol. 2, pp. 142–146). ICITA 05': In Proceedings of the Third International Conference on Information Technology and Applications. IEEE.

Reeves, B., & Read, J. L. (2009). *Total Engagement*. Brighton: Harvard Business Press.

Reinoso, A. J., & Ortega, F. (2009). A quantitative approach to the use of the Wikipedia (pp. 56–61). WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration.

Ren, Y., Harper, F. M., Drenner, S., Terveen, L. G., Kiesler, S. B., Riedl, J., & Kraut, R. E. (2012). Building Member Attachment in Online Communities - Applying Theories of Group Identity and Interpersonal Bonds. *MIS Quarterly*. http://misq.org

Ristau, K. (2011). Folklore and the Internet: Vernacular Expression in a Digital World. *Western Folklore*, *70*(3/4), 376. http://www.westernfolklore.org

Rizoiu, M.-A., Xie, L., Caetano, T., & Cebrian, M. (2016). Evolution of Privacy Loss in Wikipedia (pp. 215–224). Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. ACM.

Rogers, R., & Sendijarevic, E. (2012). Neutral or National Point of View? A Comparison of Srebrenica articles across Wikipedia's language versions. *Wikipedia Academy*.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68. http://dx.doi.org/10.1037/0022-3514.34.

Schroeder, R., & Taylor, L. (2015). Big data and Wikipedia research: social science knowledge across disciplinary divides. *Information, Communication & Society*, *18*(9), 1039–1056. http://doi.org/10.1080/1369118X.2015.1008538

Schüll, N. D. (2012). *Addiction by Design*. Princeton: Princeton University Press.

Shih, M., Pittinsky, T. L., & Ambady, N. (1999). Stereotype Susceptibility: Identity Salience and Shifts in Quantitative Performance. *Psychological Science*, *10*(1), 80–83. http://doi.org/10.1111/1467-9280.00111

Slattery, S. (2009). "edit this page": the socio-technological infrastructure of a wikipedia article. (pp. 289–296). SIGDOC 09': Proceedings of the 27th ACM international conference on Design of communication. ACm.

Smith, B. G., & Gallicano, T. D. (2015). Terms of engagement: Analyzing public engagement with organizations through social media. *Computers in Human Behavior*, *53*, 82–90. http://doi.org/10.1016/j.chb.2015.05.060

Steinkuehler, C. A. (2006). Massively multiplayer online video gaming as participation in a discourse. *Mind, culture, and activity*, *13*(1), 38-52. http://doi.org**:**10.2304/elea.2007.4.3.297

Stone, G. P., Roach, J. R., & Eicher, M. E. (1993). *Appearance and the self, dress and the social order*. New York, NY: John Wiley & Sons.

Suchman, L. (2007). Human-Machine Reconfigurations. Cambridge University Press.

Suh, B., Convertino, G., Chi, E. H., & Pirolli, P. (2009). The singularity is not near: slowing growth of Wikipedia. (p. 1). WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration. ACM.

Sutcliffe, A. (2010). Designing for User Engagement. *Synthesis Lectures on Human-Centered Informatics* (Vol. 2, pp. 47–55). Morgan & Claypool Publishers. http://doi.org/10.2200/S00210ED1V01Y200910HCI005

Tsikerdekis, M. (2015). Personal communication networks and their positive effects on online collaboration and outcome quality on Wikipedia. *Journal of the Association for Information Science and Technology*. *(67), 812–823*. http://doi.org/10.1002/asi.23429

Van Dijk, Z. (2009). Wikipedia and lesser-resourced languages. *Language Problems & Language Planning*, (3), 33. http://doi.org/10.1075/lplp.33.3.03van

Van Vugt, H. C., Konijn, E. A., Hoorn, J. F., Keur, I., & Eliëns, A. (2007). Realism is not all! User engagement with task-related interface characters. *Interacting with Computers*, *19*(2), 267–280. http://doi.org/10.1016/j.intcom.2006.08.005

Voss, J. Measuring Wikipedia. (2005). Proceedings of the International Conference of the International Society for Scientometrics and Informetrics.

Warncke-Wang, M., Ranjan, V., Terveen, L. G., & Hecht, B. (2015). Misalignment Between Supply and Demand of Quality Content in Peer Production Communities. (pp. 493–502). ICWSM '11: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.

Warncke-Wang, M., Uduwage, A., Dong, Z., & Riedl, J. (2012). In search of the ur-Wikipedia: universality, similarity, and translation in the Wikipedia inter-language link network. OpenSym '12: Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration, 20.

Webster, J., & Ahuja, J. S. (2006). Enhancing the Design of Web Navigation Systems:

The Influence of User Disorientation on Engagement and Performance. *MIS Quarterly*, *30*(3), 661–678. http://misq.org/

Webster, J., & Ho, H. (1997). Audience Engagement in Multimedia Presentations. *ACM SIGMIS Database 28.2*, *28*(2), 63–77. http://doi.org/10.1145/264701.264706

Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., & Smith, M. A. (2011). Finding social roles in Wikipedia. (pp. 122–129). WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration. ACM.

Wiebe, E. N., Lamb, A., Hardy, M., & Sharek, D. (2014). Measuring engagement in video game-based environments: Investigation of the User Engagement Scale. *Computers in Human Behavior*, *32*(C), 123–132. http://doi.org/10.1016/j.chb.2013.12.001

Wright, K. B. (2006). Researching Internet-Based Populations: Advantages and Disadvantages of Online Survey Research, Online Questionnaire Authoring Software Packages, and Web Survey Services. *Journal of Computer-Mediated Communication*, *10*(3), 00–00. http://doi.org/10.1111/j.1083-6101.2005.tb00259.x

Wulczyn, E., West, R., Zia, L., & Leskovec, J. (2016). Growing Wikipedia Across Languages via Recommendation (pp. 975–985). WWW '16: Proceedings of the 25th International Conference on World Wide Web. ACM.

Xu, B., & Li, D. (2015). An empirical study of the motivations for content contribution and community participation in Wikipedia. *Information & Management*, *52*(3), 275–286. http://doi.org/10.1016/j.im.2014.12.003

Yang, D., Halfaker, A., Kraut, R. E., & Hovy, E. H. (2016). Who Did What - Editor Role Identification in Wikipedia. ICWSM 2011: Fifth International AAAI Conference on Weblogs and Social Media.

Yang, H.-L., & Lai, C.-Y. (2010). Computers in Human Behavior. *Computers in Human Behavior*, *26*(6), 1377–1383. http://doi.org/10.1016/j.chb.2010.04.011

Yuqing Ren, Kraut, R., & Kiesler, S. (2007). Applying Common Identity and Bond Theory to Design of Online Communities. *Organization Studies*, *28*(3), 377–408. http://doi.org/10.1177/0170840607076007

Zhang, X. M., & Zhu, F. (2011). Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia. *American Economic Review*, *101*(4), 1601–1615. http://doi.org/10.1257/aer.101.4.1601

Zhang, X., & Zhu, F. (2006). Intrinsic motivation of open content contributors: The case of Wikipedia. Workshop on Information Systems and Economics, 1601–1615.

Zhu, H., Zhang, A., He, J., Kraut, R. E., & Kittur, A. (2013). Effects of peer feedback on contribution: a field experiment in Wikipedia. (pp. 2253–2262). CHI '13: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM.

All the websites have been last consulted in October 2016.

Appendix 1. Survey of Catalan Wikipedia Editors

1.1 Infographics

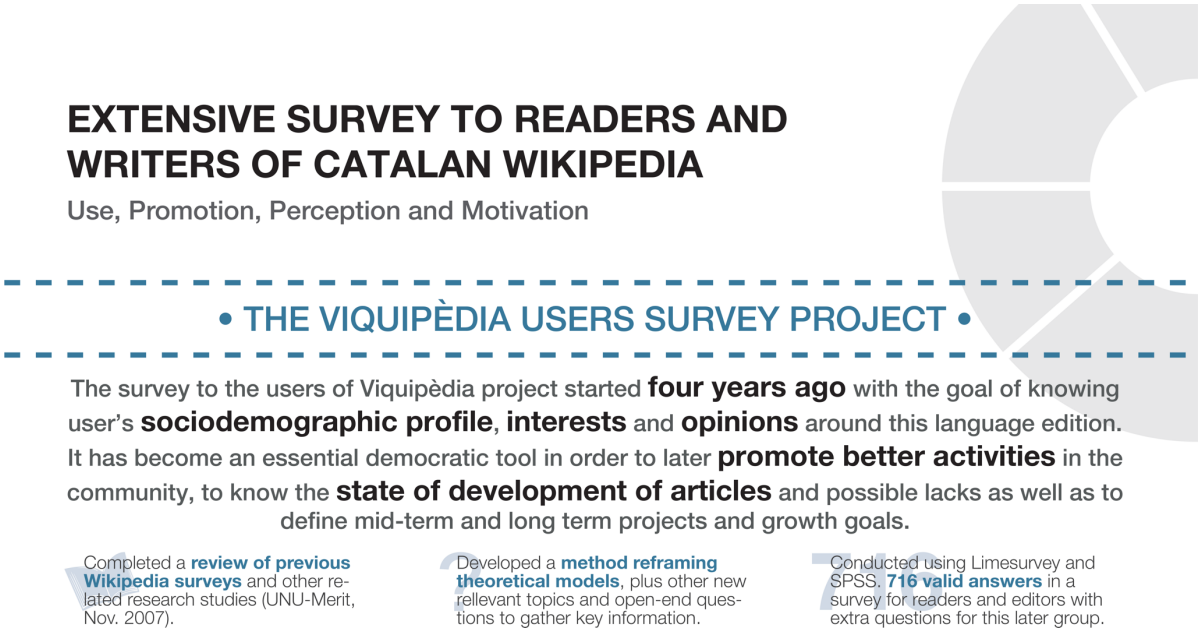
a) Highlight from the survey report: section "EDITION. ¿WHY AND HOW?":

According to the survey of Catalan Wikipedia editors, the main reason editors give to explain why they started contributing to Wikipedia is to support the idea of free knowledge (25.7%). The second reason is the pleasure for the topics they are working on (20.2%). As a third reason, there is the idea of helping other people (19.90%). **The fourth reason is the dissemination of Catalan cultural heritage** (19.6%). Other given reasons were the personal enrichment and learning about different perspectives (4.6%), the fun in the activity itself (4.5%). Finally, editors consider that editing allows them to develop their writing skills (1.2%).

"La raó principal que els editors donen per explicar per què van començar a editar Viquipèdia és perquè donen suport a la idea de que la informació ha de ser lliure (25,7%). Mentrestant, la segona raó actual es tracta del gust per la pròpia in formació sobre la qual es treballa (20,2%), que ha baixat de la primera posició en l'enquesta anterior (33,3%). Com a tercera raó trobem la idea d'ajudar als altres (19,90%), ***com a quarta la difusió del patrimoni català (19,6%)****. Després, l'enriquiment personal que pot suposar veure diferents perspectives (4,6%), la diversió de la pròpia activitat (4,5%). En darrer lloc els editors consideren que els permet desenvolupar habilitats d'escriptura (1,2%)."*

b) Poster "Extensive Survey to Readers and Writers of Catalan Wikipedia"

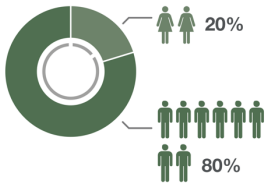
The following two pages include the poster presented in WikiAcademy Berlin 2012.



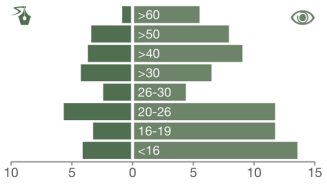
• THE AVERAGE USER PROFILE •

The average editor of Viquipèdia is a **young** Catalan speaking male, **23 years old**, born in Catalonia and with a **degree** or **master** level of studies

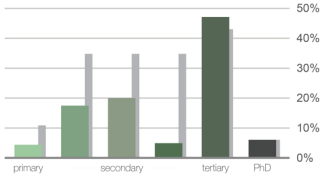
User's **gender** distribution



User's **age and role** distribution

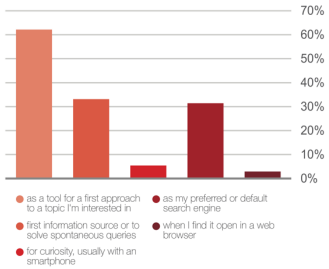


User's **level of studies** distribution



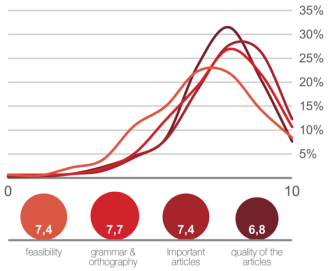
• USE AND PROMOTION •

Viquipèdia is used as the **first information source** and to **solve queries** for most of the users.

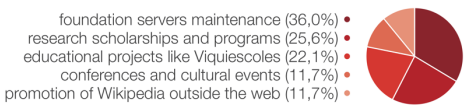


Out of the ten main areas, **humanities** has been the most queried category for the past **4 years**.

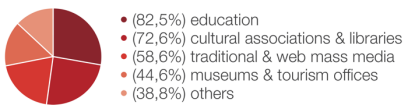
The **average mark** gave to Viquipèdia is **around 7'3**, with the lowest mark in a core area over 6'5.



Fund-raising investment preferences



Areas of promotion preferences

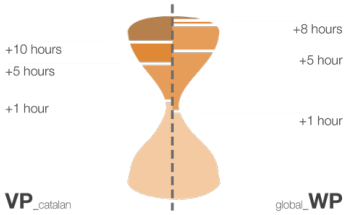


• EDITION. ¿WHY AND HOW? •

The **motivation** of the Catalan editors is **generally equivalent** to the obtained in other global surveys.

free knowledge (26%), I like the topics I write about (20%), to help others (20%), to promote catalan heritage (19%), other reasons (10%)

Catalan editors spend **two hours a week** more than their international fellows editing Wikipedia.



Over **100 suggestions and comments** where placed by the users and editors.

- Improvement and simplication of the text editor, which lacks usability when when editing.
- abc — Avoid the incomplete or wrong translations of articles from Spanish Wikipedia and focus on translating from more complete Wikipedias (FR or EN)
- Reach specialized editors with university, master & PhD studies to "supervise" the topics in which they are specialized to avoid content errors or bias.
- Avoid politic or ideological bias, specially in politics and history related topics.

• CONCLUSIONS AND SUGGESTIONS •

a We found out that the survey could be used as a **democratization and cohesion tool**. And basing it on previous research we could reframe it and extend it to different and more accurate purposes.

b We saw **Catalan Wikipedia does not differ** from other wikipedias in terms of **socio-graphics**, neither **perception** and **motivation**. Yet it can channel its interests in projects with highly motivated people: 100 quality articles, Viquiescoles, etc.

c We saw quantitative method was important to measure what we expected, but **qualitative answers gave insights** like the need of improving on usability and communication aspects, better translations or supervision by academics.

Amical Viquipèdia. 2012. Survey to readers and writers of Catalan Wikipedia Project. CC-BY-SA
David Morera Ruiz. Architect & MSc in Land Management // david.morera@mail.com
Marc Miquel Ribé. PhD Student, Telecoms Engineer & Humanist // marcmiquel@gmail.com
Joan R. Gomà Ayats. PhD Industrial Engineer // jrgoma@gmail.com

1.2 Full Report in Catalan

AMICAL VIQUIPÈDIA
Transmetent coneixement lliure

Associació Amical Viquipèdia
recerca@viquimedia.cat
www.viquimedia.cat

IV SONDEIG A LECTORS I EDITORS DE LA VIQUIPÈDIA

INFORME PRELIMINAR. GENER DE 2012.

M. Miquel (Amical Viquipèdia), D. Morera (Amical Viquipèdia).

amicalviquipèdia

1. INFORMACIÓ GENERAL

El *IV Sondeig a lector i editors de la Viquipèdia* és una iniciativa promoguda des de l'Associació Amical Viquipèdia per tal de conèixer i fer difusió del perfil socio-demogràfic, interessos i opinions dels usuaris del projecte. Sobre aquesta base es volen perfilar millor les activitats de l'associació, conèixer les inquietuds generals d'editors i lectors, determinar la difusió del projecte i plantejar objectius a curt i llarg termini.

L'edició de 2011 de l'enquesta s'ha pogut completar durant el mes de desembre a través de la plataforma virtual *limesurvey* hostatjada als servidors de Softcatalà. Durant els 30 dies que l'enquesta ha estat activa l'han respost un total de 766 usuaris, passat el filtratge s'han considerat **716 respostes vàlides**.

S'ha fet servir tots els mitjans disponibles per filtrar les respostes malintencionades, tot i així, la validesa de les respostes dels lectors té un valor purament informatiu. Els resultats a editors ofereixen un índex de confiança del 90% amb un error del 5%, al no tenir dades d'usuaris únics no es pot establir la fiabilitat genèrica.

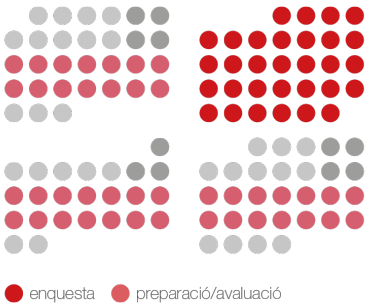
En la present edició s'ha estructurat en quatre blocs, per determinar; característiques sociodemgràfiques dels usuaris, ús de la viquipèdia, percepció del projecte i valoració del procés d'edició. El conjunt de blocs es composa de 27 preguntes de resposta múltiple o oberta.

Aquest informe preliminar desglossa sintèticament els resultats de les diferents preguntes i estableix una primera aproximació a les opinions expressades per lectors i editors.

2. CALENDARI

Les preguntes i la elaboració general de l'enquesta, que continua amb la feta durant el 2010; va tenir lloc durant les últimes setmanes de novembre de 2011. Posteriorment es va pujar el contingut al servidor web i es va activar l'enquesta des de l'1 al 31 de desembre.

L'anàlisi dels resultats s'ha desenvolupat durant el mes de gener amb programari de gestió estadística i permet donar a conèixer l'informe preliminar. A finals del mes de febrer està previst donar a conèixer l'informe definitiu dels resultats.

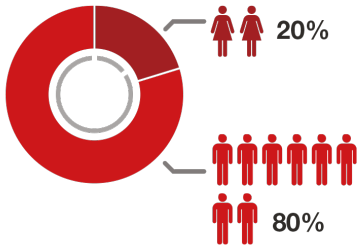


3. PERFIL SOCIO-DEMOGRÀFIC DELS USUARIS

A nivell general és continua demostrant la tendència de 2009 i 2010, els usuaris de la Viquipèdia provenen en la seva majoria de Catalunya, sense un nivell educatiu o edats específics, però amb una clara predominança d'homes.

Dades de gènere

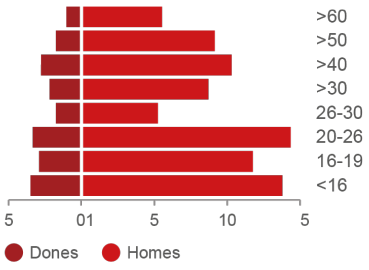
La gran majoria d'usuaris, lectors i editors, de la VP continuen sent homes (un 79,7%), tanmateix aquesta dada suposa un creixement de quasi el 5% de dones usuàries de la VP respecte el 2010.



Si es consideren només els usuaris-editors el nombre de dones es desploma fins al 8,5%, un valor que s'ha mantingut constant al llarg de totes les edicions del sondeig. A nivell de lectors la distribució s'equilibra lleugerament i s'observa com quasi un 23% dels usuaris són dones en front a un 77% d'homes.

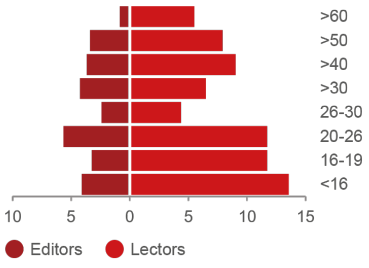
Distribució per edat

La piràmide demogràfica d'usuaris mostra de manera clara que més de la meitat dels visitants són menors de 26 anys (50,4%). Seguidament s'hi troba el grup entre 30 i 50 anys (11,1%, 13,3%, 11,2% respectivament) que acumula un 35.6% dels usuaris.



Els grups d'edat menys representats són els compresos entre els 26-30 anys i els majors de 60. En el cas de majors de 60 aquest fenomen queda clarament explicat per la barrera digital existent i l'aprimament paral·lel de la piràmide poblacional general.

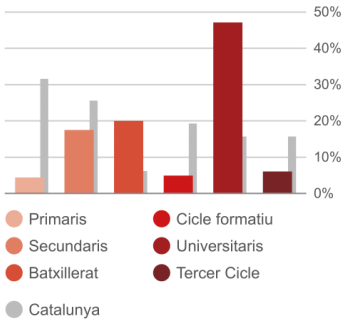
Al segregar la piràmide poblacional entre lectors i editors queda patent que hi ha dos grans grups d'editors actius de VP: Menors de 26 anys i entre 30 i 50 anys. En el cas de la gent gran es confirma que les dificultats a l'hora d'editar fan d'aquest grup el menys participatiu.



A nivell de lectors trobem que la majoria d'usuaris provenen d'entre els col·lectius més joves (37,3%) i d'una manera més o menys homogènia de les altres franges d'edat.

Nivell d'estudis

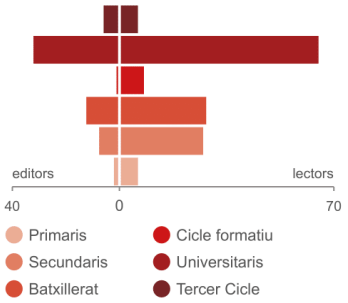
El nivell d'estudis de la mostra estudiada indica de manera molt clara que la gran majoria d'usuaris de la VP tenen una titulació universitària (47,1%), seguits pel grup amb estudis de batxillerat (20,0%) i estudis secundaris (17,5%). De manera marginal hi ha usuaris amb estudis primaris (4,4%), de cicle formatiu (4,9%) i doctorats (6,1%).



Al comparar les xifres amb el nivell d'instrucció a Catalunya es pot apreciar com hi ha un ús especialment destacat per part de persones amb formació de batxillerat o universitat. En el cas dels estudis de tercer cicle universitari (doctorat i post-doctorat), no és possible establir comparativa al no existir dades desglossades a nivell català.

En la banda contraria es pot apreciar un ús molt inferior en els grups que formen els titulats en cicles formatius i amb estudis primaris. En el cas dels usuaris amb estudis primaris es pot apreciar que en la seva majoria (71%) són menors de 12 anys, encara que a nivell català es tracta de gent major de 65 anys.

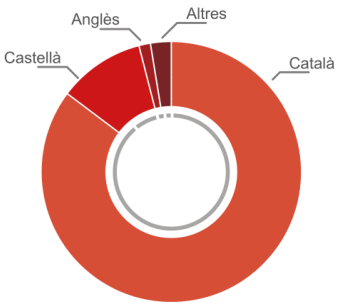
S'ha constatat que la majoria dels editors tenen formació universitària (32%) o de batxillerat (12%). En aquest sentit, tot i existir una base de lectors d'aquest nivell molt important, s'han recollit opinions d'usuaris amb formació primària o secundaria que es veuen poc capacitats per editar.



Quasi no hi ha editors de cicle formatiu (1,5%) tot i que no s'ha trobat cap raó clara que pugui explicar aquest fenomen i la relació lector/editor és la més desequilibrada de tots els grups (1-6). En canvi, a nivell de tercer cicle la proporció lector/editor és la més elevada (1-1), tot i ser el grup amb menys usuaris totals.

Llengua d'ús preferent

De manera predominant els usuaris de la VP fan servir el Català com a primera llengua (84,8%). Tanmateix, respecte les edicions anteriors del sondeig es pot observar un increment dels usuaris castellanoparlants (+4,0%) fins arribar a superar el 10,4%. Ambdós grups concentren al voltant un 25% d'editors sobre el total d'usuaris, una proporció inferior a la de col·laboradors d'altres llengües.



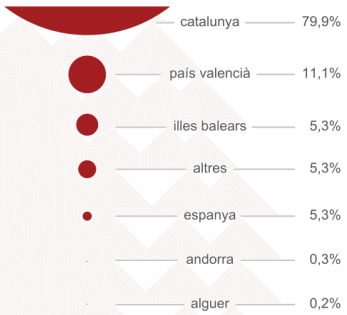
Per contra s'han perdut usuaris procedents de la llengua anglesa (-0,7%), encara que es tracta d'un grup molt minoritari (10 usuaris) que compta amb una proporció lectors/editors elevada (1-1).

Finalment, es pot apreciar que hi ha usuaris amb llengua preferent francesa, italiana, alemanya, occitana, neerlandesa i portuguesa. En aquests casos es tracta, generalment, d'usuaris editors que participen activament en diferents àrees del projecte.

Residència

Les dades de lloc de residència dels viquipedistes són clares: quasi un 80% dels usuaris provenen de Catalunya, seguits molt lluny pels originaris del País Valencià (11,1%) i les Illes Balears (5,3%). De manera marginal hi ha usuaris d'Andorra (0,3%) i l'Alguer (0,2%).

D'altres indrets de l'Estat Espanyol a són originaris un 2,3% dels usuaris. Finalment, també hi ha representació de residents d'altres països com França, el Regne Unit, Nova Zelanda, Suïssa, Mèxic, etc.



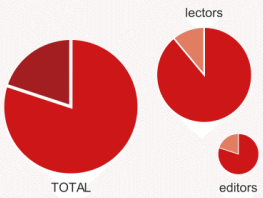
Rol

La proporció actual respostes de lector-editor és de 1-4, per cada cinc usuaris de la VP un és editor actiu (20%). Aquesta proporció s'ha reduït respecte el 2010, quan hi havia un 30% d'editors. Cal tenir em compte, però que en nombre absoluts hi ha 1.504 editors actius sobre un total de 105.466 editors registrats (1,4%).

La distribució per sexe indica clarament que la participació activa en la VP, com en el rol d'editor, és molt major en homes que en dones (92-8). Aquesta predominança s'atenua en el cas de lectors, on un 23,5% són dones

Coneixement d'Amical Viquipèdia

El coneixement de l'associació Amical Viquipèdia està clarament més estés entre els editors (40%) que els lectors (11%). A nivell global quasi 1 de cada 5 usuaris que ha respost l'enquesta coneix l'associació.



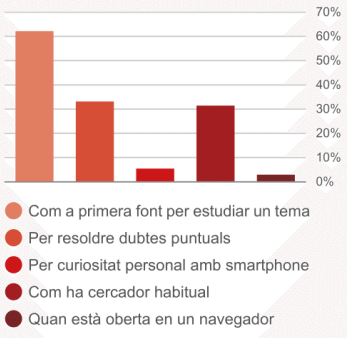
Els usuaris que coneixen l'activitat d'Amical ho han fet en base a les activitats WikiLoves Monuments (46,5%), les Xerrades (21,8%), Viquiescoles (17,8%), Altres (17,8%) i Viquifabricació (8,7%).

4. ÚS DE LA VIQUIPÈDIA

La VP és la primera font d'informació a la que recorren els usuaris quan estudien un tema concret. Quasi la meitat dels lectors del projecte ha editat algun cop i cal destacar que les temàtiques consultades es troben en un equilibri creixent, tot i predominar les ciències humanes.

Context d'ús de la Viquipèdia

L'ús principal de la VP és com a eina primària a l'hora de començar a estudiar un tema concret (62%). Així mateix, destaca com l'ús de la VP com a eina per resoldre dubtes puntuals (33%) i com a cercador general habitual (31%). De manera marginal hi ha un ús des de dispositius mòbils (5%) o quan es troba la pàgina oberta en un navegador (3%), cal destacar, però que es tracta de perfils poc predisposats.



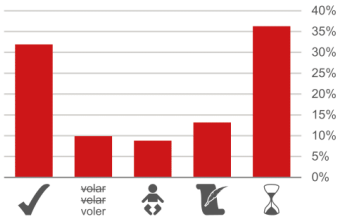
Al estudiar les dades en profunditat es pot veure com l'ús de VP com a primera font d'estudi d'un tema és un context d'ús generalitzat en totes

les franges d'edat, especialment entre els col·lectius més joves (menors de 19 anys) on arriba al 75%. De manera semblant passa amb els usos en contexts secundaris, on no hi ha un patró d'edat determinat. A nivell d'editors, la gran majoria dels editors fan servir VP com a motor de cerca habitual (63,1%).

Edició d'articles per part dels lectors

Quasi un de cada dos lectors ha provat en algun moment d'editar un article (46%), no existeix un perfil clar del lector que edita de manera puntual més enllà del ja descrit pels editors habituals.

La gran majoria dels lectors que en algun moment han fet d'editors van fer edicions puntuals (51,9%). Deixant de banda els editors puntuals, en el cas els editors que han abandonat aquest rol ho van fer perquè van deixar de tenir el temps necessari (36,3%) o perquè van trobar l'editor o la guia d'estil són massa complicades (31,9%). Seguidament trobem aquells que consideren que no tenen el nivell d'escriptura suficient per col·laborar (13,2%), els van deixar d'editar pels canvis posteriors en articles seus (9,9%) i els que no es van sentir ben acollits com a nouvinguts (8,8%).



Temàtiques consultades

No existeix una temàtica consultada amb un clar predomini sobre les altres, encara que destaquen les Ciències Humanes (17,2%). En un segon esglaó, i amb una distribució homogènia, s'hi troben les temàtiques de Ciències Bàsiques (11,5%), Tecnologia (11,2%) i Lletres (11,8%). Seguidament hi ha els temes de Cultura i Oci (10,4%) i articles relacionats amb els Països Catalans (10,1%). Finalment es consulten articles en relació a Ciències de la Terra (7,6%), Arts (7,4%), Ciències de la Vida (6,8%) i Ciències Polítiques (6,5%).

En relació a l'edició de 2010 de l'enquesta destaca un decreixement clar les Ciències Humanes, encara que queda en primer lloc. La distribució dels altres temes es manté estable respecte l'any anterior.

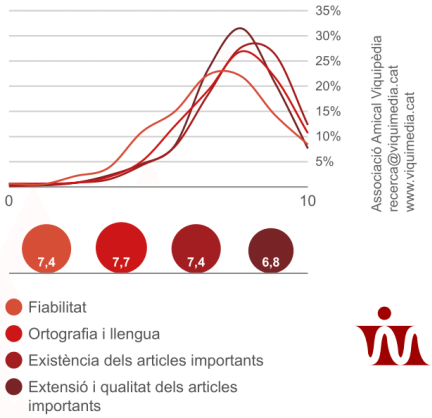
5. PERCEPCIÓ

Els usuaris consulten la Viquipèdia a la cerca d'informació per motius molt diversos, tanmateix la VP és un projecte que engloba diferents aspectes més enllà del purament enciclopèdic. En aquest sentit s'ha cercat en quins aspectes els usuaris creuen que s'hauria de promocionar la VP i invertir els diners de les donacions.

La VP està ben valorada en el conjunt dels àmbits que la composen, tanmateix hi ha altres aspectes menys coneguts que permeten verificar la qualitat d'un article que generalment no són coneguts pels usuaris.

Puntuació de la Viquipèdia per àmbit

Les respostes dels usuaris entorns valoració de quatre aspectes formals i de qualitat dels continguts han estat molt favorables. Tal com es pot apreciar a la gràfica, l'ortografia i llengua és un punt fort amb un 7,7 que obté de mitjana. En segon lloc, s'hi troben tant la fiabilitat dels articles com l'existència dels articles importants (7,4). I en darrer i quart lloc, es valora amb un 6,8 l'extensió i qualitat dels articles importants.



No obstant aquest darrer punt és el pitjor valorat, també cal destacar que és el que obté més consens en el percentatge de respostes. El que té més variació és el de la fiabilitat, un dels aspectes més discutits de l'enciclopèdia lliure.

Articles de Qualitat

Els articles de qualitat es pot veure que són poc coneguts tant per lectors com per editors. Només un 14% dels lectors ha afirmat conèixer-los i ha proposat una temàtica en la qual se'n podrien crear més. Mentre que el percentatge de coneixement augmenta a 33,3% entre editors.

Entre les temàtiques generals que es proposen articles dels àmbits socials (política, història, art, música), però també es destaca que els de ciències són millorables. També hi ha mencions a crear articles de qualitat sobre aquelles temàtiques pròpiament catalanes. Tanmateix, és difícil veure consens.

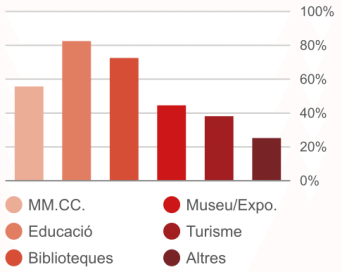
Valoració de les referències dels articles

Un dels aspectes fonamentals de Viquipèdia són les referències dels articles. Trobem que el percentatge de lectors que saben valorar-los (44%) és lleugerament superior al que no (39%), per bé que hi ha una part important que s'abstè (17%). Sens dubte, són resultats que marquen un dels aspectes en els quals s'ha de treballar i fer pedagogia per aconseguir augmentar el nivell de qualitat.

Promoció de la Viquipèdia

El primer dels àmbits en els quals els lectors creuen que Viquipèdia podria tenir major repercussió i per tant ser més promocionada és en l'educació (82,5%) – en tots els nivells. En segon lloc, també s'hauria de potenciar-ne l'ús en les associacions culturals i biblioteques (72,6%).

Posteriorment es creu que els mitjans de comunicació tradicionals també n'haurien de fer més difusió (58,6%). També als museus i exposicions es creu que podria tenir més promoció (44,6%). I en l'àmbit turístic i d'excursionisme (38,8%). Cal remarcar que una de cada cinc (21,1%) respostes ha cregut que hi ha altres àmbits en que podria ser més promocionada.



Campanyes de *Fundraising* de la Fundació

Una mica més de la meitat dels lectors coneix les campanyes de donacions individuals *fundraising* a la Fundació Wikimedia (55%). Un 35% no les coneix i un 10% s'absté de respondre.

Al preguntar en quins projectes seria més convenient invertir-hi els lectors consideren que el més prioritari és el manteniment dels servidors (36,0%). Seguidament valoren la inversió en beques de recerca (25,6%), projectes com Viquiescoles (22,1%) i conferències i esdeveniments culturals (11,7%). Només un 11,7% de les respostes han considerat important la promoció de Viquipèdia al carrer.

Suggeriments als editors

Els suggeriments fets en l'enquesta són una resposta oberta que s'ha agrupat per tal de poder donar una visió general de quines millores es creu que podrien afavorir la VP. Els punts recollits no engloben el total de respostes, tan sols aquelles més significatives.

1. Millora i simplificació de l'editor de text. Es demana més facilitat i usabilitat, permetent una interacció més visual tipus OpenOffice, MS Word o Wordpress.
2. Evitar les traduccions parcials o incorrectes d'articles de la Wikipedia en castellà i centrar les traduccions de les WP francesa i anglesa.
3. Arribar a més editors especialitzats amb estudis universitaris o doctorals per "supervisar" els temes en els quals són especialistes i evitar errors en el contingut.
4. Evitar biaixos polítics o ideològics, especialment en política i història, amb les conseqüents guerres d'edicions.

Associació Amical Viquipèdia
recerca@viquimedia.cat
www.viquimedia.cat

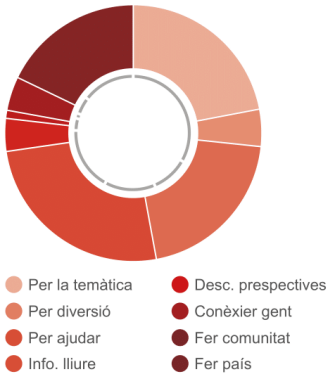
amicalviquipèdia

6. EDICIÓ

La VP és la primera font d'informació a la que recorren els usuaris quan estudien un tema concret. Quasi la meitat dels lectors del projecte ha editat algun cop i cal destacar que les temàtiques consultades es troben en un equilibri creixent, tot i predominar les ciències humanes.

Motius pels que es comença a editar

La raó principal que els editors donen per explicar per què van començar a editar Viquipèdia és perquè donen suport a la idea de que la informació ha de ser lliure (25,7%). Es tracta d'una raó que en l'anterior versió de l'enquesta ocupava la segona posició amb un 17,6%.

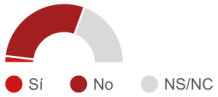


Mentrestant, la segona raó actual es tracta del gust per la pròpia informació sobre la qual es treballa (20,2%), que ha baixat de la primera posició en l'enquesta anterior (33,3%). Com a tercera raó trobem la

idea d'ajudar als altres (19,90%), com a quarta la difusió del patrimoni català (19,6%). Després, l'enriquiment personal que pot suposar veure diferents perspectives (4,6%), la diversió de la pròpia activitat (4,5%). En darrer lloc els editors consideren permetre conèixer gent i desenvolupar habilitats d'escriptura (1,2%).

Canvi en els motius per editar

Dels motius o raons només un 3,8% dels editors creu que han canviat al llarg de la seva vinculació, un 57,3% pensa que es manté pels mateixos i un 38,9% no respon. Precisament, els motius que ara motiven aquells que creuen que han canviat de raó per la qual escriuen són els tres primers que han obtingut més respostes com a raons principals.



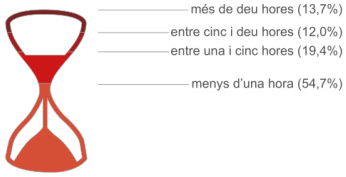
Millora de la benvinguda als editors

El sistema de benvinguda als nouvinguts és considerat correcte per la majoria dels enquestats, tanmateix hi ha diferents comentaris encarats a millorar el sistema.

1. Benvinguda més visual i simplificada, deixant clares les guies d'edició i els pilars de la VP.
2. Explicar el funcionament del projecte (discussions, viquiprojectes, taverna, etc.) i com funciona l'editor de texts.
3. Evitar les agressions i mossegades als nouvinguts, sobretot en casos d'articles al límit de l'admissibilitat.
4. Millorar i fer més entenedores les plantilles, especialment les de destrucció o manca de referències.
5. Seguiment personalitzat dels editors novells amb l'assignació d'un tutor o mentor i crear un sistema de "reconeixement" de tasques.
6. Elaboració de tutorials pas a pas per editar o començar una pàgina.
7. Fer més publicitat entre novells dels taller d'edició.

Temps dedicat a editar

Segons la dedicació dels editors que han resposta l'enquesta, podem entendre com Viquipèdia es fa en un 54,7% partir d'estones equivalents a una hora o menys. En segon lloc, un 19,4% diuen dedicar-hi entre 1 i 5 hores. La dedicació baixa a 12,0% per l'interval entre 5 i 10 hores, però torna a pujar a 13,7% de respostes per aquells que hi dediquen més de 10 hores.



Aspectes positius i negatius

S'ha rebut un total de 165 comentaris remarcant un aspecte positiu i un de negatiu. S'ha optat per a fer una interpretació general dels aspectes més destacables i remarcar per ordre de prioritat segons la repetició dels diferents tipus de comentari. Es pot trobar la llista completa de comentaris a la pàgina del quart sondeig a la Viquipèdia.

En línies generals com a aspectes positius destaquen:

1. Bon ambient i disposició a ajudar, especialment entre administradors i usuaris.
2. Organització per a fer projectes i creixement de l'enciclopèdia.
3. Control del vandalisme.
4. Valors i funcionament general de la Viquipèdia.
5. Qualitat dels articles de gran interès (Barcelona, Catalunya...).

Com a negatius destaquen:

1. Poca usabilitat de l'eina, que dificulta l'aprenentatge i fredor en els missatges.
2. Dificultats generals de les Viquipèdies: rigor, expressió, guerres d'edicions, etc.
3. Lentitud del sistema de comunicar entre editors.
4. Mancança de temes i profunditat en alguns articles.
5. Necessitat de créixer com a comunitat.

7. CONCLUSIONS

La quarta edició del sondeig als usuaris ha permès establir de manera més clara els perfils dels lectors i editors de la Viquipèdia, donant una base sobre la què actuar per tal d'incentivar aquells perfils de baixa participació o que no consulten VP.

Més enllà de les dades desglossades es pot observar com l'estat de salut del projecte continua sent bo, i augmenta la presencia de sectors socials generalment més allunyats del món digital. En aquest sentit, però, cal continuar treballant, especialment en la participació femenina.

En aquest sentit cal treballar en la recepció dels nouvinguts creant mecanismes nous o replantejant els existents per tal d'evitar que la taxa d'abandonament per "desencís" creixi i, així permetre la incorporació de nous editors, que necessitaran un temps d'adaptació abans de comprendre el complex funcionament global de VP.

L'assoliment d'aquests objectius no és senzill, tanmateix pot passar per una major promoció de la VP al carrer, essencialment en biblioteques i escoles. En aquest sentit hi ha una clara valoració positiva de projectes com Viquiescoles o Wiki Loves Monuments, que ajuden a difondre l'existència i rigorositat de la Viquipèdia.

Associació Amical Viquipèdia
recerca@viquimedia.cat
www.viquimedia.cat

amicalviquipèdia

Appendix 2. Cultural Identities Complementary Results

2.1 Table of Keywords

These tables are comprised in the file 'cira_keywords_list.csv' used to retrieve articles and categories containing them in their title.

Table 26. List of keywords (1/2). Languages are sorted alphabetically by their ISO code. *Keywords have been automatically generated using an ISO database, and for specific cases it has been extended using information from the specific article in their corresponding Wikipedia.*

Language	Wikipedia Language Code; keywords
Afrikaans	afwiki;afrikaans;namibia;suid-afrika;
Arabic	arwiki;موليبي;بناني;لبنان;الكويت;الأردن;;عراقي;الإريترية;مصر;جيبوتي;جزرالقمر;تشاد;البحرين;البحريني;الجزائر;العربية;الإمارات;سودان;التونسية;صومالي,الصومال;سوريا,سوري;سودان;الصومال;عربيسعودي;فلسطين;قطر;عمان;مغربي;ريتانيا;الصحراوية;اليَمَن,اليمني,يمني;ي;
Catalan	cawiki;catala;catalunya;balear,mallorca,menorca,eivissa;andorra;valencia;alguer;franja_de_ponent;
Cebuano	cebwiki;cebuano;cebu;bohol;negroe;masbate;biliran;eastern_samar;leyte;northern_samar;western_samar;guimaras;camiguin;marinduque;sulu;tawi-tawi;
Chinese	cswiki;čeština;cestina;česk;
Czech	dawiki;dansk;danmark;slesvig-holsten;grønland;færøe,føroyar;
German	dewiki;deutsch;Österreich;Liechtenstein;Luxembourg;Aargau;Ausserrhoden;Innerrhoden;Basel-Landschaft;Basel-Stadt;Glarus;Luzern;Nidwalden;Obwalden;Schaffhausen;Schwyz;Solothurn;Sankt_Gallen;Thurgau;Uri;Zug;Zürich;Bern;Fribourg;Valais;Swiss;Svizra;
Greek	elwiki;Ελληνικά;Ελλάδα;Κύπρος,Κύπριοι,Κιβρις;
English	enwiki;english;British,United_Kingdom;Australia;American,United_States;New_Zealand,Aotearoa;Canadian,Canada;Indian;Irish,Ireland;Swazi;Zambia;Vanuatu;Tonga;Tuvalu;Tanzania;Sudan;South_Africa;Solomon;Sierra_Leone;Seychellois,Seychelles;Samoa;Saint_Lucia;Rwanda;Philippines;New_Guinea;Palau;Niuean,Niuē;Nauru;Mauritian,Maurice;Marshallese,Marshall_Islands;Maltese,Malta;Malawian,Malaŵi;Liberia;Kiribati,Ribaberiki_Kiribati;Gambian,Gambia;Fijian,Viti;Micronesian,Micronesia;Eritrea;Cameroonian,Cameroun;Motswana,Botswana;Pakistan;Zimbabwe;Uganda;Trinidad;Saint_Vincent_and_the_Grenadines;Saint_Kitts_and_Nevis;Nigeria;Namibia;Mosotho,Lesotho;Kenyan;Jamaica;Guyana;Grenada;Ghana;Dominica;Kūki_Āirani;Belizean,Belice;Barbadian,Barbados;Bahamian,Bahamas;Antiguan,Barbuda;Malaysia;Brunei;Bangladesh;Puerto_Rico;Tokelau;Cocos_Islander,Cocos_(Keeling)_Islands;Saint_Helenian,Ascension_and_Tristan_da_Cunha;Montserrat;Guernesey;Indian,British_Indian_Ocean_Territory;American_Samoan,Sāmoa;Anguilla;Bermuda;Virgin_Islander,British_Virgin_Islands;Caymanian,Cayman_Islands;Christmas_Island,Christmas_Island;Falkland_Islander,Falkland_Islands;Gibraltar,Gibraltar;Guamanian,Guam;Hong_Kong;Manx,Ellan_Vannin_or_Mannin;Channel_Islander,Jersey;American,Northern_Mariana_Islands;Norfolk_Islander;St._Maartener,Saint-Martin;American,Northern_Mariana_Islands;Pitcairn_Islander,Pitcairn_Islands;Virgin_Islander,United_States_Virgin_Islands;Turks_and_Caicos_Islander,Turks_and_Caicos_Islands;
Spanish	eswiki;español;españa;venezuela;Uruguay;peru;Paraguay;Panama;paname;Nicaragua;Honduras;Guatemal;Guinea_Ecuatorial;El_Salvador;ecuatorian,Ecuador;ecuator;Dominican,República_Dominicana;Chile;Cuba;Costa_Ricens,Costa_Rica;Colombia;Bolivia;Argentin;
Estonian	etwiki;eesti;
Basque	euwiki;euskara;euskal;nafarroa;
Persian	fawiki;فارسی;ایرانی;ایران;افغانستان;تاجیک;
Finnish	fiwiki;suomi;finland;
French	frwiki;français;france;wallonne;bruxelles-Capitale;bénin;Burkinabé,Burkina_Faso;burundaise,burundi;camerounais,Cameroun,Québec,Ontario,Nouveau-Brunswick,Manitoba,Afrique_centrale;tchadien,Tchad;comorien,Udzima_wa_Komori;ivoirien,Côte_d'Ivoire;congolais,Congo;Djibouti;équato-guinéen,Guiné_Equatoriale;gabonaise,Gabon;guinéen,Guinée;haïtienne,haïti;luxembourg;malgache,Madagascar;malien,Mali;monégasque;Monaco;nigérienne,Niger;rwandais,Rwanda;Senegalese,Sénégal;Seychellois,Seychelles;Vanuatu;Togo;Genève;Vaud;Neuchâtel;Jura;Fribourg;Valais;
Guarani	gnwiki;Guarani;bolivia;corrientes;paragua;brasil;
Hebrew	hewiki;יהודי,יהדות,ישראל,עברית;
Hungarian	huwiki;magyar;magyarország;vojvodina;lendva;vajdaság;
Indonesian	idwiki;indonesia;
Icelandic	iswiki;íslenska;Ísland,íslen;

Table 27. List of keywords (2/2). Languages are sorted alphabetically by their ISO code. *Keywords have been automatically generated using an ISO database, and for specific cases it has been extended using information from the specific article in their corresponding Wikipedia.*

Language	Wikipedia Language Code; keywords
Italian	itwiki;italia;Ticin;vatican;Sammarines,san_marin;
Japanese	jawiki;日本;パラオ;
Korean	kowiki;한국,북한;
Macedonian	mkwiki;македон;
Malay	mswiki;Melayu;Malaysia;Brunei;Singapor;Cocos;Indonesia;
Nepali	newiki;नेपाली;नपल;नेपाल;
Dutch	nlwiki;nederlands;Brussels;Vlaams;Suriname;St._Maartener,Sint_Maarten;Aruban,Aruba;
Norwegian	nowiki;norsk,bokmål,Norge;
Polish	plwiki;polsk;
Portuguese	ptwiki;português;portug;Angola;Brasil;Cabo_Verde;timorenses,Timor-Leste;guineense,Guiné-Bissau;guineense_Equatorial,Guiné_Ecuatorial,Moçambique;são-tomense,São_Tomé_e_Príncipe;
Romanian	rowiki;român;moldov;vojvodina;
Russian	ruwiki;русский;Россия;Белоруссия;Казахстан;Киргизия;росси,русск,сове́тски,украин,белорус,беларус,сибир,урал,грузин,абхаз,молдав,армян,казах,бурят,дагенстан,осетин,чечен,татар,алтай,адыгей,калмык,кыргыз,абха́з,приднестро́вс,гагауз,крым,таджики,азербайдж,саха́,яку́тия,Приднестров
Serbian	srwiki;срвики;Српски;Србија;Босна,Херцеговина;Комори;
Swedish	svwiki;svensk;sverige;ahvenanmaan_maakunta;närpes,larsmo,åland;
Swahili	swwiki;swahili;tanzania;Kenya;Ugandan,Uganda;burundi;
Turkish	trwiki;Türk;kıbrıs;
Ukranian	ukwiki;україн;
Viatnamese	viwiki;Việt;Việt_Nam;
Waray	warwiki;Waray;Eastern_Samar;Northern_Samar;Western_Samar;Biliran;Leyte;Masbate;Sorsogon;
Chinese	zhwiki;中国,中文,香港,台湾,澳門,臺灣,澳门,新加坡;

2.2 CIRA Geolocated

As seen in Section 7.3.2 (Table 8), the average percentage of CIRA Geolocated articles does not exceed the 5% of all articles in each Wikipedia. Those with a higher percentage are languages like Guarani (13,96%) and Nepali (11,77%), followed by Russian (10,98%), Persian (10,33%) and English (9,84%). Many geolocated articles do not respond to geographic entities with a political administration such as a city or region, but are historical places, architectural elements or event locations. To further inspect the creation of CIRA geolocated articles, I counted the number of geolocated articles by municipalities or cities using the same reverse geocoder[90]. A municipality is an urban administrative division which vary according to political criteria. Yet, this aggregation gives an interesting insight on how geolocated content is concentrated in specific locations and across very different language contexts. In Figure 51, I rank the cities per language according to the density of articles assigned to their municipality. The first positions (on the right) are usually occupied by country capitals in all languages. The slope for each language indicate the number of articles dedicated to the rest of the territories of each language edition. In languages like Japanese and Basque, one city accounts for the 5 or 10%. In others like English, the first in number of articles is Stanley and only accounts for a 0.37% of all CIRA Geolocated articles. This shows that every context where cultural identities are generated and evolve is sociologically and geographically very different.

[90] https://pypi.python.org/pypi/pygeocoder

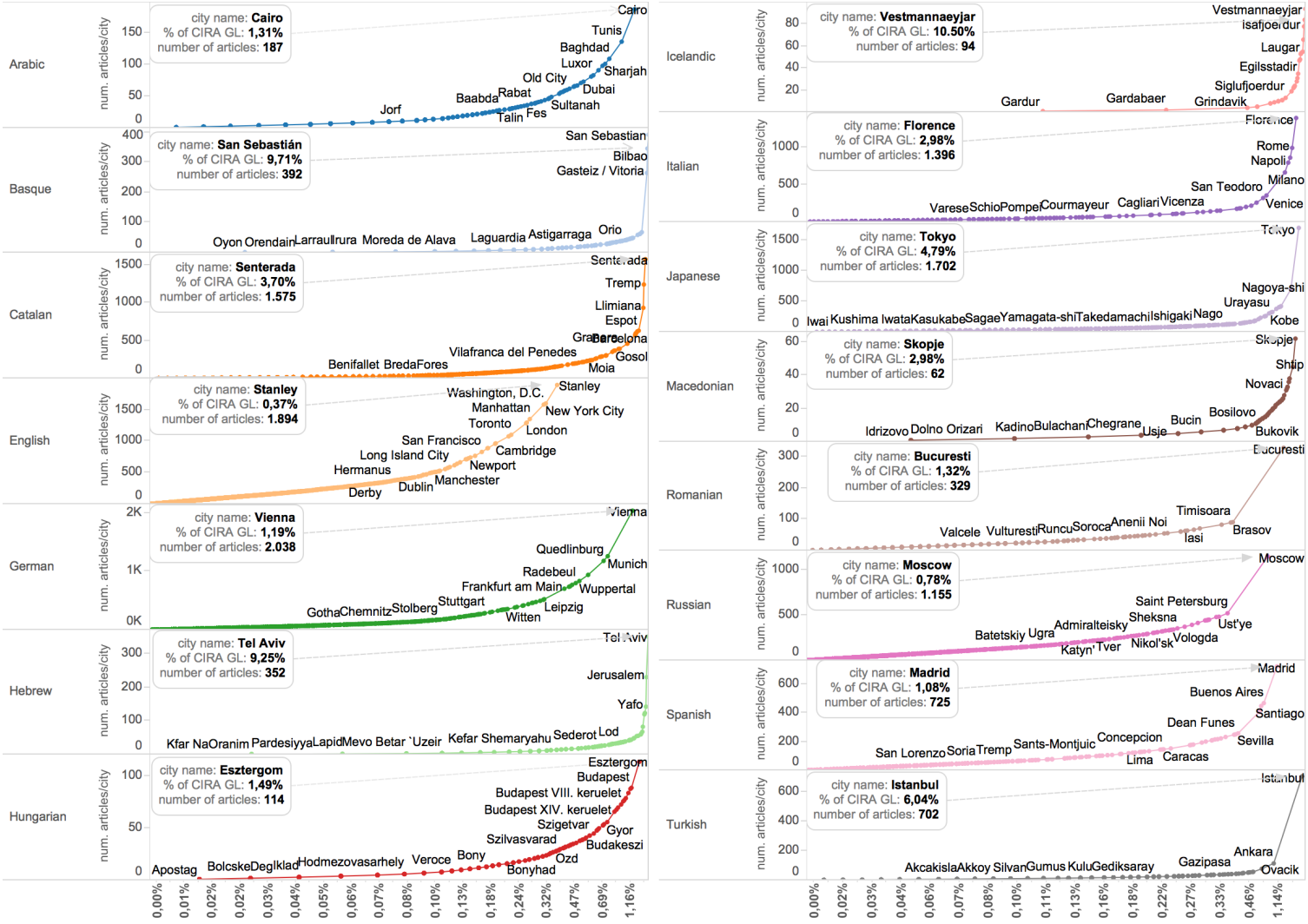


Figure 51. Ranking of cities by CIRA Geolocated articles in them, for each Wikipedia language edition.

2.3 CIRA Editor and Reader Engagement

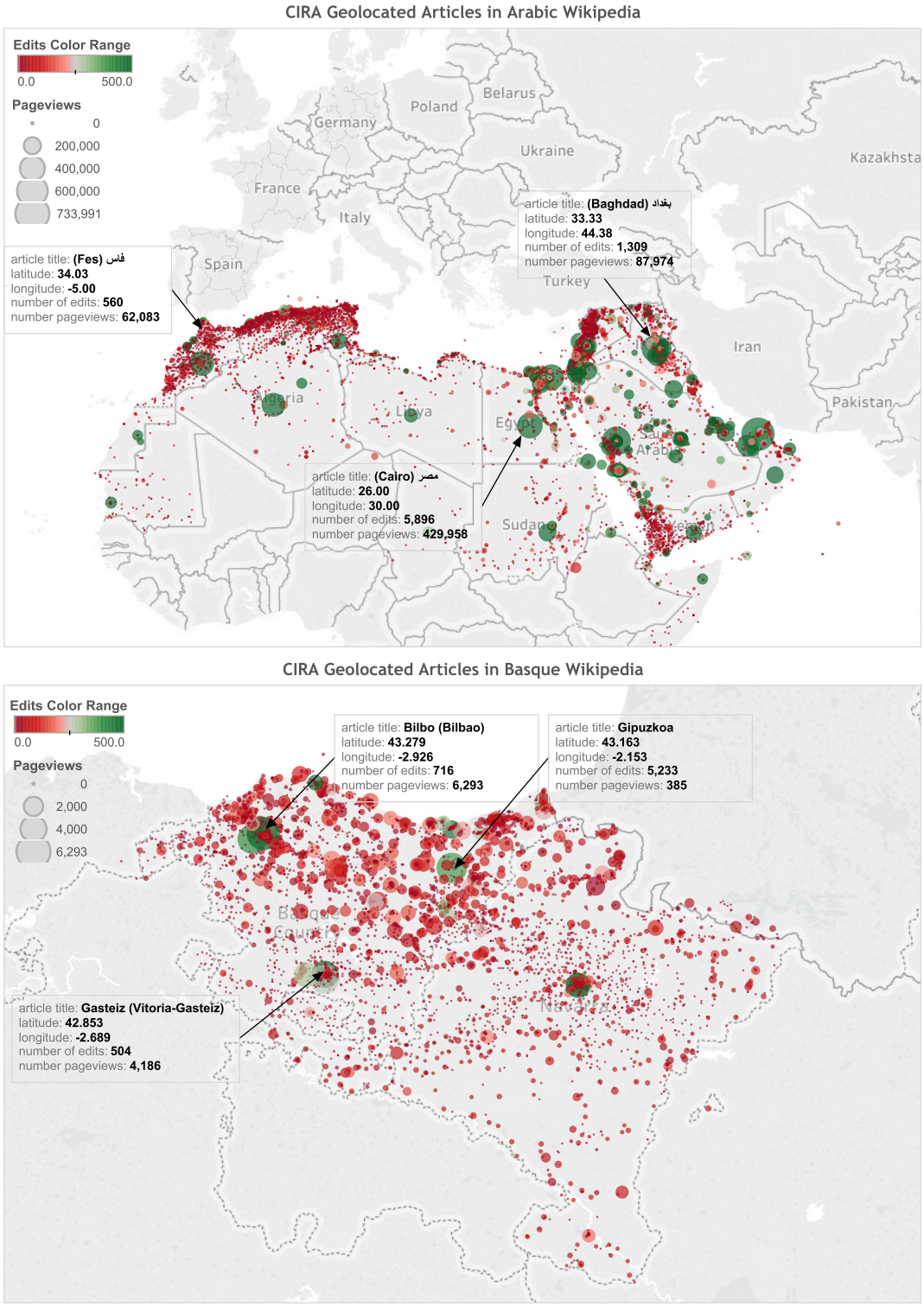


Figure 52. Editor and reader engagement in CIRA Geolocated articles from Arabic and Basque Wikipedia (top and bottom respectively). *Each point is a CIRA geolocated article. Colour represents the number of edits, depicted as a continuum from red to green with a middle point of 250 edits in colour beige. Size represents the number of page views. Important geolocated articles are marked with infoboxes.*

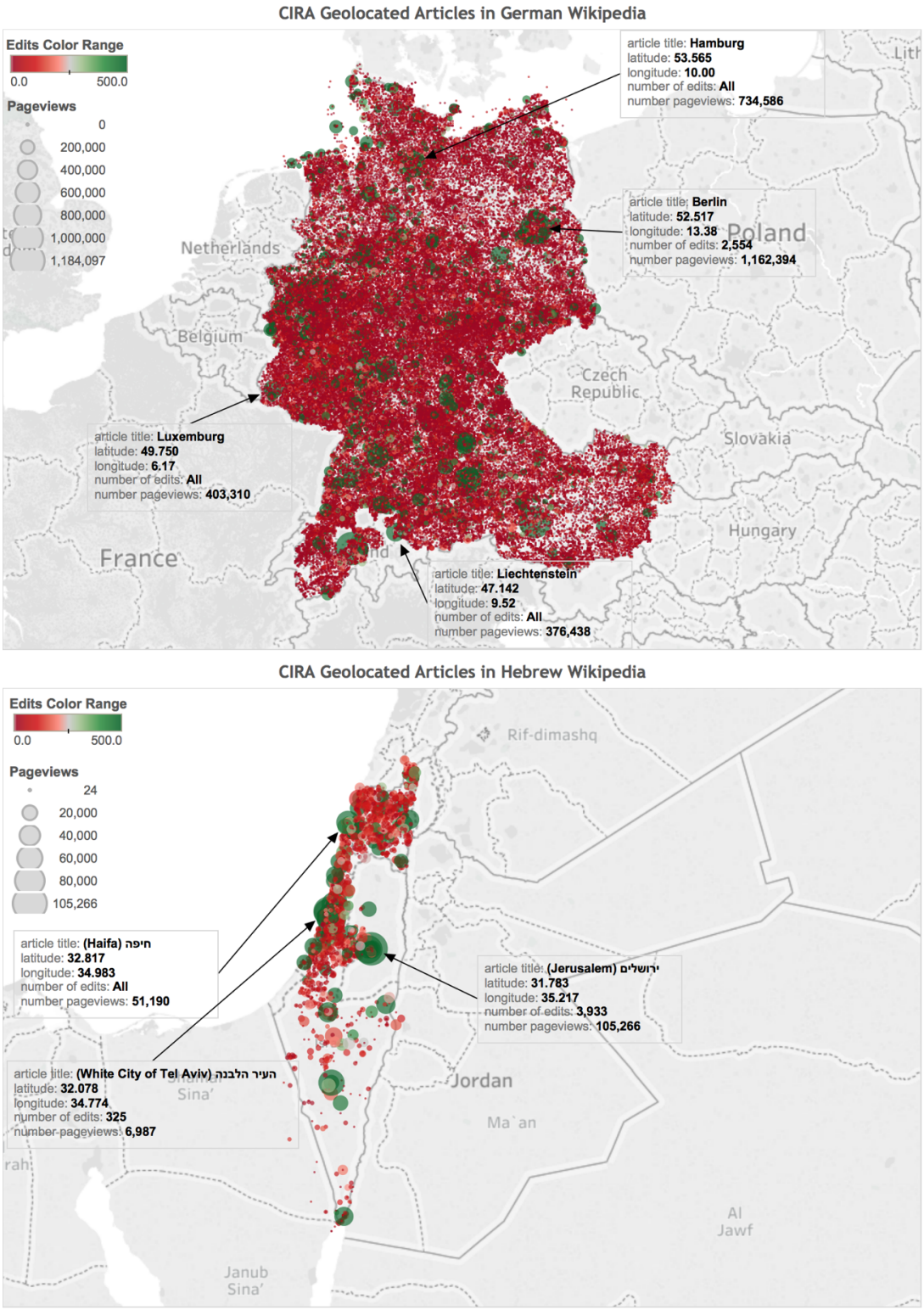


Figure 53. Editor and reader engagement in CIRA Geolocated articles from the German and Hebrew Wikipedia (top and bottom respectively). *Each point is a CIRA geolocated article. Colour represents the number of edits, depicted as a continuum from red to green with a middle point of 250 edits in colour beige. Size represents the number of page views. Important geolocated articles are marked with infoboxes.*

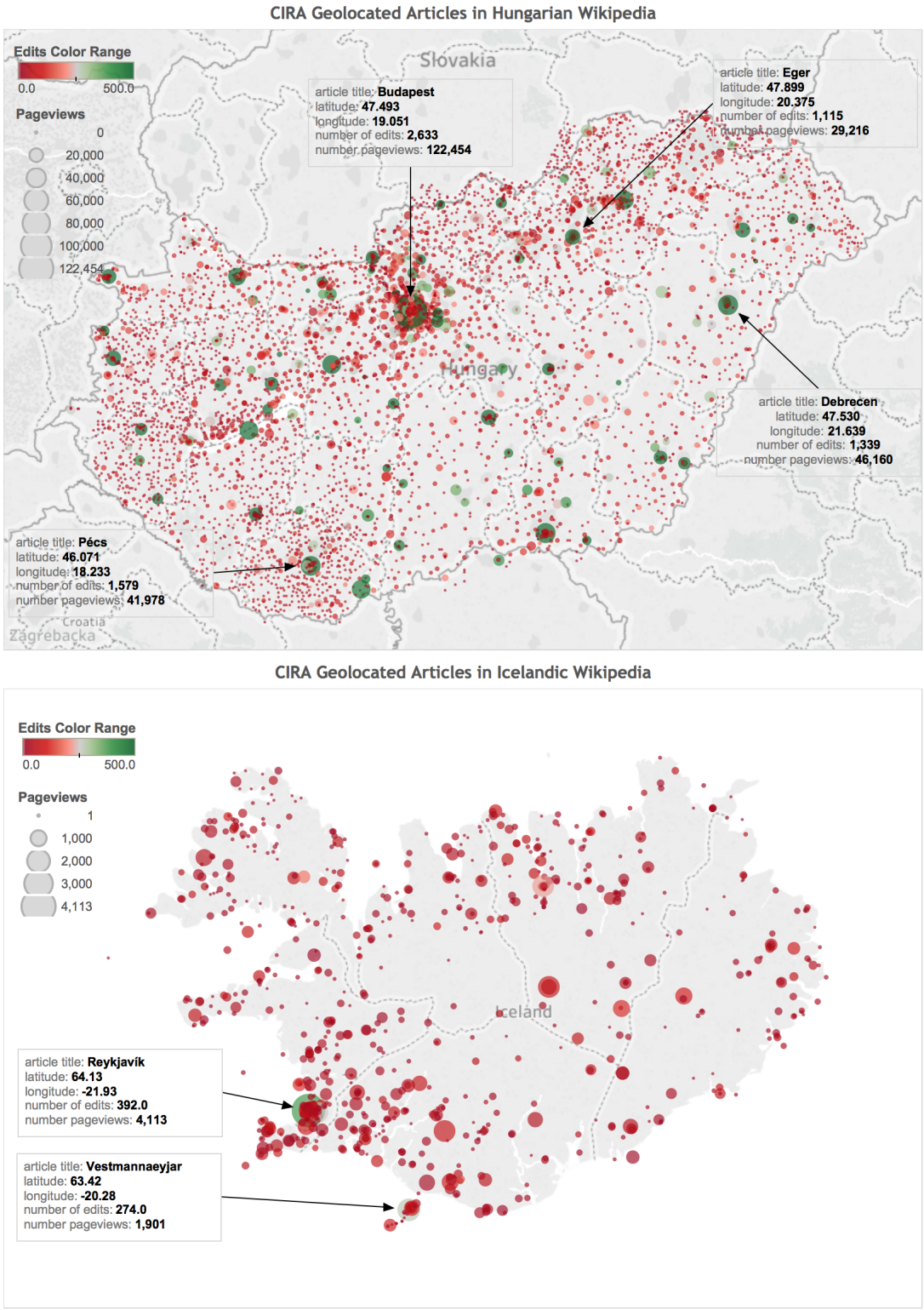


Figure 54. Editor and reader engagement in CIRA Geolocated articles from Hungarian and Icelandic Wikipedia (top and bottom respectively). *Each point is a CIRA geolocated article. Colour represents the number of edits, depicted as a continuum from red to green with a middle point of 250 edits in colour beige. Size represents the number of page views. Important geolocated articles are marked with infoboxes.*

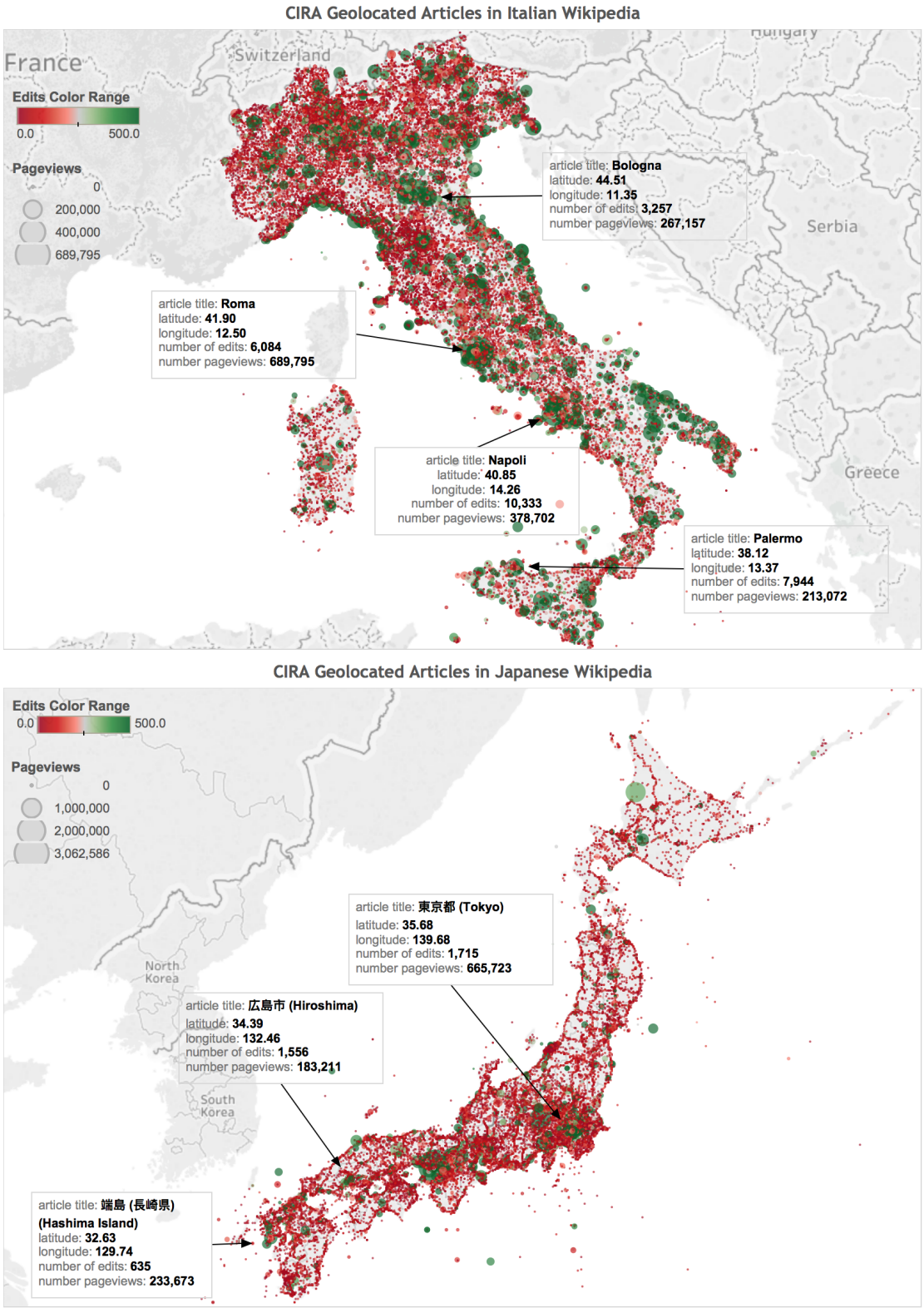


Figure 55. Editor and reader engagement in CIRA Geolocated articles from Italian and Japanese Wikipedia (top and bottom respectively). *Each point is a CIRA geolocated article. Colour represents the number of edits, depicted as a continuum from red to green with a middle point of 250 edits in colour beige. Size represents the number of page views. Important geolocated articles are marked with infoboxes.*

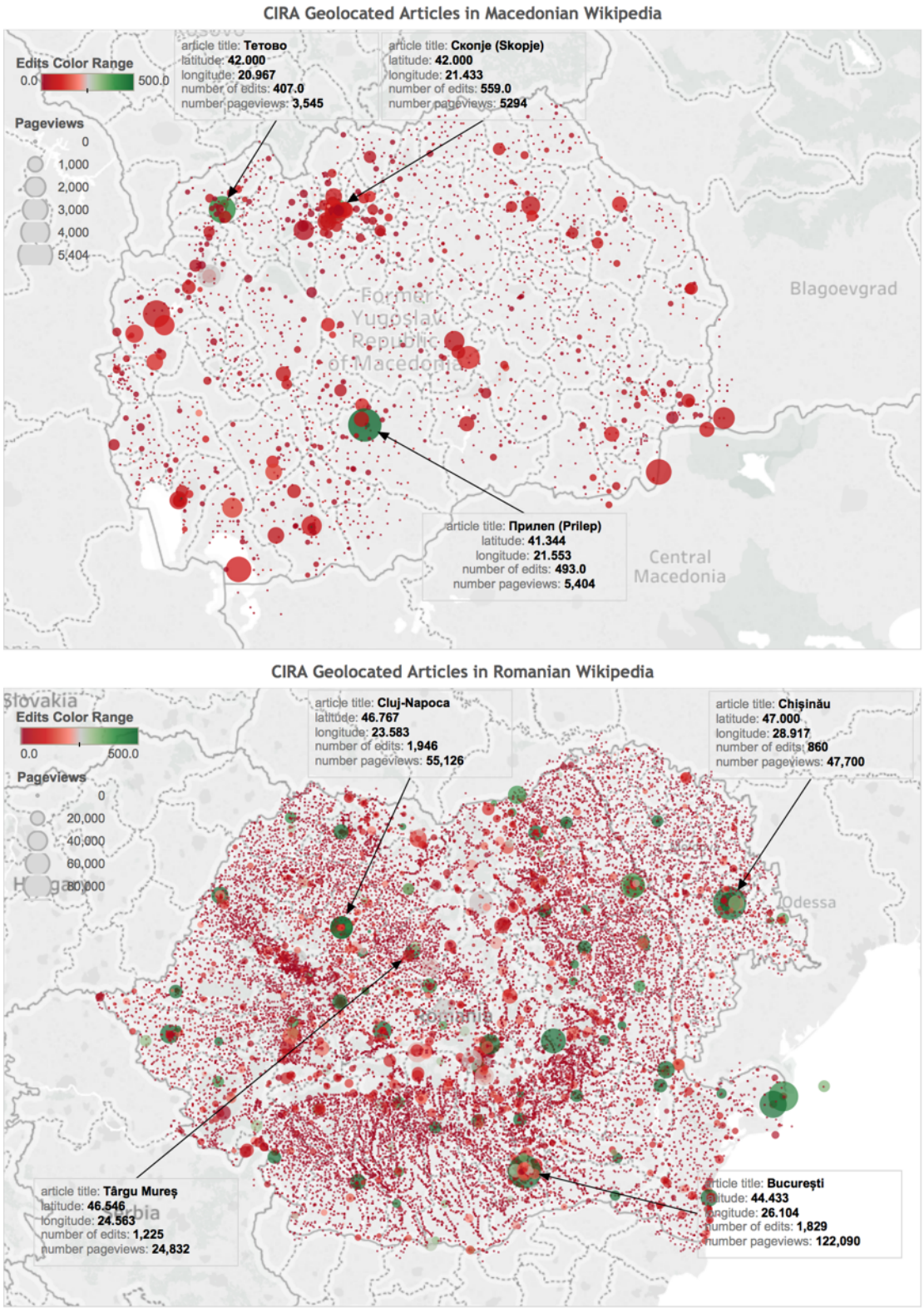


Figure 56. Editor and reader engagement in CIRA Geolocated articles from Macedonian and Romanian Wikipedia (top and bottom respectively). *Each point is a CIRA geolocated article. Colour represents the number of edits, depicted as a continuum from red to green with a middle point of 250 edits in colour beige. Size represents the number of page views. Important geolocated articles are marked with infoboxes.*

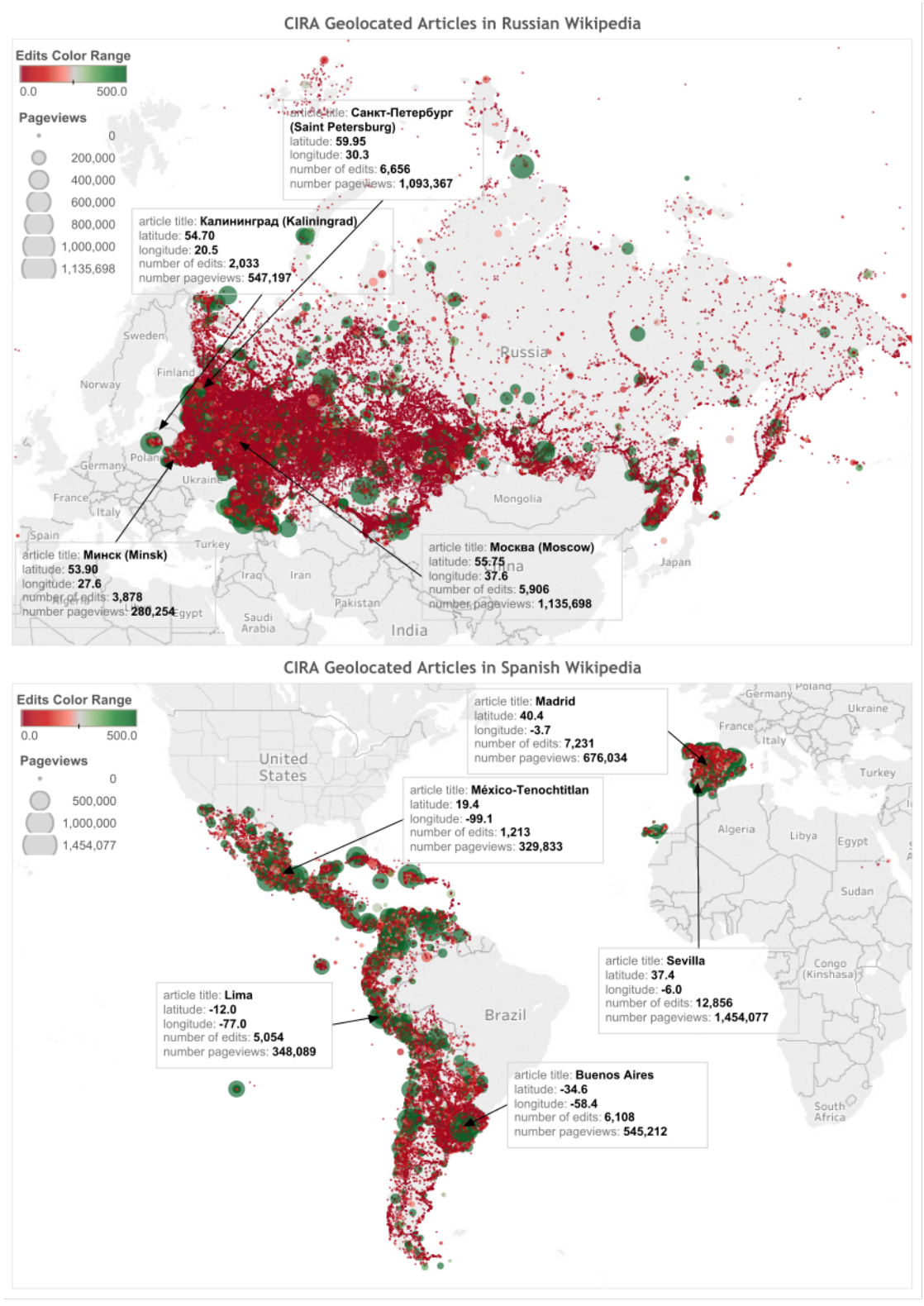


Figure 57. Editor and reader engagement in CIRA Geolocated articles from Russian and Spanish Wikipedia (top and bottom respectively). *Each point is a CIRA geolocated article. Colour represents the number of edits, depicted as a continuum from red to green with a middle point of 250 edits in colour beige. Size represents the number of page views. Important geolocated articles are marked with infoboxes.*

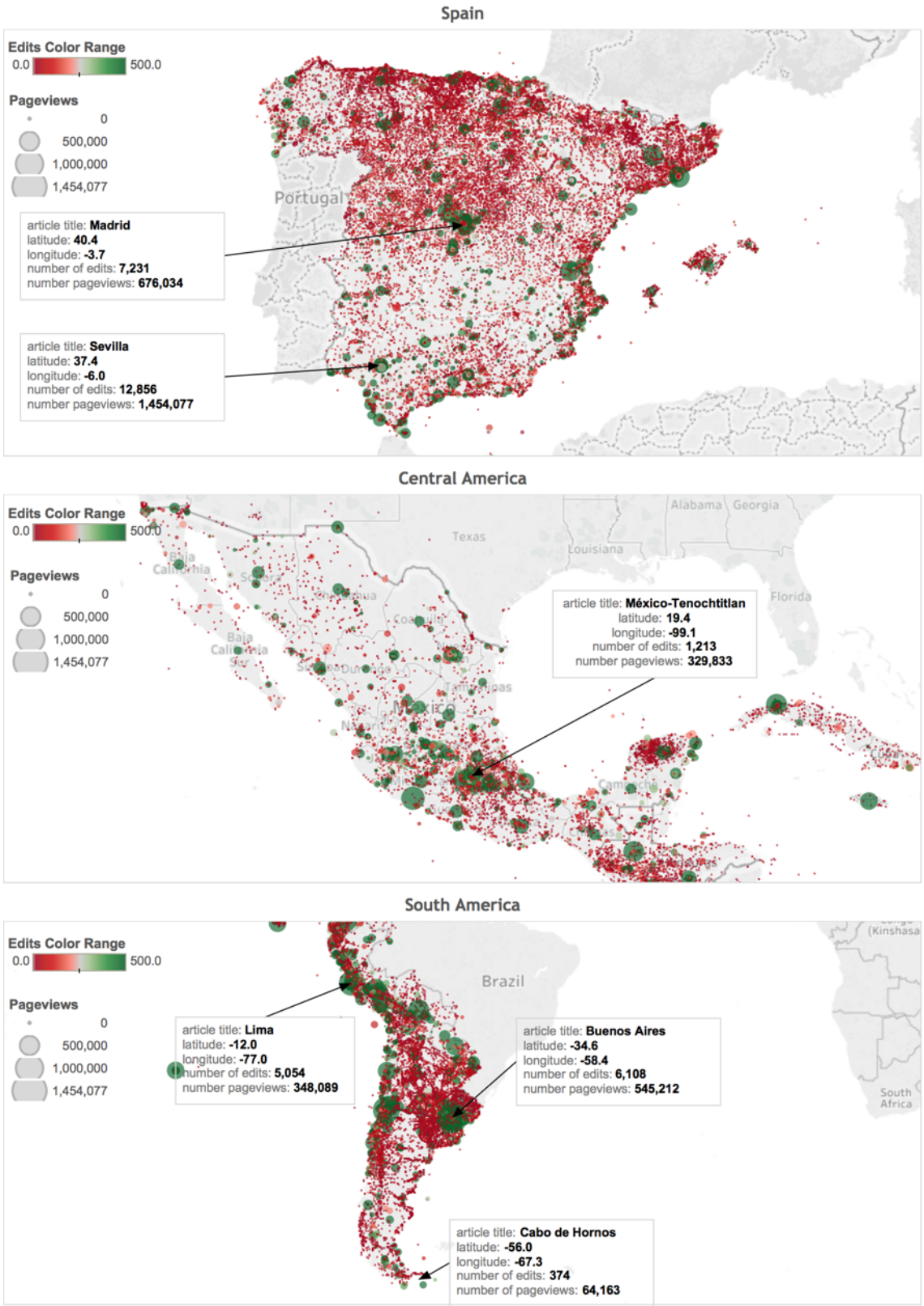


Figure 58. Editor and reader engagement in CIRA Geolocated articles from Spanish Wikipedia (zooms on Spain, Central America and South America). *Each point is a CIRA geolocated article. Colour represents the number of edits, depicted as a continuum from red to green with a middle point of 250 edits in colour beige. Size represents the number of page views. Important geolocated articles are marked with infoboxes.*

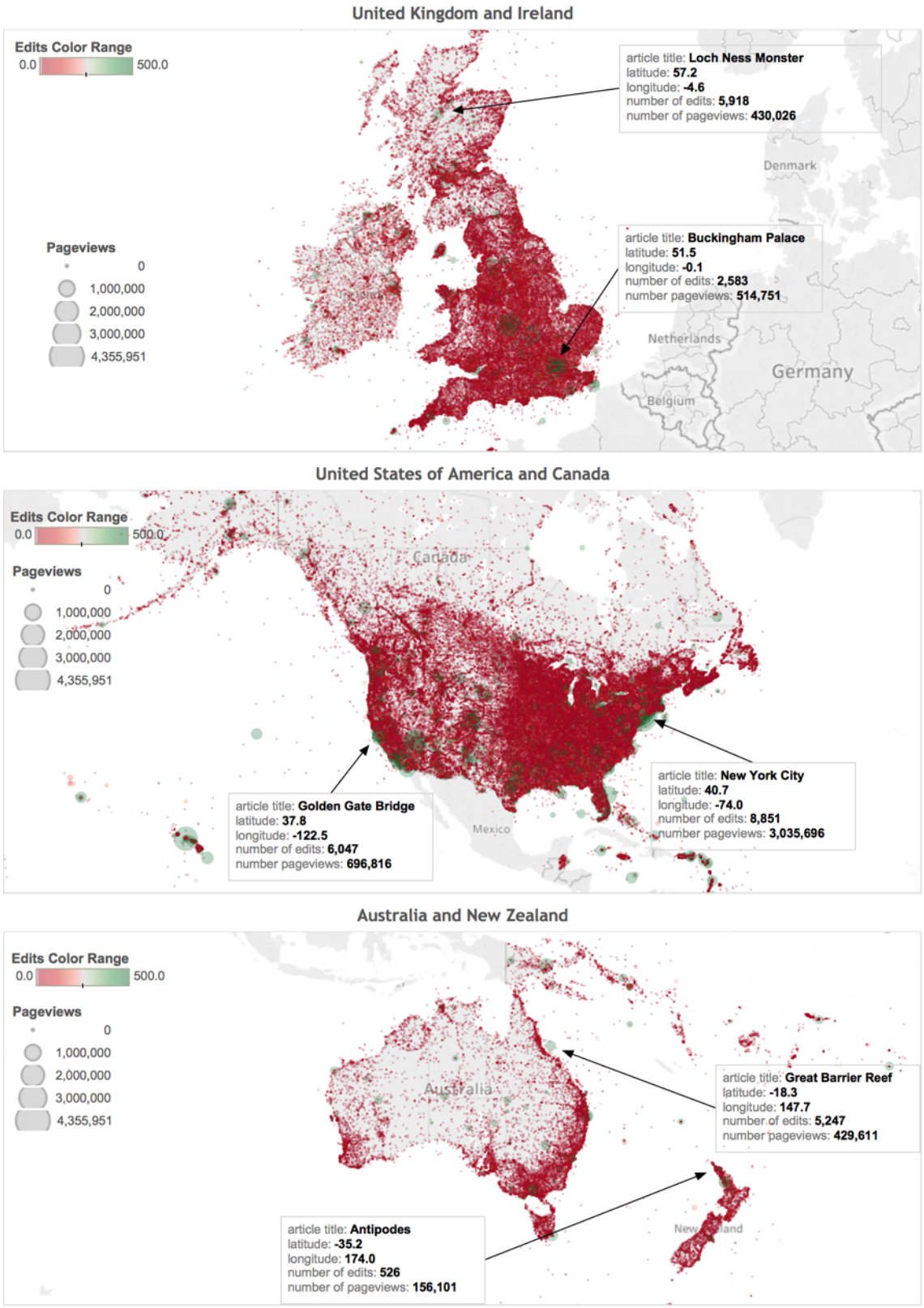


Figure 59. Editor and reader engagement in CIRA Geolocated articles from English Wikipedia (zooms on United Kingdom, United States of America and Canada, Australia and New Zealand). *Each point is a CIRA geolocated article. Colour represents the number of edits, depicted as a continuum from red to green with a middle point of 250 edits in colour beige. Size represents the number of page views. Important geolocated articles are marked with infoboxes.*

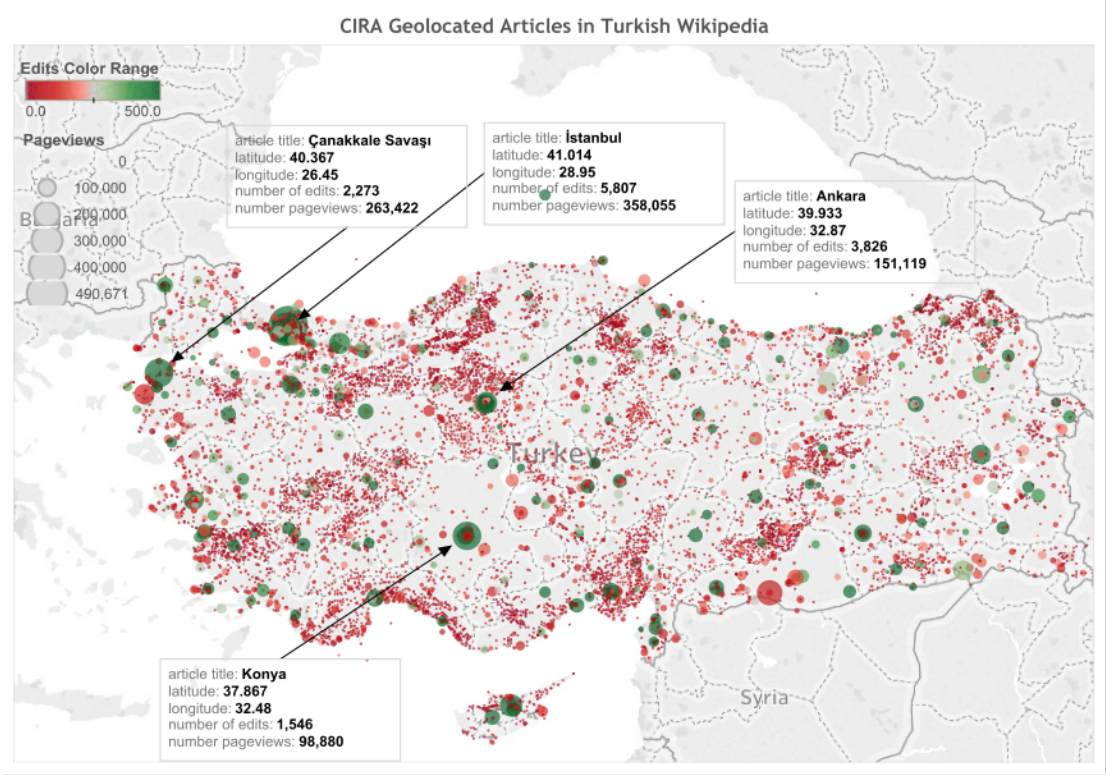


Figure 60. Editor and reader engagement in CIRA Geolocated articles from Turkish Wikipedia. *Each point is a CIRA geolocated article. Colour represents the number of edits, depicted as a continuum from red to green with a middle point of 250 edits in colour beige. Size represents the number of page views. Important geolocated articles are marked with infoboxes.*

2.4 Prioritising the Culture Gap

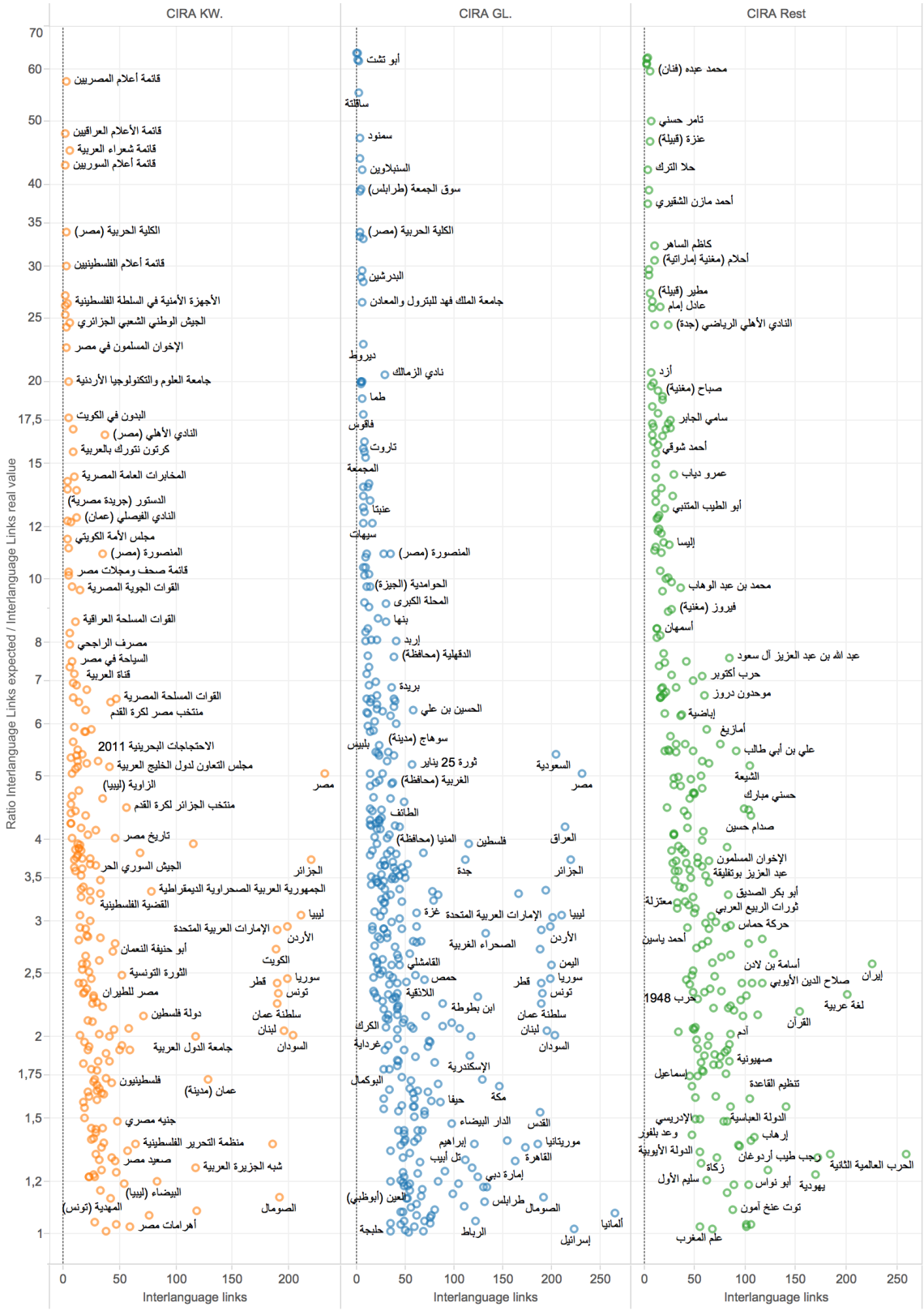


Figure 61. Arabic Wikipedia CIRA articles according to the priority ratio (Interlanguage Links expected / Interlanguage Links current value), by Interlanguage links current value and coloured by CIRA segment. *Only the top 5 articles for each Interlanguage Link are shown.*

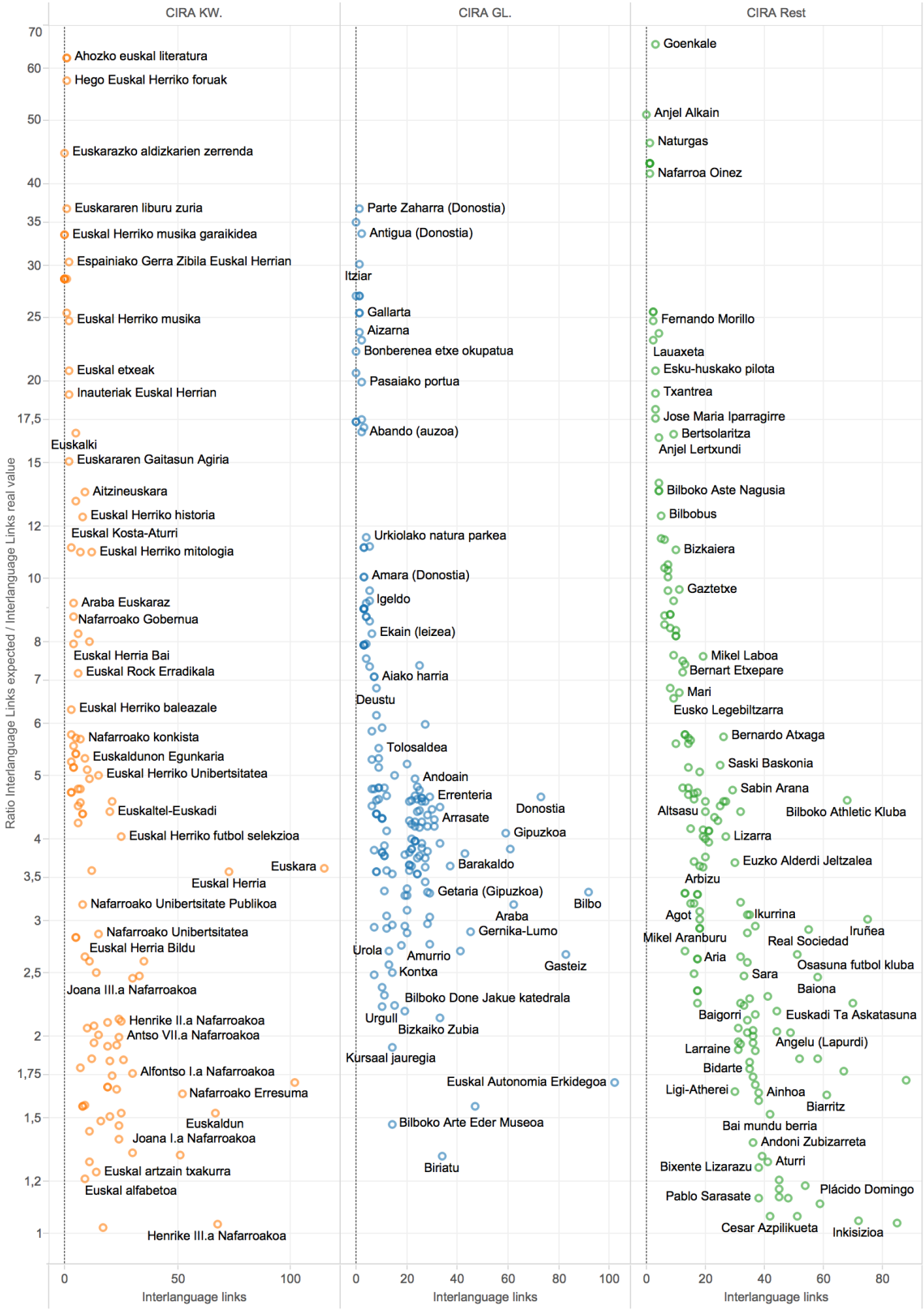


Figure 62. Basque Wikipedia CIRA articles according to the priority ratio (Interlanguage Links expected / Interlanguage Links current value), by Interlanguage links current value and coloured by CIRA segment. *Only the top 5 articles for each Interlanguage Link are shown.*

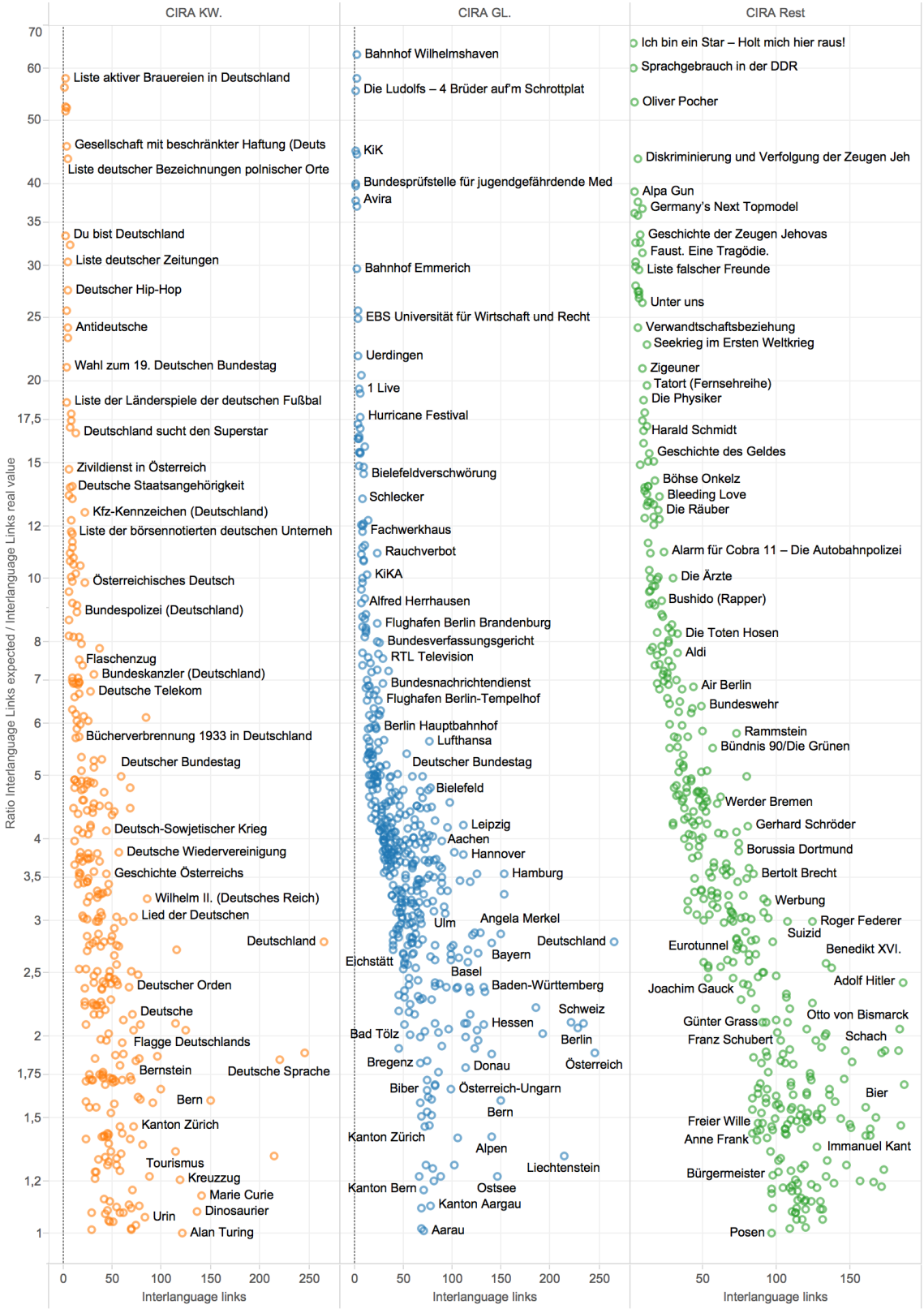


Figure 63. German Wikipedia CIRA articles according to the priority ratio (Interlanguage Links expected / Interlanguage Links current value), by Interlanguage links current value and coloured by CIRA segment. *Only the top 5 articles for each Interlanguage Link are shown.*

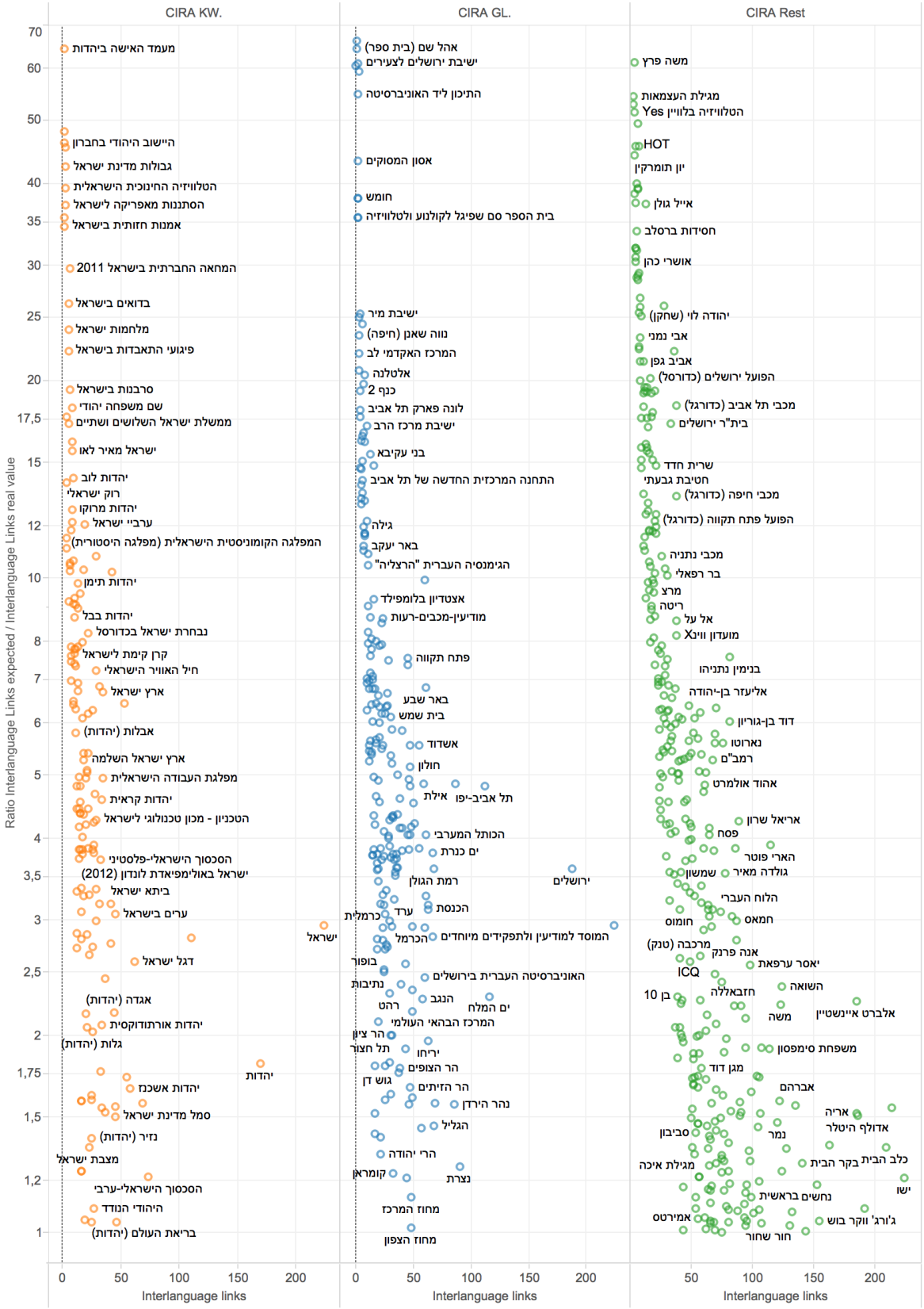


Figure 64. Hebrew Wikipedia CIRA articles according to the priority ratio (Interlanguage Links expected / Interlanguage Links current value), by Interlanguage links current value and coloured by CIRA segment. *Only the top 5 articles for each Interlanguage Link are shown.*

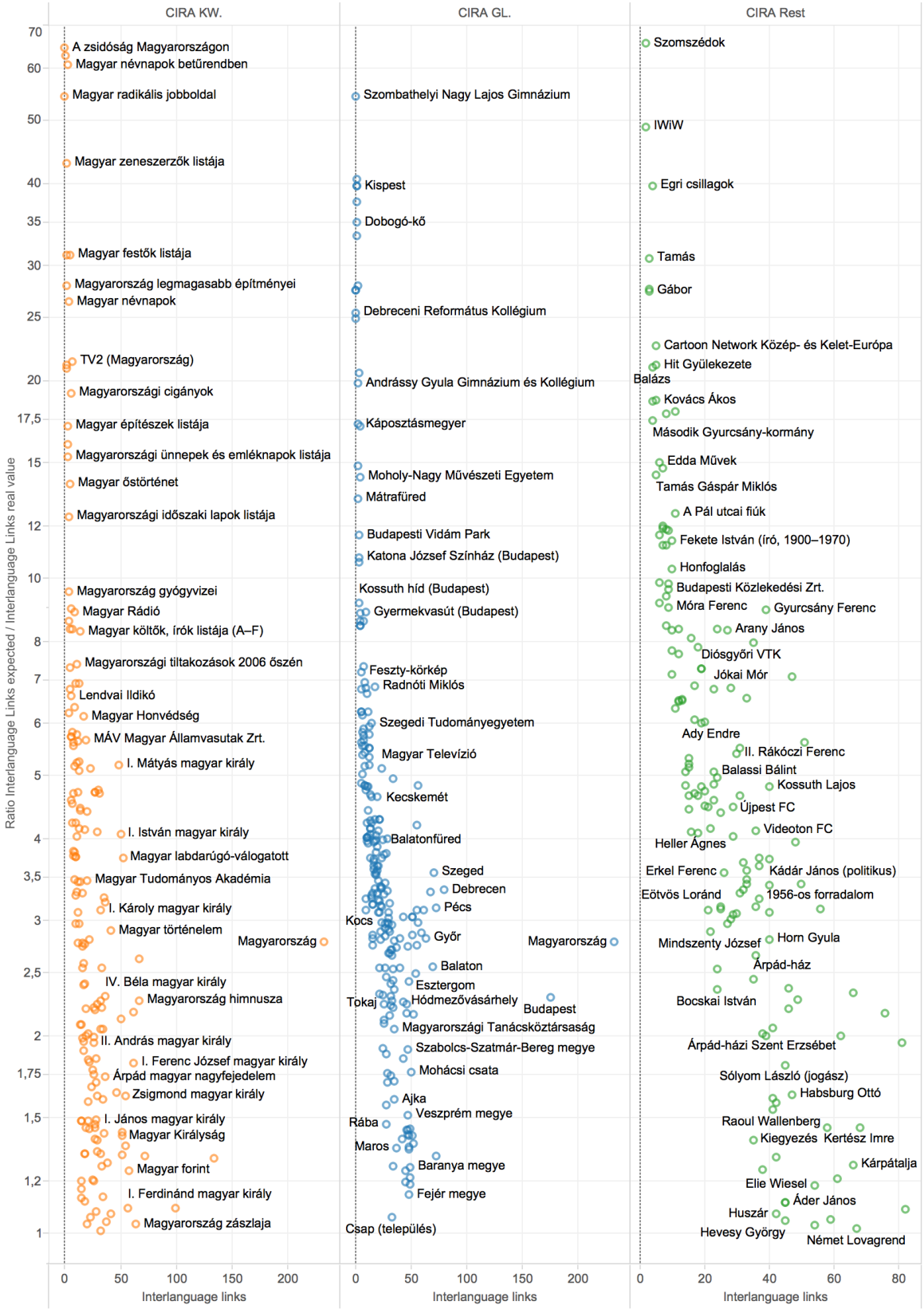


Figure 65. Hungarian Wikipedia CIRA articles according to the priority ratio (Interlanguage Links expected / Interlanguage Links current value), by Interlanguage links current value and coloured by CIRA segment. *Only the top 5 articles for each Interlanguage Link are shown.*

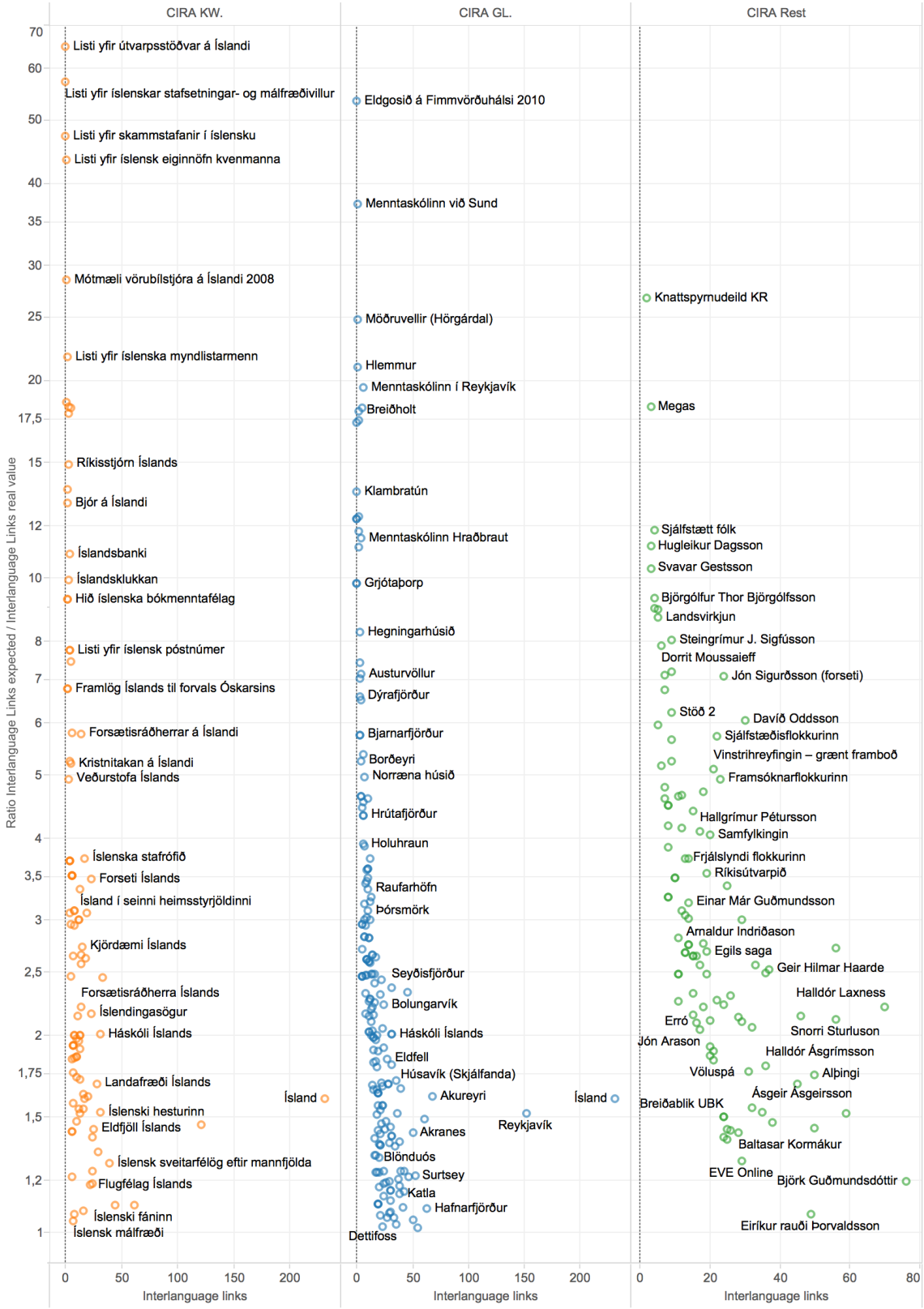


Figure 66. Icelandic Wikipedia CIRA articles according to the priority ratio (Interlanguage Links expected / Interlanguage Links current value), by Interlanguage links current value and coloured by CIRA segment. *Only the top 5 articles for each Interlanguage Link are shown.*

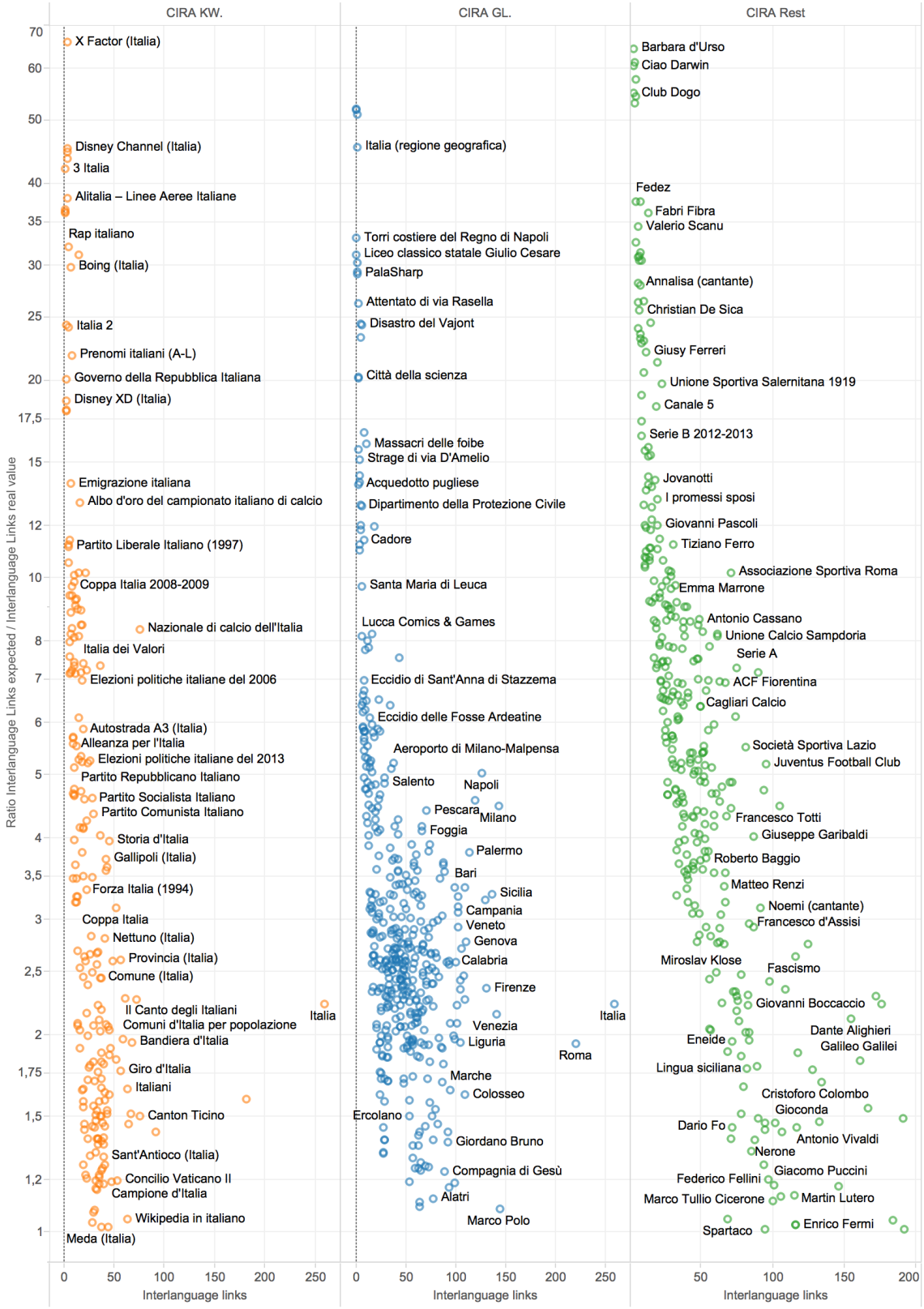


Figure 67. Italian Wikipedia CIRA articles according to the priority ratio (Interlanguage Links expected / Interlanguage Links current value), by Interlanguage links current value and coloured by CIRA segment. *Only the top 5 articles for each Interlanguage Link are shown.*

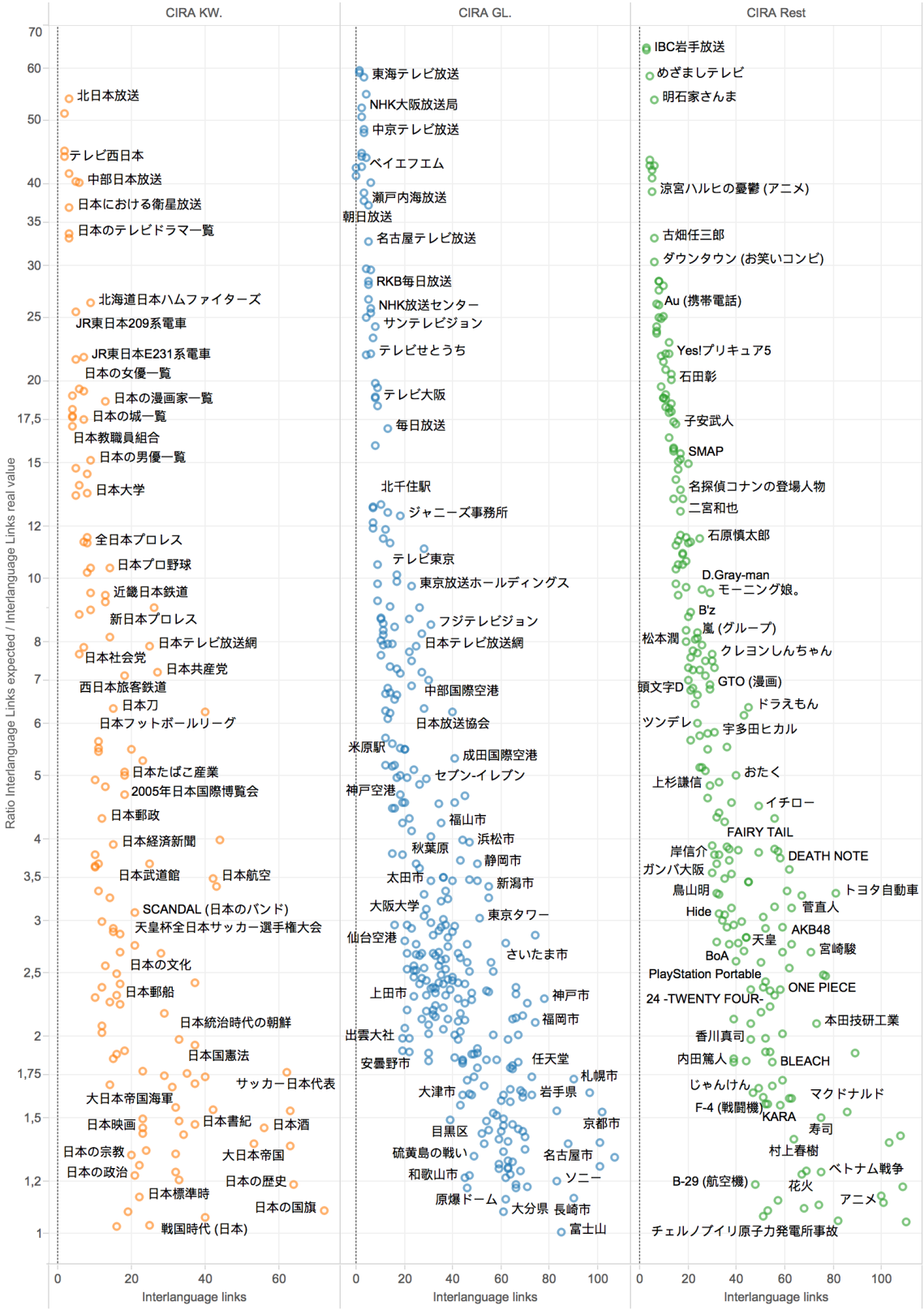


Figure 68. Japanese Wikipedia CIRA articles according to the priority ratio (Interlanguage Links expected / Interlanguage Links current value), by Interlanguage links current value and coloured by CIRA segment. *Only the top 5 articles for each Interlanguage Link are shown.*

Figure 69. Macedonian Wikipedia CIRA articles according to the priority ratio (Interlanguage Links expected / Interlanguage Links current value), by Interlanguage links current value and coloured by CIRA segment. *Only the top 5 articles for each Interlanguage Link are shown.*

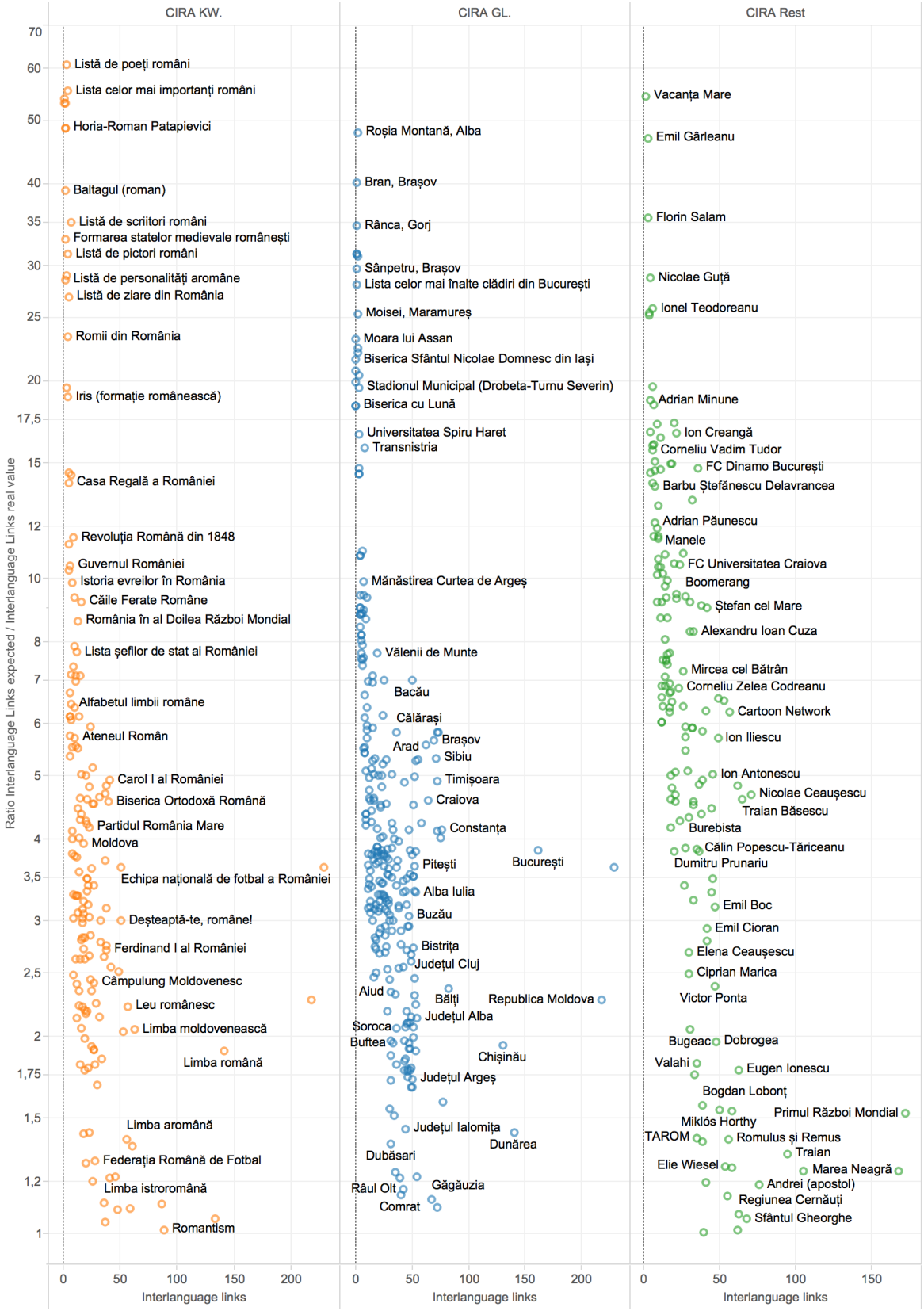


Figure 70. Romanian Wikipedia CIRA articles according to the priority ratio (Interlanguage Links expected / Interlanguage Links current value), by Interlanguage links current value and coloured by CIRA segment. *Only the top 5 articles for each Interlanguage Link are shown.*

Figure 71. Russian Wikipedia CIRA articles according to the priority ratio (Interlanguage Links expected / Interlanguage Links current value), by Interlanguage links current value and coloured by CIRA segment. *Only the top 5 articles for each Interlanguage Link are shown.*

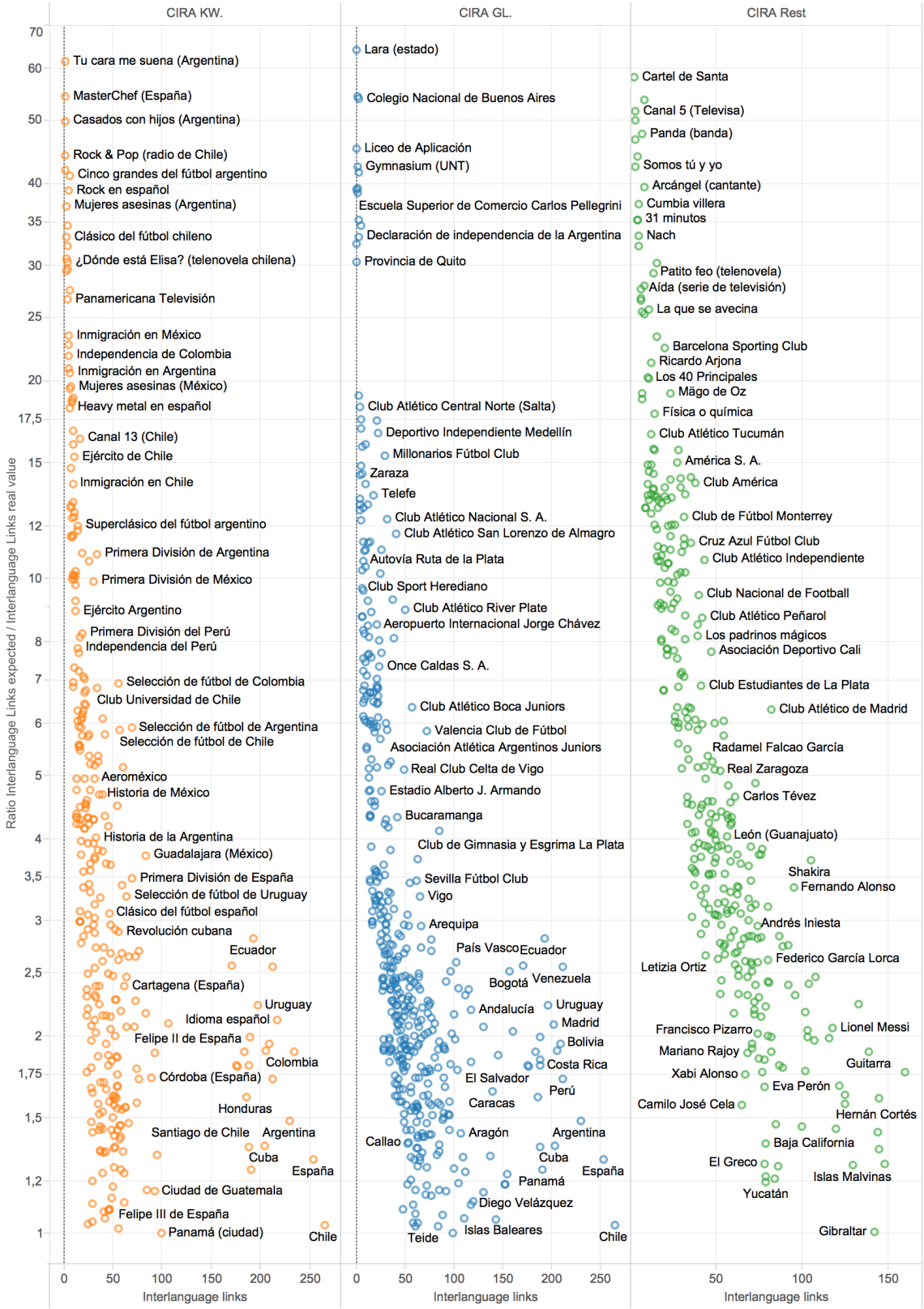


Figure 72. Spanish Wikipedia CIRA articles according to the priority ratio (Interlanguage Links expected / Interlanguage Links current value), by Interlanguage links current value and coloured by CIRA segment. *Only the top 5 articles for each Interlanguage Link are shown.*

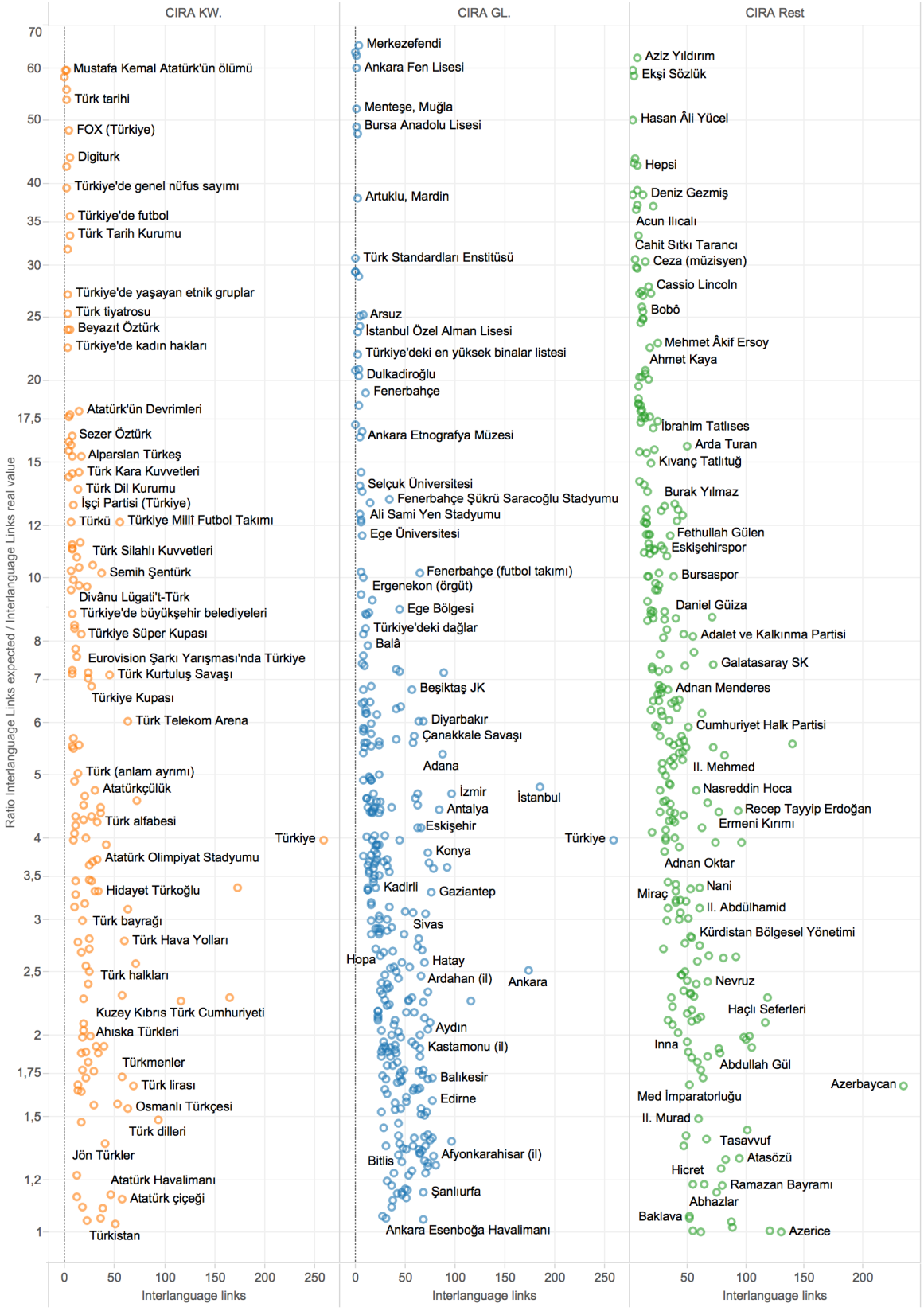


Figure 73. Turkish Wikipedia CIRA articles according to the priority ratio (Interlanguage Links expected / Interlanguage Links current value), by Interlanguage links current value and coloured by CIRA segment. *Only the top 5 articles for each Interlanguage Link are shown.*

2.5 CIRA Exported Articles

Figure 74. Top 50 Arabic Wikipedia CIRA exported articles by number of exporter editors in non-primary languages (top) and by times created in non- primary languages by exporters (bottom)

Figure 75. Top 50 Basque Wikipedia CIRA exported articles by number of exporter editors in non-primary languages (top) and by times created in non- primary languages by exporters (bottom)

Figure 76. Top 50 German Wikipedia CIRA exported articles by number of exporter editors in non-primary languages (top) and by times created in non- primary languages by exporters (bottom)

Figure 77. Top 50 Hebrew Wikipedia CIRA exported articles by number of exporter editors in non-primary languages (top) and by times created in non- primary languages by exporters (bottom)

Figure 78. Top 50 Hungarian Wikipedia CIRA exported articles by number of exporter editors in non-primary languages (top) and by times created in non- primary languages by exporters (bottom)

Figure 79. Top 50 Icelandic Wikipedia CIRA exported articles by number of exporter editors in non-primary languages (top) and by times created in non- primary languages by exporters (bottom)

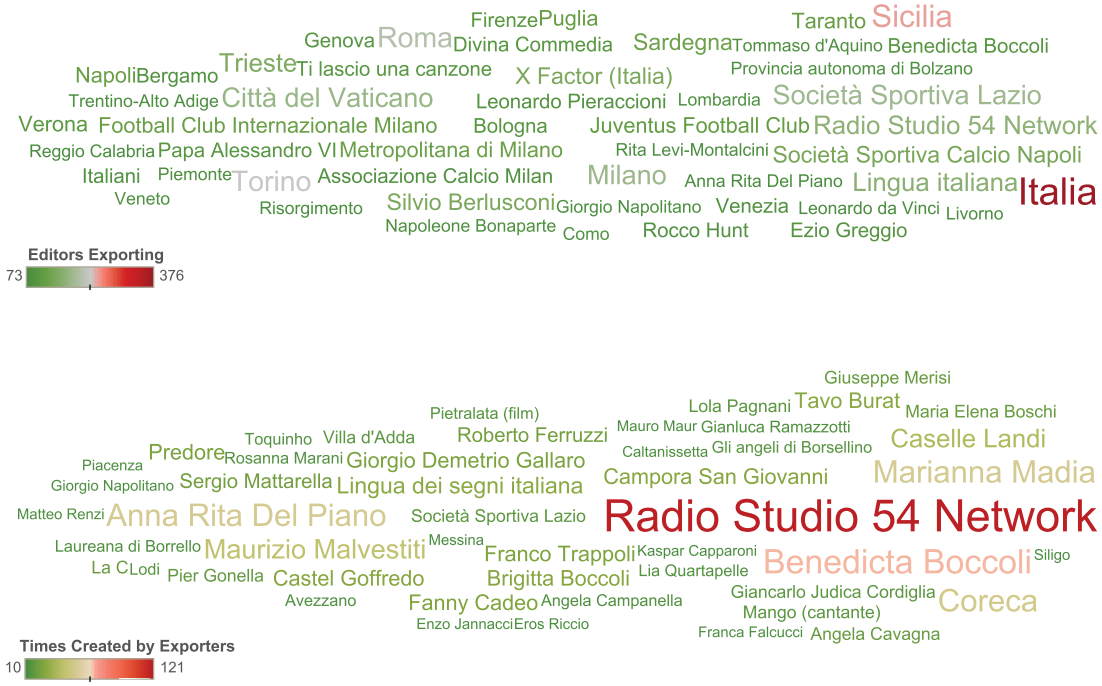


Figure 80. Top 50 Italian Wikipedia CIRA exported articles by number of exporter editors in non-primary languages (top) and by times created in non- primary languages by exporters (bottom)

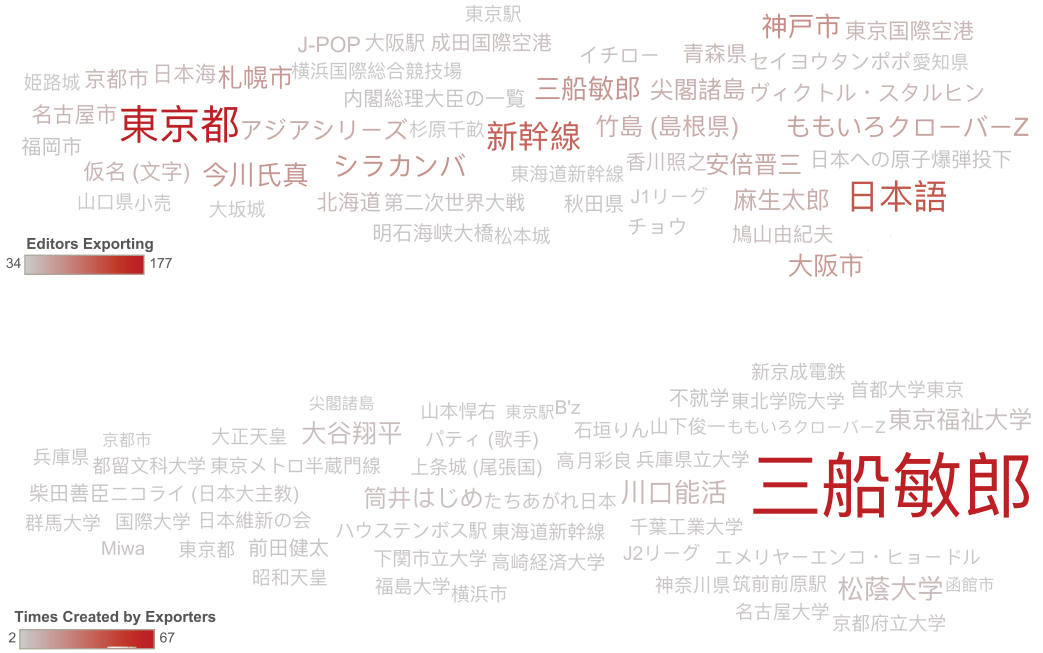


Figure 81. Top 50 Japanese Wikipedia CIRA exported articles by number of exporter editors in non-primary languages (top) and by times created in non- primary languages by exporters (bottom)

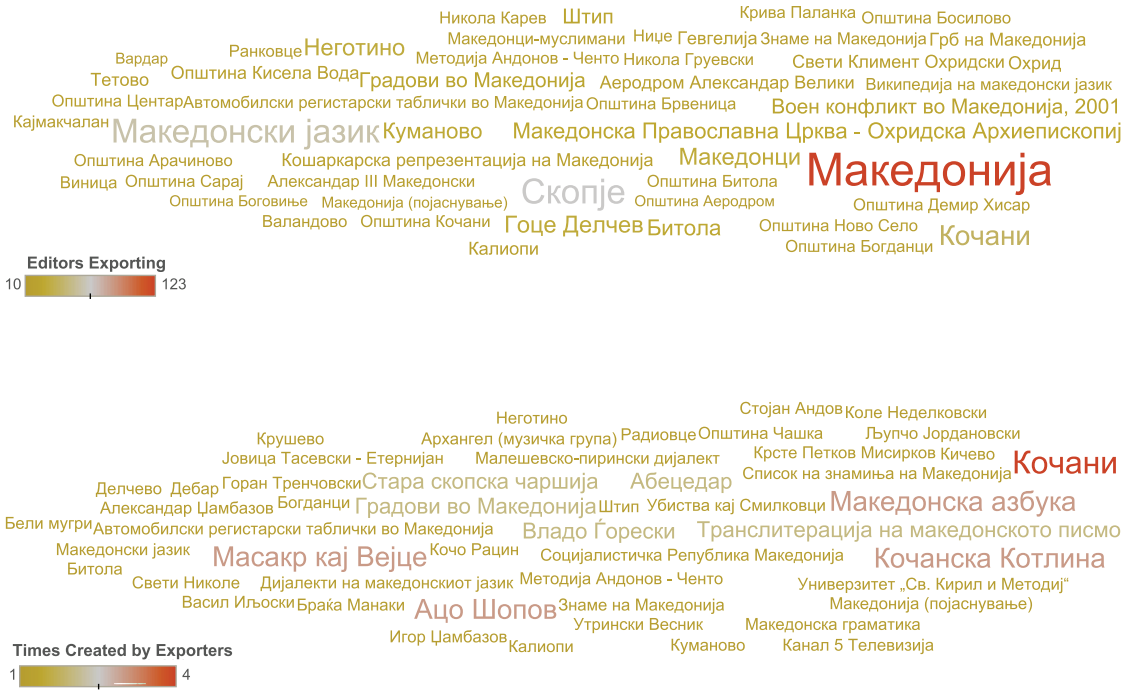


Figure 82. Top 50 Macedonian Wikipedia CIRA exported articles by number of exporter editors in non-primary languages (top) and by times created in non- primary languages by exporters (bottom)

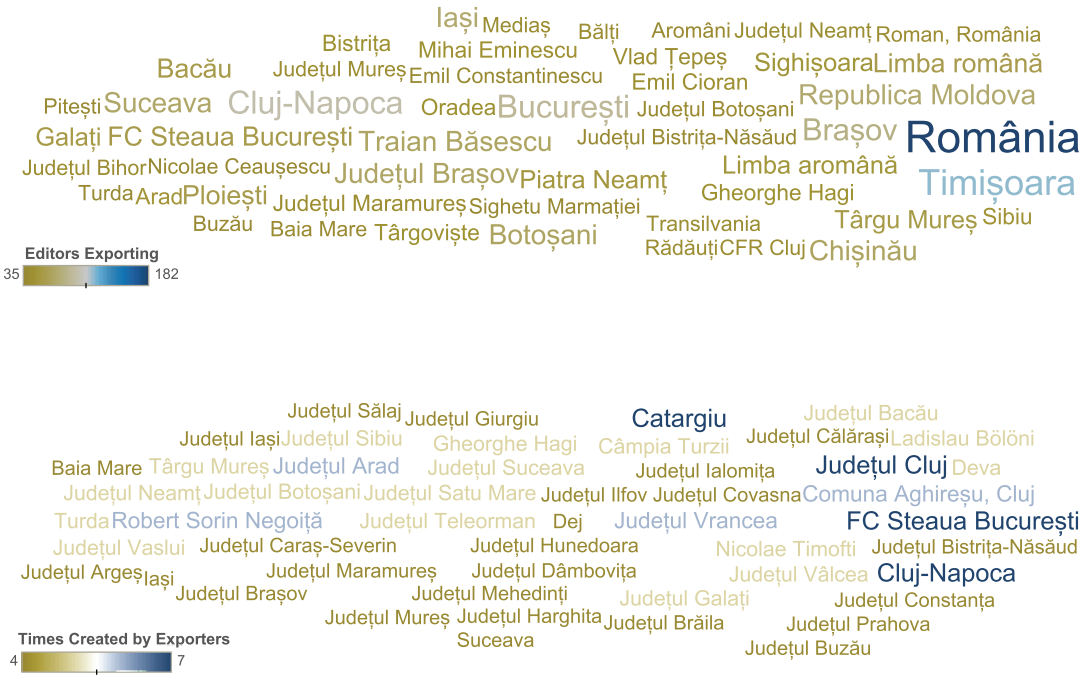


Figure 83. Top 50 Romania Wikipedia CIRA exported articles by number of exporter editors in non-primary languages (top) and by times created in non- primary languages by exporters (bottom)

Figure 84. Top 50 Russian Wikipedia CIRA exported articles by number of exporter editors in non-primary languages (top) and by times created in non- primary languages by exporters (bottom)

Figure 85. Top 50 Spanish Wikipedia CIRA exported articles by number of exporter editors in non-primary languages (top) and by times created in non- primary languages by exporters (bottom)

Figure 86. Top 50 Turkish Wikipedia CIRA exported articles by number of exporter editors in non-primary languages (top) and by times created in non- primary languages by exporters (bottom)

Appendix 3. Statistical Tests Results

The following tables report the results for the pairwise comparisons (Dunn's test with the Bonferroni correction) for the Kruskal-Wallis tests employed in chapters 6 and 8.

3.1 Editor Session Characteristics

Table 28. Editor session characteristics (edits, Bytes. session and inter-session time) by edit buckets and editor types. *Values correspond to Kruskal-Wallis test mean ranks for each group and p-values for the pairwise comparison among groups according to Dunn's test with Bonferroni correction.*

Session Characteristics	**Edits**			**Bytes**			**Duration**			**Absence Time**		
pairwise comparisons: edit buckets	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
(1-100) - (101-1000)	0.000	279538	300330	0.589	308942	310833	0.000	283889	307692	0.000	381130	494190
(1-100) - (1001-5000)	0.000	279538	337189	0.000	308942	356048	0.000	283889	307692	0.000	381130	440476
(1-100) - (5001-10000)	0.000	279538	334769	0.000	308942	355642	0.000	283889	346014	0.027	381130	384254
(1-100) - (10001+)	0.000	279538	430953	0.000	308942	412660	0.000	283889	425793	0.000	381130	328839
(101-1000) - (1001-5000)	0.000	300330	337189	0.000	310833	356048	0.000	307692	307692	0.000	494190	440476
(101-1000) - (5001-10000)	0.000	300330	334769	0.000	310833	355642	0.000	307692	346014	0.000	494190	384254
(101-1000) - (10001+)	0.000	300330	430953	0.000	310833	412660	0.000	307692	425793	0.000	494190	328839
(1001-5000) - (5001-10000)	0.211	337189	334769	1.000	356048	355642	0.000	338497	346014	0.000	440476	384254
(1001-5000) - (10001+)	0.000	337189	430953	0.000	356048	412660	0.000	338497	425793	0.000	440476	328839
(5001-10000) - (10001+)	0.000	334769	430953	0.000	355642	412660	0.000	346014	425793	0.000	384254	328839
Session Characteristics	**Edits**			**Bytes**			**Duration**			**Absence Time**		
pairwise comparisons: editor types	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
Anonymous - Registered	0.000	430015	565833	0.000	419103	590717	0.000	423572	574565	0.000	407725	407725
Anonymous - Security Force	0.000	430015	695251	0.000	419103	704263	0.000	423572	700154	0.000	407725	407725
Anonymous - Production Force	0.000	430015	741127	0.000	419103	749921	0.000	423572	743078	0.000	407725	407725
Anonymous - Quality Patrol	0.000	430015	743760	0.000	419103	722324	0.000	423572	737007	0.000	407725	407725
Anonymous - Administrators	0.000	430015	821405	0.000	419103	747026	0.000	423572	805855	0.000	407725	407725
Registered - Security Force	0.000	565833	695251	0.000	590717	704263	0.000	574565	700154	0.000	732586	732586
Registered - Production Force	0.000	565833	741127	0.000	590717	749921	0.000	574565	743078	0.000	732586	732586
Registered - Quality Patrol	0.000	565833	743760	0.000	590717	722324	0.000	574565	737007	0.000	732586	732586
Registered - Administrators	0.000	565833	821405	0.000	590717	747026	0.000	574565	805855	0.000	732586	732586
Security Force - Production Force	0.000	695251	741127	0.000	704263	749921	0.000	700154	743078	0.000	624802	624802
Security Force - Quality Patrol	0.000	695251	743760	0.002	704263	722324	0.000	700154	737007	0.002	624802	624802
Security Force - Administrators	0.000	695251	821405	0.000	704263	747026	0.000	700154	805855	0.000	624802	624802
Production Force - Quality Patrol	1.000	741127	743760	0.000	749921	722324	0.001	743078	737007	0.000	650775	650775
Production Force - Administrators	0.000	741127	821405	0.413	749921	747026	0.000	743078	805855	0.413	650775	650775
Quality Patrol - Administrators	0.000	743760	821405	0.000	722324	747026	0.000	737007	805855	0.000	627681	627681

3.2 Multilingualism and Primary Language Edits

Table 29. Editor proportion of edits in primary language in relation to all edits by edit bucket. *Values correspond to Kruskal-Wallis test mean ranks for each group and p-values for the pairwise comparison among groups according to Dunn's test with Bonferroni correction.*

Primary Language Edits %	**Arabic**			**Basque**			**Catalan**			**English**			**German**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
(1-100) - (101-1000)	0.000	3440	5798	0.000	273	408	0.000	2297	3643	0.000	59407	99119	0.000	16088	28995
(1-100) - (1001-5000)	0.000	3440	6491	0.000	273	505	0.000	2297	4278	0.000	59407	112111	0.000	16088	32959
(1-100) - (5001-10000)	0.000	3440	6338	0.000	273	559	0.000	2297	4150	0.000	59407	118245	0.000	16088	34410
(1-100) - (10001+)	0.000	3440	6698	0.000	273	545	0.000	2297	4356	0.000	59407	120649	0.000	16088	35554
(101-1000) - (1001-5000)	0.000	5798	6491	0.279	408	505	0.000	3643	4278	0.000	99119	112111	0.000	28995	32959
(101-1000) - (5001-10000)	0.344	5798	6338	0.240	408	559	0.212	3643	4150	0.000	99119	118245	0.000	28995	34410
(101-1000) - (10001+)	0.000	5798	6698	0.031	408	545	0.000	3643	4356	0.000	99119	120649	0.000	28995	35554
(1001-5000) - (5001-10000)	1.000	6491	6338	1.000	505	559	1.000	4278	4150	0.001	112111	118245	0.009	32959	34410
(1001-5000) - (10001+)	1.000	6491	6698	1.000	505	545	1.000	4278	4356	0.000	112111	120649	0.000	32959	35554
(5001-10000) - (10001+)	1.000	6338	6698	1.000	559	545	1.000	4150	4356	1.000	118245	120649	0.198	34410	35554
Primary Language Edits %	**Hebrew**			**Hungarian**			**Icelandic**			**Italian**			**Japanese**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
(1-100) - (101-1000)	0.000	1888	3331	0.000	1548	2650	0.000	162	283	0.000	5683	10226	0.000	2881	5484
(1-100) - (1001-5000)	0.000	1888	3910	0.000	1548	3038	0.000	162	266	0.000	5683	11372	0.000	2881	6399
(1-100) - (5001-10000)	0.000	1888	3965	0.000	1548	3181	1.000	162	174	0.000	5683	11888	0.000	2881	6323
(1-100) - (10001+)	0.000	1888	4260	0.000	1548	3383	0.020	162	259	0.000	5683	12570	0.000	2881	6553
(101-1000) - (1001-5000)	0.000	3331	3910	0.000	2650	3038	1.000	283	266	0.000	10226	11372	0.000	5484	6399
(101-1000) - (5001-10000)	0.000	3331	3965	0.000	2650	3181	1.000	283	174	0.000	10226	11888	0.000	5484	6323
(101-1000) - (10001+)	0.000	3331	4260	0.000	2650	3383	1.000	283	259	0.000	10226	12570	0.000	5484	6553
(1001-5000) - (5001-10000)	1.000	3910	3965	1.000	3038	3181	1.000	266	174	0.599	11372	11888	1.000	6399	6323
(1001-5000) - (10001+)	0.049	3910	4260	0.021	3038	3383	1.000	266	259	0.000	11372	12570	1.000	6399	6553
(5001-10000) - (10001+)	0.860	3965	4260	1.000	3181	3383	1.000	174	259	0.229	11888	12570	1.000	6323	6553
Primary Language Edits %	**Macedonian**			**Romanian**			**Russian**			**Spanish**			**Turkish**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
(1-100) - (101-1000)	0.000	263	412	0.000	1708	2781	0.000	162386	106808	0.000	15983	27165	0.000	3099	5233
(1-100) - (1001-5000)	0.000	263	468	0.000	1708	3063	0.000	162386	64034	0.000	15983	30554	0.000	3099	5752
(1-100) - (5001-10000)	0.005	263	485	0.000	1708	2833	0.000	162386	44238	0.000	15983	31560	0.000	3099	6206
(1-100) - (10001+)	0.000	263	494	0.000	1708	3248	0.000	162386	40294	0.000	15983	32417	0.000	3099	6149
(101-1000) - (1001-5000)	1.000	412	468	0.104	2781	3063	0.000	106808	64034	0.000	27165	30554	0.004	5233	5752
(101-1000) - (5001-10000)	1.000	412	485	1.000	2781	2833	0.000	106808	44238	0.000	27165	31560	0.000	5233	6206
(101-1000) - (10001+)	1.000	412	494	0.013	2781	3248	0.000	106808	40294	0.000	27165	32417	0.002	5233	6149
(1001-5000) - (5001-10000)	1.000	468	485	1.000	3063	2833	0.000	64034	44238	1.000	30554	31560	0.846	5752	6206
(1001-5000) - (10001+)	1.000	468	494	1.000	3063	3248	0.000	64034	40294	0.002	30554	32417	1.000	5752	6149
(5001-10000) - (10001+)	1.000	485	494	0.976	2833	3248	0.750	44238	40294	1.000	31560	32417	1.000	6206	6149

Table 30. Editor proportion of edits in primary language in relation to all edits by editor type. *Values correspond to Kruskal-Wallis test mean ranks for each group and p-values for the pairwise comparison among groups according to Dunn's test with Bonferroni correction.*

Primary Language Edits %	**Arabic**			**Basque**			**Catalan**			**English**			**German**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
Registered - Security Force	0.509	3673	5652	-	-	-	1.000	2527	2624	0.000	16268	38096	0.000	65441	129603
Registered - Production Force	0.000	3673	6264	-	-	-	0.000	2527	4357	0.000	16268	30445	0.006	65441	125331
Registered - Quality Patrol	0.000	3673	6852	-	-	-	0.000	2527	4890		16268	0	0.000	65441	116927
Registered - Administrators	0.000	3673	7082	0.008	305.82	473.63	0.000	2527	4344	0.000	16268	35245	0.000	65441	126823
Security Force - Production Force	1.000	5652	6264	-	-	-	1.000	2624	4357	0.285	38096	30445	1.000	129603	125331
Security Force - Quality Patrol	1.000	5652	6852	-	-	-	1.000	2624	4890		38096	0	1.000	129603	116927
Security Force - Administrators	1.000	5652	7082	-	-	-	0.604	2624	4344	1.000	38096	35245	1.000	129603	126823
Production Force - Quality Patrol	0.124	6264	6852	-	-	-	1.000	4357	4890		30445	0	1.000	125331	116927
Production Force - Administrators	0.619	6264	7082	-	-	-	1.000	4357	4344	0.000	30445	35245	1.000	125331	126823
Quality Patrol - Administrators	1.000	6852	7082	-	-	-	1.000	4890	4344			35245	1.000	116927	126823
Primary Language Edits %	**Hebrew**			**Hungarian**			**Icelandic**			**Italian**			**Japanese**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
Registered - Security Force	0.055	2240	4086	0.069	1763	3131	-	-	-	0.005	7054	13669	0.010	4077	7253
Registered - Production Force	0.000	2240	3895	0.000	1763	3092	-	-	-	0.000	7054	12149			
Registered - Quality Patrol	0.000	2240	4359	-	-	-	-	-	-	0.000	7054	12977	1.000	4077	5085
Registered - Administrators	0.000	2240	4274	0.000	1763	3328	0.001	181.10	265.76	0.000	7054	12641	0.000	4077	6507
Security Force - Production Force	1.000	4086	3895	1.000	3131	3092	-	-	-	1.000	13669	12149			
Security Force - Quality Patrol	0.149	4086	4359	-	-	-	-	-	-	1.000	13669	12977	1.000	7253	5085
Security Force - Administrators	1.000	4086	4274	1.000	3131	3328	-	-	-	1.000	13669	12641	1.000	7253	6507
Production Force - Quality Patrol	1.000	3895	4359	-	-	-	-	-	-	1.000	12149	12977			
Production Force - Administrators	1.000	3895	4274	1.000	3092	3328	-	-	-	1.000	12149	12641			
Quality Patrol - Administrators	1.000	4359	4274	-	-	-	-	-	-	1.000	12977	12641	1.000	5085	6507
Primary Language Edits %	**Macedonian**			**Romanian**			**Russian**			**Spanish**			**Turkish**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
Registered - Security Force	-	-	-	-	-	-	0.000	12354	23714	0.000	18448	34693	0.013	3332	6601
Registered - Production Force	0.000	276	474	0.000	1877.65	3133.13	0.000	12354	20502	0.000	18448	31451	0.000	3332	5421
Registered - Quality Patrol	0.000	276	483	-	-	-	0.000	12354	21761	0.000	18448	32985	0.000	3332	5889
Registered - Administrators	0.000	276	429	0.000	1877.65	3197.65	0.000	12354	22276	0.000	18448	32896	0.000	3332	6106
Security Force - Production Force	-	-	-	-	-	-	1.000	23714	20502	1.000	34693	31451	1.000	6601	5421
Security Force - Quality Patrol	-	-	-	-	-	-	1.000	23714	21761	1.000	34693	32985	1.000	6601	5889
Security Force - Administrators	-	-	-	-	-	-	1.000	23714	22276	1.000	34693	32896	1.000	6601	6106
Production Force - Quality Patrol	1.000	474	483	-	-	-	0.000	20502	21761	0.798	31451	32985	0.082	5421	5889
Production Force - Administrators	1.000	474	429	1.000	3133.13	3197.65	0.752	20502	22276	1.000	31451	32896	1.000	5421	6106
Quality Patrol - Administrators	1.000	483	429	-	-	-	1.000	21761	22276	1.000	32985	32896	1.000	5889	6106

3.3 Community Oriented Activities

Table 31. Editor proportion of edits in Data Spaces by edit bucket. *Values correspond to Kruskal-Wallis test mean ranks for each group and p-values for the pairwise comparison among groups according to Dunn's test with Bonferroni correction.*

Data Spaces	**Arabic**			**Basque**			**Catalan**			**English**			**German**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
(1-100) - (101-1000)	0.000	78075	125191	0.000	1439	2253	0.000	17487	23887	0.000	3405808	5045774	0.000	331455	474550
(1-100) - (1001-5000)	0.000	78075	149190	0.000	1439	2698	0.000	17487	32902	0.000	3405808	6315309	0.000	331455	612883
(1-100) - (5001-10000)	0.000	78075	153707	0.000	1439	2747	0.000	17487	33904	0.000	3405808	6605100	0.000	331455	643937
(1-100) - (10001+)	0.000	78075	154216	0.000	1439	2911	0.000	17487	34858	0.000	3405808	6670068	0.000	331455	651429
(101-1000) - (1001-5000)	0.000	125191	149190	0.000	2253	2698	0.000	23887	32902	0.000	5045774	6315309	0.000	474550	612883
(101-1000) - (5001-10000)	0.000	125191	153707	0.003	2253	2747	0.000	23887	33904	0.000	5045774	6605100	0.000	474550	643937
(101-1000) - (10001+)	0.000	125191	154216	0.000	2253	2911	0.000	23887	34858	0.000	5045774	6670068	0.000	474550	651429
(1001-5000) - (5001-10000)	0.694	149190	153707	1.000	2698	2747	0.694	32902	33904	0.000	6315309	6605100	0.000	612883	643937
(1001-5000) - (10001+)	0.000	149190	154216	0.614	2698	2911	0.000	32902	34858	0.000	6315309	6670068	0.000	612883	651429
(5001-10000) - (10001+)	1.000	153707	154216	1.000	2747	2911	1.000	33904	34858	0.000	6605100	6670068	0.107	643937	651429
Data Spaces	**Hebrew**			**Hungarian**			**Icelandic**			**Italian**			**Japanese**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
(1-100) - (101-1000)	0.000	35597	54677	0.000	32183	49477	0.000	1481	2240	0.000	95809	138783	0.000	142119	201706
(1-100) - (1001-5000)	0.000	35597	67012	0.000	32183	59951	0.000	1481	2923	0.000	95809	176468	0.000	142119	263428
(1-100) - (5001-10000)	0.000	35597	68562	0.000	32183	62037	0.000	1481	2926	0.000	95809	186287	0.000	142119	277841
(1-100) - (10001+)	0.000	35597	68907	0.000	32183	62732	0.000	1481	2916	0.000	95809	188440	0.000	142119	282086
(101-1000) - (1001-5000)	0.000	54677	67012	0.000	49477	59951	0.000	2240	2923	0.000	138783	176468	0.000	201706	263428
(101-1000) - (5001-10000)	0.000	54677	68562	0.000	49477	62037	0.000	2240	2926	0.000	138783	186287	0.000	201706	277841
(101-1000) - (10001+)	0.000	54677	68907	0.000	49477	62732	0.001	2240	2916	0.000	138783	188440	0.000	201706	282086
(1001-5000) - (5001-10000)	1.000	67012	68562	0.201	59951	62037	1.000	2923	2926	0.000	176468	186287	0.000	263428	277841
(1001-5000) - (10001+)	0.254	67012	68907	0.003	59951	62732	1.000	2923	2916	0.000	176468	188440	0.000	263428	282086
(5001-10000) - (10001+)	1.000	68562	68907	1.000	62037	62732	1.000	2926	2916	1.000	186287	188440	0.638	277841	282086
Data Spaces	**Macedonian**			**Romanian**			**Russian**			**Spanish**			**Turkish**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
(1-100) - (101-1000)	0.000	2273	3707	0.000	21121	33355	0.000	153402	234743	0.000	326096	405020	0.000	69977	113984
(1-100) - (1001-5000)	0.000	2273	4363	0.000	21121	40632	0.000	153402	293735	0.000	326096	560807	0.000	69977	134534
(1-100) - (5001-10000)	0.000	2273	4368	0.000	21121	41478	0.000	153402	302582	0.000	326096	623098	0.000	69977	136189
(1-100) - (10001+)	0.000	2273	4450	0.000	21121	41460	0.000	153402	303984	0.000	326096	642982	0.000	69977	137615
(101-1000) - (1001-5000)	0.000	3707	4363	0.000	33355	40632	0.000	234743	293735	0.000	405020	560807	0.000	113984	134534
(101-1000) - (5001-10000)	0.086	3707	4368	0.000	33355	41478	0.000	234743	302582	0.000	405020	623098	0.000	113984	136189
(101-1000) - (10001+)	0.002	3707	4450	0.000	33355	41460	0.000	234743	303984	0.000	405020	642982	0.000	113984	137615
(1001-5000) - (5001-10000)	1.000	4363	4368	1.000	40632	41478	0.000	293735	302582	0.000	560807	623098	1.000	134534	136189
(1001-5000) - (10001+)	1.000	4363	4450	1.000	40632	41460	0.000	293735	303984	0.000	560807	642982	0.504	134534	137615
(5001-10000) - (10001+)	1.000	4368	4450	1.000	41478	41460	1.000	302582	303984	0.000	623098	642982	1.000	136189	137615

Table 32. Editor proportion of edits in Community Communication by edit bucket. *Values correspond to Kruskal-Wallis test mean ranks for each group and p-values for the pairwise comparison among groups according to Dunn's test with Bonferroni correction.*

Community Communication	**Arabic**			**Basque**			**Catalan**			**English**			**German**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
(1-100) - (101-1000)	0.000	78239	118348	0.000	1444	2240	0.000	17462	25541	0.000	3421905	4662045	0.000	335887	425435
(1-100) - (1001-5000)	0.000	78239	133504	0.000	1444	2496	0.000	17462	29963	0.000	3421905	4907233	0.000	335887	440801
(1-100) - (5001-10000)	0.000	78239	136260	0.000	1444	2520	0.000	17462	30236	0.000	3421905	5031618	0.000	335887	450251
(1-100) - (10001+)	0.000	78239	136173	0.000	1444	2656	0.000	17462	30747	0.000	3421905	5112037	0.000	335887	457736
(101-1000) - (1001-5000)	0.000	118348	133504	0.123	2240	2496	0.000	25541	29963	0.000	4662045	4907233	0.000	425435	440801
(101-1000) - (5001-10000)	0.000	118348	136260	1.000	2240	2520	0.000	25541	30236	0.000	4662045	5031618	0.000	425435	450251
(101-1000) - (10001+)	0.000	118348	136173	0.014	2240	2656	0.000	25541	30747	0.000	4662045	5112037	0.000	425435	457736
(1001-5000) - (5001-10000)	1.000	133504	136260	1.000	2496	2520	1.000	29963	30236	0.000	4907233	5031618	0.856	440801	450251
(1001-5000) - (10001+)	1.000	133504	136173	1.000	2496	2656	1.000	29963	30747	0.000	4907233	5112037	0.006	440801	457736
(5001-10000) - (10001+)	1.000	136260	136173	1.000	2520	2656	1.000	30236	30747	0.147	5031618	5112037	1.000	450251	457736
Community Communication	**Hebrew**			**Hungarian**			**Icelandic**			**Italian**			**Japanese**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
(1-100) - (101-1000)	0.000	35700	54955	0.000	32345	47051	0.000	1485	2269	0.000	95733	145585	0.000	141812	212399
(1-100) - (1001-5000)	0.000	35700	59053	0.000	32345	53464	0.000	1485	2600	0.000	95733	163463	0.000	141812	247122
(1-100) - (5001-10000)	0.000	35700	60598	0.000	32345	54762	0.000	1485	2650	0.000	95733	166130	0.000	141812	255688
(1-100) - (10001+)	0.000	35700	60608	0.000	32345	55760	0.000	1485	2696	0.000	95733	169018	0.000	141812	261486
(101-1000) - (1001-5000)	1.000	54955	59053	0.000	47051	53464	0.068	2269	2600	0.000	145585	163463	0.000	212399	247122
(101-1000) - (5001-10000)	1.000	54955	60598	0.000	47051	54762	1.000	2269	2650	0.000	145585	166130	0.000	212399	255688
(101-1000) - (10001+)	1.000	54955	60608	0.000	47051	55760	0.080	2269	2696	0.000	145585	169018	0.000	212399	261486
(1001-5000) - (5001-10000)	1.000	59053	60598	1.000	53464	54762	1.000	2600	2650	1.000	163463	166130	0.009	247122	255688
(1001-5000) - (10001+)	1.000	59053	60608	0.540	53464	55760	1.000	2600	2696	0.024	163463	169018	0.000	247122	261486
(5001-10000) - (10001+)	1.000	60598	60608	1.000	54762	55760	1.000	2650	2696	1.000	166130	169018	0.896	255688	261486
Community Communication	**Macedonian**			**Romanian**			**Russian**			**Spanish**			**Turkish**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
(1-100) - (101-1000)	0.000	2292	3331	0.000	21206	30975	0.000	153905	233968	0.000	325029	468945	0.000	70282	99831
(1-100) - (1001-5000)	0.000	2292	3864	0.000	21206	37057	0.000	153905	261058	0.000	325029	526378	0.000	70282	111524
(1-100) - (5001-10000)	0.000	2292	3870	0.000	21206	37807	0.000	153905	266380	0.000	325029	540550	0.000	70282	114665
(1-100) - (10001+)	0.000	2292	3803	0.000	21206	37771	0.000	153905	267383	0.000	325029	546741	0.000	70282	115431
(101-1000) - (1001-5000)	0.872	3331	3864	0.000	30975	37057	0.000	233968	261058	0.000	468945	526378	0.000	99831	111524
(101-1000) - (5001-10000)	0.153	3331	3870	0.000	30975	37807	0.000	233968	266380	0.000	468945	540550	0.000	99831	114665
(101-1000) - (10001+)	1.000	3331	3803	0.000	30975	37771	0.000	233968	267383	0.000	468945	546741	0.000	99831	115431
(1001-5000) - (5001-10000)	1.000	3864	3870	1.000	37057	37807	0.637	261058	266380	0.281	526378	540550	1.000	111524	114665
(1001-5000) - (10001+)	1.000	3864	3803	1.000	37057	37771	0.146	261058	267383	0.002	526378	546741	1.000	111524	115431
(5001-10000) - (10001+)	1.000	3870	3803	1.000	37807	37771	1.000	266380	267383	1.000	540550	546741	1.000	114665	115431

Table 33. Editor proportion of edits in Personal Communication by edit bucket. *Values correspond to Kruskal-Wallis test mean ranks for each group and p-values for the pairwise comparison among groups according to Dunn's test with Bonferroni correction.*

Personal Communication	**Arabic**			**Basque**			**Catalan**			**English**			**German**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
(1-100) - (101-1000)	0.000	78778	91432	0.000	1450	2235	0.000	17562	24325	0.000	3417806	4810645	0.000	334903	449545
(1-100) - (1001-5000)	0.000	78778	97395	0.000	1450	2281	0.000	17562	26422	0.000	3417806	5088740	0.000	334903	449633
(1-100) - (5001-10000)	0.000	78778	99229	0.000	1450	2323	0.000	17562	26346	0.000	3417806	5164337	0.000	334903	446300
(1-100) - (10001+)	0.000	78778	98985	0.001	1450	2303	0.000	17562	26213	0.000	3417806	5189198	0.000	334903	442615
(101-1000) - (1001-5000)	0.194	91432	97395	1.000	2235	2281	0.194	24325	26422	0.000	4810645	5088740	1.000	449545	449633
(101-1000) - (5001-10000)	0.625	91432	99229	1.000	2235	2323	0.625	24325	26346	0.000	4810645	5164337	1.000	449545	446300
(101-1000) - (10001+)	0.004	91432	98985	1.000	2235	2303	0.004	24325	26213	0.000	4810645	5189198	1.000	449545	442615
(1001-5000) - (5001-10000)	1.000	97395	99229	1.000	2281	2323	1.000	26422	26346	0.033	5088740	5164337	1.000	449633	446300
(1001-5000) - (10001+)	1.000	97395	98985	1.000	2281	2303	1.000	26422	26213	0.000	5088740	5189198	1.000	449633	442615
(5001-10000) - (10001+)	1.000	99229	98985	1.000	2323	2303	1.000	26346	26213	1.000	5164337	5189198	1.000	446300	442615
Personal Communication	**Hebrew**			**Hungarian**			**Icelandic**			**Italian**			**Japanese**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
(1-100) - (101-1000)	0.000	35929	51921	0.000	32315	48763	0.000	1500	2010	0.000	96052	142269	0.000	142020	209772
(1-100) - (1001-5000)	0.000	35929	53259	0.000	32315	51740	0.000	1500	2244	0.000	96052	153576	0.000	142020	241267
(1-100) - (5001-10000)	0.000	35929	53298	0.000	32315	52063	0.060	1500	2296	0.000	96052	155077	0.000	142020	251279
(1-100) - (10001+)	0.000	35929	53279	0.000	32315	52185	0.000	1500	2297	0.000	96052	155817	0.000	142020	253209
(101-1000) - (1001-5000)	1.000	51921	53259	0.002	48763	51740	1.000	2010	2244	0.000	142269	153576	0.000	209772	241267
(101-1000) - (5001-10000)	1.000	51921	53298	0.130	48763	52063	1.000	2010	2296	0.000	142269	155077	0.000	209772	251279
(101-1000) - (10001+)	1.000	51921	53279	0.018	48763	52185	1.000	2010	2297	0.000	142269	155817	0.000	209772	253209
(1001-5000) - (5001-10000)	1.000	53259	53298	1.000	51740	52063	1.000	2244	2296	1.000	153576	155077	0.002	241267	251279
(1001-5000) - (10001+)	1.000	53259	53279	1.000	51740	52185	1.000	2244	2297	1.000	153576	155817	0.000	241267	253209
(5001-10000) - (10001+)	1.000	53298	53279	1.000	52063	52185	1.000	2296	2297	1.000	155077	155817	1.000	251279	253209
Personal Communication	**Macedonian**			**Romanian**			**Russian**			**Spanish**			**Turkish**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
(1-100) - (101-1000)	0.000	2299	3222	0.000	21244	30292	0.000	154415	225707	0.000	324801	480301	0.000	70228	103408
(1-100) - (1001-5000)	0.000	2299	3568	0.000	21244	34211	0.000	154415	251939	0.000	324801	525977	0.000	70228	112438
(1-100) - (5001-10000)	0.000	2299	3569	0.000	21244	34853	0.000	154415	253033	0.000	324801	534746	0.000	70228	113538
(1-100) - (10001+)	0.004	2299	3505	0.000	21244	35028	0.000	154415	252778	0.000	324801	538381	0.000	70228	114752
(101-1000) - (1001-5000)	1.000	3222	3568	0.000	30292	34211	0.000	225707	251939	0.000	480301	525977	0.000	103408	112438
(101-1000) - (5001-10000)	1.000	3222	3569	0.000	30292	34853	0.000	225707	253033	0.000	480301	534746	0.016	103408	113538
(101-1000) - (10001+)	1.000	3222	3505	0.000	30292	35028	0.000	225707	252778	0.000	480301	538381	0.000	103408	114752
(1001-5000) - (5001-10000)	1.000	3568	3569	1.000	34211	34853	1.000	251939	253033	1.000	525977	534746	1.000	112438	113538
(1001-5000) - (10001+)	1.000	3568	3505	1.000	34211	35028	1.000	251939	252778	0.294	525977	538381	1.000	112438	114752
(5001-10000) - (10001+)	1.000	3569	3505	1.000	34853	35028	1.000	253033	252778	1.000	534746	538381	1.000	113538	114752

3.4 Editor and Reader Engagement with CIRA

Table 34. Page views by content type (CIRA KW-GL, CIRA KW, CIRA GL, CIRA rest and WP rest). *Values correspond to Kruskal-Wallis test mean ranks for each group and p-values for the pairwise comparison among groups according to Dunn's test with Bonferroni correction.*

Page views	**Arabic**			**Basque**			**Catalan**			**English**			**German**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	0.000	174928	249623	0.000	100220	149178	0.000	226112	344451	0.000	2203191	2714416	0.000	850468	980990
WP rest - CIRA rest	0.000	174928	211799	0.000	100220	139770	0.000	226112	320783	0.000	2203191	2799549	0.000	850468	1027248
WP rest - CIRA GL	0.000	174928	270021	0.000	100220	144771	0.000	226112	221513	0.000	2203191	2566498	0.000	850468	1051316
WP rest - CIRA KW-GL	0.000	174928	298913	0.000	100220	185864	0.000	226112	372024	0.000	2203191	2621479	0.000	850468	1191907
CIRA KW - CIRA rest	0.000	249623	211799	0.000	149178	139770	0.000	344451	320783	0.000	2714416	2799549	0.000	980990	1027248
CIRA KW - CIRA GL	0.000	249623	270021	0.000	149178	144771	0.000	344451	221513	0.000	2714416	2566498	0.000	980990	1051316
CIRA KW - CIRA KW-GL	0.000	249623	298913	0.002	149178	185864	0.000	344451	372024	0.000	2714416	2621479	0.000	980990	1191907
CIRA rest - CIRA GL	0.000	211799	270021	0.566	139770	144771	0.000	320783	221513	0.000	2799549	2566498	0.000	1027248	1051316
CIRA rest - CIRA KW-GL	0.000	211799	298913	0.009	139770	185864	0.000	320783	372024	0.000	2799549	2621479	0.000	1027248	1191907
CIRA GL - CIRA KW-GL	0.000	270021	298913	0.032	144771	185864	0.000	221513	372024	0.000	2566498	2621479	0.000	1051316	1191907
Page views	**Hebrew**			**Hungarian**			**Icelandic**			**Italian**			**Japanese**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	0.000	82149	103719	0.000	152022	196386	0.000	21209.44	24116.82	0.000	562063	755232	0.000	451398	555455
WP rest - CIRA rest	0.000	82149	94977	0.000	152022	202851	0.000	21209.44	15082.70	0.000	562063	775188	0.000	451398	513959
WP rest - CIRA GL	0.000	82149	127252	0.000	152022	266990	0.000	21209.44	29156.77	0.000	562063	839191	0.000	451398	640870
WP rest - CIRA KW-GL	0.000	82149	121954	0.000	152022	267768	0.000	21209.44	32894.48	0.000	562063	1005777	0.000	451398	752618
CIRA KW - CIRA rest	0.000	103719	94977	0.000	196386	202851	0.000	24116.82	15082.70	0.000	755232	775188	0.000	555455	513959
CIRA KW - CIRA GL	0.000	103719	127252	0.000	196386	266990	0.000	24116.82	29156.77	0.000	755232	839191	0.000	555455	640870
CIRA KW - CIRA KW-GL	0.000	103719	121954	0.000	196386	267768	0.000	24116.82	32894.48	0.000	755232	1005777	0.000	555455	752618
CIRA rest - CIRA GL	0.015	94977	127252	0.000	202851	266990	0.000	15082.70	29156.77	0.000	775188	839191	0.000	513959	640870
CIRA rest - CIRA KW-GL	0.000	94977	121954	0.000	202851	267768	0.000	15082.70	32894.48	0.000	775188	1005777	0.000	513959	752618
CIRA GL - CIRA KW-GL	1.000	127252	121954	1.000	266990	267768	0.733	29156.77	32894.48	0.000	839191	1005777	0.000	640870	752618
Page views	**Macedonian**			**Romanian**			**Russian**			**Spanish**			**Turkish**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	1.000	41257	43466	0.000	150158	250070	0.000	604667.48	768412.16	0.000	520390.25	728382.02	0.000	118873	136848
WP rest - CIRA rest	0.000	41257	41377	0.000	150158	210767	0.000	604667.48	710883.96	0.000	520390.25	707356.51	0.000	118873	136489
WP rest - CIRA GL	0.001	41257	43001	0.000	150158	213301	0.000	604667.48	513234.10	0.000	520390.25	730249.69	0.000	118873	128604
WP rest - CIRA KW-GL	1.000	41257	44326	0.000	150158	237757	0.000	604667.48	618330.47	0.000	520390.25	836553.61	0.000	118873	155726
CIRA KW - CIRA rest	0.003	43466	41377	0.010	250070	210767	0.432	768412.16	710883.96	0.000	728382.02	707356.51	0.000	136848	155726
CIRA KW - CIRA GL	0.005	43466	43001	0.000	250070	213301	0.000	768412.16	513234.10	0.000	728382.02	730249.69	0.000	136848	128604
CIRA KW - CIRA KW-GL	1.000	43466	44326	0.000	250070	237757	0.000	768412.16	618330.47	0.000	728382.02	836553.61	0.000	136848	155726
CIRA rest - CIRA GL	1.000	41377	43001	0.000	210767	213301	0.000	710883.96	513234.10	1.000	707356.51	730249.69	1.000	136489	128604
CIRA rest - CIRA KW-GL	1.000	41377	44326	0.000	210767	237757	0.000	710883.96	618330.47	0.000	707356.51	836553.61	0.002	136489	155726
CIRA GL - CIRA KW-GL	1.000	43001	44326	0.190	213301	237757	0.000	513234.10	618330.47	0.000	730249.69	836553.61	0.003	128604	155726

Table 35. Edits by content type (CIRA KW-GL, CIRA KW, CIRA GL, CIRA rest and WP rest). *Values correspond to Kruskal-Wallis test mean ranks for each group and p-values for the pairwise comparison among groups according to Dunn's test with Bonferroni correction.*

Edits	**Arabic**			**Basque**			**Catalan**			**English**			**German**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	0.000	173681	246308	0.000	97755	166467	0.000	221303	349292	0.000	2211878	2623270	0.000	853503	1003962
WP rest - CIRA rest	0.000	173681	216993	0.000	97755	160888	0.000	221303	334169	0.000	2211878	2795003	0.000	853503	1027417
WP rest - CIRA GL	0.000	173681	266367	0.000	97755	172850	0.000	221303	259115	0.000	2211878	2562090	0.000	853503	1025320
WP rest - CIRA KW-GL	0.000	173681	290758	0.000	97755	197027	0.000	221303	382121	0.000	2211878	2498244	0.000	853503	1145545
CIRA KW - CIRA rest	0.000	246308	216993	0.058	166467	160888	0.000	349292	334169	0.000	2623270	2795003	0.000	1003962	1027417
CIRA KW - CIRA GL	0.000	246308	266367	0.000	166467	172850	0.000	349292	259115	0.000	2623270	2562090	0.000	1003962	1025320
CIRA KW - CIRA KW-GL	0.000	246308	290758	0.020	166467	197027	0.000	349292	382121	0.000	2623270	2498244	0.000	1003962	1145545
CIRA rest - CIRA GL	0.000	216993	266367	0.037	160888	172850	0.000	334169	259115	0.000	2795003	2562090	1.000	1027417	1025320
CIRA rest - CIRA KW-GL	0.000	216993	290758	0.099	160888	197027	0.000	334169	382121	0.000	2795003	2498244	0.000	1027417	1145545
CIRA GL - CIRA KW-GL	0.000	266367	290758	0.393	172850	197027	0.000	259115	382121	0.000	2498244	2498244	0.000	1003962	1145545
Edits	**Hebrew**			**Hungarian**			**Icelandic**			**Italian**			**Japanese**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	0.000	78834	112869	0.000	150715	193815	0.000	19240	26267	0.000	553683	833045	0.000	420352	577648
WP rest - CIRA rest	0.000	78834	102682	0.000	150715	210794	0.000	19240	20042	0.000	553683	805224	0.000	420352	545199
WP rest - CIRA GL	0.000	78834	125163	0.000	150715	261092	0.000	19240	28703	0.000	553683	888808	0.000	420352	686429
WP rest - CIRA KW-GL	0.000	78834	129083	0.000	150715	279864	0.000	19240	32655	0.000	553683	1055822	0.000	420352	761386
CIRA KW - CIRA rest	0.000	112869	102682	0.000	193815	210794	0.000	26267	20042	0.000	833045	805224	0.000	577648	545199
CIRA KW - CIRA GL	0.000	112869	125163	0.000	193815	261092	0.000	26267	28703	0.000	833045	888808	0.000	577648	686429
CIRA KW - CIRA KW-GL	0.000	112869	129083	0.000	193815	279864	0.000	26267	32655	0.000	833045	1055822	0.000	577648	761386
CIRA rest - CIRA GL	0.000	102682	125163	0.000	210794	261092	0.001	20042	28703	0.000	805224	888808	0.000	545199	686429
CIRA rest - CIRA KW-GL	0.047	102682	129083	0.000	210794	279864	0.024	20042	32655	0.000	805224	1055822	0.000	545199	761386
CIRA GL - CIRA KW-GL	1.000	125163	129083	0.088	261092	279864	0.577	28703	32655	0.000	888808	1055822	0.000	686429	761386
Edits	**Macedonian**			**Romanian**			**Russian**			**Spanish**			**Turkish**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	0.000	38754	64058	0.000	145402	249560	0.000	591888	804657	0.000	505647	756783	0.000	110832.1	129430.6
WP rest - CIRA rest	0.000	38754	52545	0.000	145402	230892	0.000	591888	754215	0.000	505647	743856	0.000	110832.1	150433.2
WP rest - CIRA GL	0.000	38754	63015	0.000	145402	230016	0.000	591888	515294	0.000	505647	783634	0.000	110832.1	166207.3
WP rest - CIRA KW-GL	0.000	38754	70510	0.000	145402	261977	0.000	591888	583347	0.000	505647	833199	0.000	110832.1	158899.8
CIRA KW - CIRA rest	0.000	64058	52545	1.000	249560	230892	1.000	804657	754215	0.000	756783	743856	0.000	129430.6	150433.2
CIRA KW - CIRA GL	0.000	64058	63015	0.000	249560	230016	0.000	804657	515294	0.000	756783	783634	0.000	129430.6	166207.3
CIRA KW - CIRA KW-GL	0.000	64058	70510	0.000	249560	261977	0.000	804657	583347	0.000	756783	833199	0.000	129430.6	158899.8
CIRA rest - CIRA GL	1.000	52545	63015	0.000	230892	230016	0.000	754215	515294	0.000	743856	783634	0.953	150433.2	166207.3
CIRA rest - CIRA KW-GL	0.677	52545	70510	0.000	230892	261977	0.000	754215	583347	0.000	743856	833199	0.000	150433.2	158899.8
CIRA GL - CIRA KW-GL	1.000	63015	70510	0.157	230016	261977	0.000	515294	583347	0.000	783634	833199	1.000	166207.3	158899.8

3.5 CIRA Article Features Analysis

Table 36. Bytes by content type (CIRA KW-GL, CIRA KW, CIRA GL, CIRA rest and WP rest). *Values correspond to Kruskal-Wallis test mean ranks for each group and p-values for the pairwise comparison among groups according to Dunn's test with Bonferroni correction.*

Bytes	**Arabic**			**Basque**			**Catalan**			**English**			**German**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	0.000	180807	216901	0.000	102555	144308	0.000	235238	307325	0.000	2090731	3008833	0.000	819056	1093038
WP rest - CIRA rest	0.000	180807	203973	0.000	102555	117786	0.000	235238	248388	0.000	2090731	2897407	0.000	819056	1077530
WP rest - CIRA GL	0.000	180807	211588	0.000	102555	125051	0.000	235238	196526	0.000	2090731	2753012	0.000	819056	1106223
WP rest - CIRA KW-GL	0.000	180807	244726	0.000	102555	158247	0.000	235238	316679	0.000	2090731	2955132	0.000	819056	1212649
CIRA KW - CIRA rest	0.000	216901	203973	0.000	144308	117786	0.000	307325	248388	0.000	3008833	2897407	0.000	1093038	1077530
CIRA KW - CIRA GL	0.000	216901	211588	0.000	144308	125051	0.000	307325	196526	0.000	3008833	2753012	0.000	1093038	1106223
CIRA KW - CIRA KW-GL	0.000	216901	244726	0.010	144308	158247	0.000	307325	316679	0.000	3008833	2955132	0.000	1093038	1212649
CIRA rest - CIRA GL	0.006	203973	211588	0.000	117786	125051	0.000	307325	196526	0.000	2897407	2753012	0.001	1077530	1106223
CIRA rest - CIRA KW-GL	0.000	203973	244726	0.071	117786	158247	0.000	307325	316679	0.000	2897407	2955132	0.000	1077530	1212649
CIRA GL - CIRA KW-GL	0.000	211588	244726	1.000	125051	158247	1.000	196526	316679	0.000	2753012	2955132	0.000	1106223	1212649
Bytes	**Hebrew**			**Hungarian**			**Icelandic**			**Italian**			**Japanese**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	0.000	82153	114617	0.000	155182	223555	0.000	17940	26652	0.000	570527	887202	0.000	468203	587482
WP rest - CIRA rest	0.000	82153	96636	0.000	155182	189917	0.000	17940	23810	0.000	570527	740435	0.000	468203	495769
WP rest - CIRA GL	0.000	82153	97780	0.000	155182	215896	0.000	17940	22839	0.000	570527	773619	0.000	468203	620582
WP rest - CIRA KW-GL	0.000	82153	118740	0.000	155182	242347	0.000	17940	30219	0.000	570527	925618	0.000	468203	721346
CIRA KW - CIRA rest	1.000	114617	96636	0.000	223555	189917	0.179	26652	23810	0.000	887202	740435	0.000	587482	495769
CIRA KW - CIRA GL	0.000	114617	97780	0.000	223555	215896	0.000	26652	22839	0.000	887202	773619	0.000	587482	620582
CIRA KW - CIRA KW-GL	0.001	114617	118740	0.000	223555	242347	0.004	26652	30219	0.000	887202	925618	0.000	587482	721346
CIRA rest - CIRA GL	0.000	96636	97780	0.000	189917	215896	0.000	23810	22839	0.000	740435	773619	0.000	495769	620582
CIRA rest - CIRA KW-GL	0.002	96636	118740	0.002	189917	242347	0.018	23810	30219	0.000	740435	925618	0.000	495769	721346
CIRA GL - CIRA KW-GL	1.000	97780	118740	0.092	215896	242347	0.903	22839	30219	0.031	773619	925618	0.000	620582	721346
Bytes	**Macedonian**			**Romanian**			**Russian**			**Spanish**			**Turkish**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	0.000	41742	54849	0.000	159434	223391	0.000	562474	838433	0.000	528597	786504	0.000	117493	143878
WP rest - CIRA rest	0.000	41742	36395	0.000	159434	180257	0.000	562474	787254	0.000	528597	683283	0.000	117493	136038
WP rest - CIRA GL	0.000	41742	45148	0.000	159434	169238	0.000	562474	637266	0.000	528597	692432	0.000	117493	149487
WP rest - CIRA KW-GL	0.000	41742	63361	0.000	159434	197772	0.000	562474	630487	0.000	528597	787955	0.000	117493	150316
CIRA KW - CIRA rest	0.000	54849	36395	0.000	223391	180257	1.000	838433	787254	0.000	786504	683283	0.000	143878	136038
CIRA KW - CIRA GL	0.000	54849	45148	0.000	223391	169238	0.000	838433	637266	0.000	786504	692432	0.000	143878	149487
CIRA KW - CIRA KW-GL	0.000	54849	63361	0.000	223391	197772	0.000	838433	630487	0.000	786504	787955	0.049	143878	150316
CIRA rest - CIRA GL	0.000	36395	45148	0.005	180257	169238	0.000	787254	637266	0.000	683283	692432	0.000	136038	149487
CIRA rest - CIRA KW-GL	0.000	36395	63361	0.000	180257	197772	0.000	787254	630487	0.000	683283	787955	1.000	136038	150316
CIRA GL - CIRA KW-GL	0.408	45148	63361	0.000	169238	197772	0.000	637266	630487	1.000	692432	787955	1.000	149487	150316

Table 37. Discussion Bytes by content type (CIRA KW-GL, CIRA KW, CIRA GL, CIRA rest and WP rest). *Values correspond to Kruskal-Wallis test mean ranks for each group and p-values for the pairwise comparison among groups according to Dunn's test with Bonferroni correction.*

Discussion Bytes	**Arabic**			**Basque**			**Catalan**			**English**			**German**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	0.000	183770	196210	0.000	103787	115371	0.000	236036	252938	0.000	2144559	2695197	0.000	878448	987030
WP rest - CIRA rest	0.000	183770	197852	0.000	103787	108668	0.000	236036	228217	0.000	2144559	2855946	0.000	878448	983864
WP rest - CIRA GL	0.000	183770	200803	0.000	103787	109265	0.000	236036	212407	0.000	2144559	2694333	0.000	878448	980571
WP rest - CIRA KW-GL	0.000	183770	213184	0.000	103787	124462	0.000	236036	263869	0.000	2144559	2617742	0.000	878448	1038480
CIRA KW - CIRA rest	0.095	196210	197852	0.175	115371	108668	0.000	252938	228217	0.000	2695197	2855946	0.081	987030	983864
CIRA KW - CIRA GL	0.000	196210	200803	0.000	115371	109265	0.000	252938	212407	0.000	2695197	2694333	0.161	987030	980571
CIRA KW - CIRA KW-GL	0.000	196210	213184	0.000	115371	124462	0.000	252938	263869	0.000	2695197	2617742	0.000	987030	1038480
CIRA rest - CIRA GL	0.000	197852	200803	0.000	108668	109265	0.000	228217	212407	1.000	2855946	2694333	1.000	983864	980571
CIRA rest - CIRA KW-GL	0.000	197852	213184	0.000	108668	124462	0.000	228217	263869	0.000	2855946	2617742	0.000	983864	1038480
CIRA GL - CIRA KW-GL	0.000	200803	213184	0.011	109265	124462	0.101	212407	263869	0.000	2617742	2617742	0.000	980571	1038480
Discussion Bytes	**Hebrew**			**Hungarian**			**Icelandic**			**Italian**			**Japanese**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	0.000	81694	108239	0.000	158717	205414	0.002	18519	21940	0.000	586433	685346	0.000	482261	519264
WP rest - CIRA rest	0.000	81694	96811	0.000	158717	174805	0.000	18519	22843	0.000	586433	659731	0.000	482261	489626
WP rest - CIRA GL	0.000	81694	114537	0.000	158717	199616	0.000	18519	19650	0.000	586433	785534	0.000	482261	511864
WP rest - CIRA KW-GL	0.000	81694	122987	0.000	158717	205322	0.013	18519	23567	0.000	586433	990000	0.000	482261	626757
CIRA KW - CIRA rest	0.000	108239	96811	0.000	205414	174805	0.000	21940	22843	0.000	685346	659731	0.000	519264	489626
CIRA KW - CIRA GL	0.000	108239	114537	0.000	205414	199616	0.000	21940	19650	0.000	685346	785534	0.000	519264	511864
CIRA KW - CIRA KW-GL	0.000	108239	122987	0.000	205414	205322	0.144	21940	23567	0.000	685346	990000	0.000	519264	626757
CIRA rest - CIRA GL	0.000	96811	114537	1.000	174805	199616	0.174	22843	19650	0.000	659731	785534	0.002	489626	511864
CIRA rest - CIRA KW-GL	0.058	96811	122987	0.010	174805	205322	1.000	22843	23567	0.000	659731	990000	0.000	489626	626757
CIRA GL - CIRA KW-GL	1.000	114537	122987	1.000	199616	205322	1.000	19650	23567	0.000	785534	990000	0.000	511864	626757
Discussion Bytes	**Macedonian**			**Romanian**			**Russian**			**Spanish**			**Turkish**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	0.018	41155	43604	0.000	159225	195933	0.000	624374	517095	0.000	557360	650916	0.000	110617	122865
WP rest - CIRA rest	0.000	41155	40373	0.000	159225	183146	0.000	624374	598314	0.000	557360	606584	0.000	110617	150287
WP rest - CIRA GL	0.000	41155	51311	0.000	159225	170531	0.000	624374	517095	0.000	557360	647033	0.000	110617	173297
WP rest - CIRA KW-GL	0.000	41155	58965	0.000	159225	185705	0.000	624374	598314	0.000	557360	672978	0.000	110617	166061
CIRA KW - CIRA rest	0.010	43604	40373	0.000	195933	183146	0.000	517095	598314	0.000	650916	606584	0.000	122865	150287
CIRA KW - CIRA GL	0.000	43604	51311	0.000	195933	170531	0.000	517095	517095	0.000	650916	647033	0.000	122865	173297
CIRA KW - CIRA KW-GL	0.000	43604	58965	0.000	195933	185705	0.000	517095	598314	0.000	650916	672978	0.000	122865	166061
CIRA rest - CIRA GL	0.000	40373	51311	1.000	183146	170531	0.000	598314	517095	0.536	606584	647033	0.004	150287	173297
CIRA rest - CIRA KW-GL	0.002	40373	58965	0.000	183146	185705	0.000	598314	598314	0.000	606584	672978	0.000	150287	166061
CIRA GL - CIRA KW-GL	0.603	51311	58965	0.005	170531	185705	0.000	517095	598314	0.000	647033	672978	1.000	173297	166061

Table 38. Images by content type (CIRA KW-GL, CIRA KW, CIRA GL, CIRA rest and WP rest). *Values correspond to Kruskal-Wallis test mean ranks for each group and p-values for the pairwise comparison among groups according to Dunn's test with Bonferroni correction.*

Images	**Arabic**			**Basque**			**Catalan**			**English**			**German**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	0.000	186487	173691	1.000	107063	66342	0.000	241104	189358	0.000	2457625	2515258	0.000	929258	850571
WP rest - CIRA rest	0.000	186487	186294	0.000	107063	67213	0.000	241104	150451	0.000	2457625	2224949	0.000	929258	749425
WP rest - CIRA GL	0.000	186487	231585	0.000	107063	143115	0.000	241104	236125	0.000	2457625	3259721	0.000	929258	1336484
WP rest - CIRA KW-GL	0.000	186487	243895	0.000	107063	135567	0.000	241104	295188	0.000	2457625	3285670	0.000	929258	1324662
CIRA KW - CIRA rest	1.000	173691	186294	0.000	66342	67213	0.000	189358	150451	0.000	2515258	2224949	0.000	850571	749425
CIRA KW - CIRA GL	0.000	173691	231585	0.000	66342	143115	0.000	189358	236125	0.000	2515258	3259721	0.000	850571	1336484
CIRA KW - CIRA KW-GL	0.000	173691	243895	0.000	66342	135567	0.000	189358	295188	0.000	2515258	3285670	0.000	850571	1324662
CIRA rest - CIRA GL	0.000	186294	231585	0.194	67213	143115	0.000	150451	236125	0.000	2224949	3259721	0.000	749425	1336484
CIRA rest - CIRA KW-GL	0.000	186294	243895	0.000	67213	135567	0.000	150451	295188	0.000	2224949	3285670	0.000	749425	1324662
CIRA GL - CIRA KW-GL	0.032	231585	243895	1.000	143115	135567	0.000	236125	295188	0.119	3259721	3285670	1.000	1336484	1324662
Images	**Hebrew**			**Hungarian**			**Icelandic**			**Italian**			**Japanese**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	0.000	89282	87609	0.000	169635	141505	0.000	21767	19535	0.000	593749	713633	0.000	500689	416219
WP rest - CIRA rest	0.000	89282	77912	0.000	169635	113969	0.000	21767	14293	0.000	593749	601062	0.000	500689	454134
WP rest - CIRA GL	0.000	89282	135021	0.000	169635	262666	0.000	21767	24376	0.000	593749	862838	0.000	500689	733822
WP rest - CIRA KW-GL	0.000	89282	127907	0.000	169635	270199	0.000	21767	24896	0.000	593749	983212	0.000	500689	706529
CIRA KW - CIRA rest	0.803	87609	77912	0.000	141505	113969	0.000	19535	14293	0.000	713633	601062	0.000	416219	454134
CIRA KW - CIRA GL	0.000	87609	135021	0.000	141505	262666	0.000	19535	24376	0.000	713633	862838	0.000	416219	733822
CIRA KW - CIRA KW-GL	0.000	87609	127907	0.000	141505	270199	0.084	19535	24896	0.000	713633	983212	0.000	416219	706529
CIRA rest - CIRA GL	0.000	77912	135021	0.000	113969	262666	0.000	14293	24376	0.000	601062	862838	0.000	454134	733822
CIRA rest - CIRA KW-GL	0.000	77912	127907	0.000	113969	270199	1.000	14293	24896	0.000	601062	983212	0.000	454134	733822
CIRA GL - CIRA KW-GL	1.000	135021	127907	1.000	262666	270199	1.000	24376	24896	0.000	862838	983212	1.000	733822	733822
Images	**Macedonian**			**Romanian**			**Russian**			**Spanish**			**Turkish**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	0.000	44329	44329	0.000	167854	120565	0.000	574920	595625	0.000	575098	642799	0.433	124747	116878
WP rest - CIRA rest	0.000	44329	45170	0.000	167854	107440	0.000	574920	625888	0.000	575098	502628	0.000	124747	118993
WP rest - CIRA GL	0.000	44329	66348	0.000	167854	217815	0.000	574920	856783	0.000	575098	825015	0.000	124747	159276
WP rest - CIRA KW-GL	0.000	44329	62089	0.000	167854	237537	0.000	574920	785753	0.000	575098	843614	0.000	124747	155474
CIRA KW - CIRA rest	1.000	44329	45170	0.000	120565	107440	0.000	595625	625888	0.000	642799	502628	0.000	116878	118993
CIRA KW - CIRA GL	0.000	44329	66348	0.000	120565	217815	0.000	595625	856783	0.000	642799	825015	0.000	116878	159276
CIRA KW - CIRA KW-GL	0.000	44329	62089	0.000	120565	237537	0.000	595625	785753	0.000	642799	843614	0.000	116878	155474
CIRA rest - CIRA GL	0.000	45170	66348	0.000	107440	217815	0.000	625888	856783	0.000	502628	825015	0.000	118993	159276
CIRA rest - CIRA KW-GL	0.000	45170	62089	0.000	107440	237537	0.000	625888	785753	0.000	502628	843614	0.000	118993	155474
CIRA GL - CIRA KW-GL	1.000	66348	62089	0.001	217815	237537	0.000	856783	785753	0.039	825015	843614	1.000	159276	155474

Table 39. External references by content type (CIRA KW-GL, CIRA KW, CIRA GL, CIRA rest and WP rest). *Values correspond to Kruskal-Wallis test mean ranks for each group and p-values for the pairwise comparison among groups according to Dunn's test with Bonferroni correction.*

External References	**Arabic**			**Basque**			**Catalan**			**English**			**German**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	0.001	189238	170680	0.076	107865	71431	0.000	236391	202655	0.000	2111299	2436205	0.000	803412	918187
WP rest - CIRA rest	0.000	189238	175317	0.000	107865	65819	0.000	236391	185404	0.000	2111299	2745786	0.000	803412	1093126
WP rest - CIRA GL	0.000	189238	244129	0.000	107865	104692	0.000	236391	251907	0.000	2111299	3315235	0.000	803412	1205243
WP rest - CIRA KW-GL	0.000	189238	257558	0.000	107865	141523	0.000	236391	316163	0.000	2111299	3402319	0.000	803412	1268672
CIRA KW - CIRA rest	0.000	170680	175317	0.000	71431	65819	0.000	202655	185404	0.000	2436205	2745786	0.000	918187	1093126
CIRA KW - CIRA GL	0.000	170680	244129	0.000	71431	104692	0.000	202655	251907	0.000	2436205	3315235	0.000	918187	1205243
CIRA KW - CIRA KW-GL	0.000	170680	257558	0.000	71431	141523	0.000	202655	316163	0.000	2436205	3402319	0.000	918187	1268672
CIRA rest - CIRA GL	0.000	175317	244129	0.020	65819	104692	0.000	185404	251907	0.000	2745786	3315235	0.000	1093126	1205243
CIRA rest - CIRA KW-GL	0.000	175317	257558	0.026	65819	141523	0.000	185404	316163	0.000	2745786	3402319	0.000	1093126	1268672
CIRA GL - CIRA KW-GL	0.011	244129	257558	0.057	104692	141523	0.000	251907	316163	0.000	3315235	3402319	0.000	1205243	1268672
External References	**Hebrew**			**Hungarian**			**Icelandic**			**Italian**			**Japanese**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	0.000	77688	110351	0.000	160679	142403	0.000	17896	25640	0.000	603272	538524	0.000	487935	487930
WP rest - CIRA rest	0.000	77688	104819	0.000	160679	163278	0.000	17896	23613	0.000	603272	573657	0.000	487935	464993
WP rest - CIRA GL	0.000	77688	135107	0.000	160679	251937	0.000	17896	27282	0.000	603272	791251	0.000	487935	759839
WP rest - CIRA KW-GL	0.000	77688	144099	0.000	160679	266776	0.000	17896	34985	0.000	603272	981074	0.000	487935	797062
CIRA KW - CIRA rest	0.000	110351	104819	0.000	142403	163278	0.000	25640	23613	0.000	538524	573657	1.000	487930	464993
CIRA KW - CIRA GL	0.000	110351	135107	0.000	142403	251937	0.000	25640	27282	0.000	538524	791251	0.000	487930	759839
CIRA KW - CIRA KW-GL	0.000	110351	144099	0.000	142403	266776	0.000	25640	34985	0.000	538524	981074	0.000	487930	797062
CIRA rest - CIRA GL	0.000	104819	135107	0.000	163278	251937	0.036	23613	27282	0.000	573657	791251	0.000	464993	759839
CIRA rest - CIRA KW-GL	0.000	104819	144099	0.000	163278	266776	0.000	23613	34985	0.000	573657	981074	0.000	464993	797062
CIRA GL - CIRA KW-GL	1.000	135107	144099	0.371	251937	266776	0.000	27282	34985	0.000	791251	981074	0.299	759839	797062
External References	**Macedonian**			**Romanian**			**Russian**			**Spanish**			**Turkish**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	0.000	43188	35191	0.000	161817	172103	0.000	560675	595162	0.000	566833	609908	0.000	166482	125904
WP rest - CIRA rest	0.000	43188	29338	0.000	161817	145609	0.000	560675	579289	0.000	566833	558799	0.000	166482	128767
WP rest - CIRA GL	0.000	43188	40472	0.000	161817	209773	0.000	560675	1022590	0.000	566833	721241	0.000	166482	166482
WP rest - CIRA KW-GL	0.000	43188	53151	0.000	161817	262646	0.000	560675	969570	0.000	566833	778247	0.000	166482	163100
CIRA KW - CIRA rest	0.000	35191	29338	0.000	172103	145609	0.000	595162	579289	0.000	609908	558799	0.057	166482	128767
CIRA KW - CIRA GL	0.000	35191	40472	0.000	172103	209773	0.000	595162	1022590	0.000	609908	721241	0.000	166482	166482
CIRA KW - CIRA KW-GL	0.000	35191	53151	0.000	172103	262646	0.000	595162	969570	0.000	609908	778247	0.000	166482	163100
CIRA rest - CIRA GL	0.000	29338	40472	0.000	145609	209773	0.000	579289	1022590	0.000	558799	721241	0.000	128767	166482
CIRA rest - CIRA KW-GL	0.017	29338	53151	0.000	145609	262646	0.000	579289	969570	0.000	558799	778247	0.000	128767	163100
CIRA GL - CIRA KW-GL	0.130	40472	53151	0.000	209773	262646	0.000	1022590	969570	0.000	721241	778247	1.000	166482	163100

Table 40. Redirects by content type (CIRA KW-GL, CIRA KW, CIRA GL, CIRA rest and WP rest). *Values correspond to Kruskal-Wallis test mean ranks for each group and p-values for the pairwise comparison among groups according to Dunn's test with Bonferroni correction.*

Redirects	**Arabic**			**Basque**			**Catalan**			**English**			**German**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	0.000	195966	188709	0.000	102467	129276	0.000	232642	269525	0.000	2508588	2575053	0.000	930261	931976
WP rest - CIRA rest	0.000	195966	158634	0.000	102467	121190	0.000	232642	268619	0.000	2508588	2377581	0.000	930261	904170
WP rest - CIRA GL	0.000	195966	188173	0.000	102467	116968	0.000	232642	208602	0.000	2508588	2427049	0.000	930261	855898
WP rest - CIRA KW-GL	0.000	195966	233883	0.000	102467	171664	0.000	232642	313036	0.000	2508588	2670577	0.000	930261	978657
CIRA KW - CIRA rest	1.000	188709	158634	0.000	129276	121190	0.000	269525	268619	0.000	2575053	2427049	0.000	931976	904170
CIRA KW - CIRA GL	0.000	188709	188173	0.000	129276	116968	0.000	269525	208602	0.000	2575053	2427049	0.000	931976	855898
CIRA KW - CIRA KW-GL	0.000	188709	233883	0.000	129276	171664	0.000	269525	313036	0.000	2575053	2670577	0.000	931976	978657
CIRA rest - CIRA GL	0.000	158634	188173	0.000	121190	116968	1.000	268619	208602	0.000	2377581	2427049	1.000	904170	855898
CIRA rest - CIRA KW-GL	0.000	158634	233883	0.000	121190	171664	0.000	268619	313036	0.000	2377581	2670577	0.000	904170	978657
CIRA GL - CIRA KW-GL	0.000	188173	233883	0.000	116968	171664	0.000	208602	313036	0.000	2427049	2670577	0.000	855898	978657
Redirects	**Hebrew**			**Hungarian**			**Icelandic**			**Italian**			**Japanese**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	0.040	86448	98690	0.000	164934	190162	0.000	20952	21664	0.000	604890	682145	0.000	472067	491188
WP rest - CIRA rest	0.000	86448	87138	0.000	164934	146841	0.000	20952	16696	0.000	604890	594109	0.000	472067	472067
WP rest - CIRA GL	0.000	86448	97172	0.000	164934	167415	0.000	20952	18675	0.000	604890	650058	0.000	472067	513509
WP rest - CIRA KW-GL	0.000	86448	107731	0.000	164934	224268	0.000	20952	30269	0.000	604890	678599	0.000	472067	639253
CIRA KW - CIRA rest	0.000	98690	87138	0.111	190162	146841	0.000	21664	16696	0.000	682145	594109	0.029	491188	472067
CIRA KW - CIRA GL	0.000	98690	97172	0.000	190162	167415	0.000	21664	18675	0.000	682145	650058	0.000	491188	513509
CIRA KW - CIRA KW-GL	0.000	98690	107731	0.000	190162	224268	0.000	21664	30269	0.000	682145	678599	0.000	491188	639253
CIRA rest - CIRA GL	1.000	87138	97172	0.000	146841	167415	0.766	16696	18675	0.032	594109	650058	0.000	472067	513509
CIRA rest - CIRA KW-GL	0.373	87138	107731	0.000	146841	224268	0.000	16696	30269	0.000	594109	678599	0.000	472067	639253
CIRA GL - CIRA KW-GL	0.757	97172	107731	0.000	167415	224268	0.000	18675	30269	1.000	650058	678599	0.000	513509	639253
Redirects	**Macedonian**			**Romanian**			**Russian**			**Spanish**			**Turkish**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	0.003	41053	54033	0.000	153744	193359	0.000	585276	596538	0.000	547183	671247	0.024	121837	124757
WP rest - CIRA rest	0.000	41053	42327	0.000	153744	183234	0.000	585276	811799	0.000	547183	626927	0.000	121837	127694
WP rest - CIRA GL	0.000	41053	39614	0.000	153744	230158	0.000	585276	475201	0.000	547183	702217	0.000	121837	144545
WP rest - CIRA KW-GL	0.000	41053	66156	0.000	153744	227158	0.000	585276	513942	0.000	547183	719789	0.000	121837	168516
CIRA KW - CIRA rest	0.000	54033	42327	0.000	193359	183234	0.000	596538	811799	0.000	671247	626927	0.028	124757	127694
CIRA KW - CIRA GL	0.000	54033	39614	0.000	193359	230158	0.000	596538	475201	0.000	671247	702217	0.000	124757	144545
CIRA KW - CIRA KW-GL	0.000	54033	66156	0.000	193359	227158	0.000	596538	513942	0.000	671247	719789	0.000	124757	168516
CIRA rest - CIRA GL	0.000	42327	39614	0.000	183234	230158	0.003	811799	475201	0.000	626927	702217	0.000	127694	144545
CIRA rest - CIRA KW-GL	0.000	42327	66156	0.000	183234	227158	0.000	811799	513942	0.000	626927	719789	0.000	127694	168516
CIRA GL - CIRA KW-GL	0.001	39614	66156	1.000	230158	227158	0.000	475201	513942	0.034	702217	719789	0.000	144545	168516

Table 41. Categories by content type (CIRA KW-GL, CIRA KW, CIRA GL, CIRA rest and WP rest). *Values correspond to Kruskal-Wallis test mean ranks for each group and p-values for the pairwise comparison among groups according to Dunn's test with Bonferroni correction.*

Categories	**Arabic**			**Basque**			**Catalan**			**English**			**German**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	0.000	184357	192144	0.000	99409	125924	0.000	220344	281146	0.000	2017590	2263784	0.000	768812	958757
WP rest - CIRA rest	0.000	184357	197415	0.000	99409	155201	0.000	220344	342367	0.000	2017590	3169861	0.000	768812	1246658
WP rest - CIRA GL	0.000	184357	192134	0.000	99409	120302	0.000	220344	268985	0.000	2017590	2381995	0.000	768812	992115
WP rest - CIRA KW-GL	0.000	184357	224048	0.000	99409	145966	0.000	220344	298471	0.000	2017590	2265595	0.000	768812	1031234
CIRA KW - CIRA rest	1.000	192144	197415	0.086	125924	155201	0.000	281146	298471	1.000	2263784	3169861	0.000	958757	1246658
CIRA KW - CIRA GL	0.000	192144	192134	0.245	125924	120302	0.000	281146	268985	0.000	2263784	2381995	0.000	958757	992115
CIRA KW - CIRA KW-GL	0.000	192144	224048	0.000	125924	145966	0.000	281146	298471	0.000	2263784	2265595	0.000	958757	1031234
CIRA rest - CIRA GL	0.000	197415	192134	0.822	155201	120302	0.019	342367	268985	0.000	3169861	2381995	0.000	1246658	992115
CIRA rest - CIRA KW-GL	0.000	197415	224048	0.000	155201	145966	0.000	342367	298471	0.000	3169861	2265595	0.000	1246658	1031234
CIRA GL - CIRA KW-GL	0.000	192134	224048	1.000	120302	145966	0.000	268985	298471	0.000	2265595	2265595	0.000	992115	1031234
Categories	**Hebrew**			**Hungarian**			**Icelandic**			**Italian**			**Japanese**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	0.000	77991	93850	0.007	148029	178695	0.000	19311	21667	0.000	842875	544987	0.000	411657.92	434795.06
WP rest - CIRA rest	0.000	77991	106817	0.512	148029	241706	0.000	19311	20528	0.000	842875	842875	0.000	411657.92	566137.75
WP rest - CIRA GL	0.000	77991	112119	0.000	148029	144180	0.000	19311	23530	0.000	842875	527489	0.000	411657.92	582159.52
WP rest - CIRA KW-GL	0.000	77991	116838	0.000	148029	157387	0.061	19311	24557	0.000	842875	395591	0.000	411657.92	661975.30
CIRA KW - CIRA rest	0.000	93850	106817	1.000	178695	241706	0.136	21667	20528	0.001	544987	842875	0.000	434795.06	566137.75
CIRA KW - CIRA GL	0.000	93850	112119	0.000	178695	144180	0.000	21667	23530	0.000	544987	527489	0.000	434795.06	582159.52
CIRA KW - CIRA KW-GL	0.000	93850	116838	0.000	178695	157387	0.353	21667	24557	0.000	544987	395591	0.000	434795.06	661975.30
CIRA rest - CIRA GL	0.000	106817	112119	0.017	241706	144180	0.013	20528	23530	0.000	842875	527489	0.000	566137.75	582159.52
CIRA rest - CIRA KW-GL	0.717	106817	116838	0.000	241706	157387	1.000	20528	24557	0.000	842875	395591	0.000	566137.75	661975.30
CIRA GL - CIRA KW-GL	1.000	112119	116838	0.000	144180	157387	1.000	23530	24557	0.000	527489	395591	0.000	582159.52	661975.30
Categories	**Macedonian**			**Romanian**			**Russian**			**Spanish**			**Turkish**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
WP rest - CIRA KW	0.000	39209	31831	0.000	155008	168851	1.000	563056	510806	1.000	522153	555329	0.000	102895	102895
WP rest - CIRA rest	0.000	39209	53286	0.000	155008	225275	0.387	563056	868658	0.000	522153	756581	0.000	136479	136479
WP rest - CIRA GL	0.000	39209	60132	0.000	155008	146937	0.000	563056	522063	0.000	522153	569025	0.000	102895	114674
WP rest - CIRA KW-GL	0.000	39209	48568	0.000	155008	193091	0.000	563056	508201	0.000	522153	519039	0.000	136479	127348
CIRA KW - CIRA rest	0.180	31831	53286	0.000	168851	225275	0.010	510806	868658	0.000	555329	756581	0.000	102895	136479
CIRA KW - CIRA GL	0.000	31831	60132	0.000	168851	146937	0.000	510806	522063	0.000	555329	569025	0.122	102895	114674
CIRA KW - CIRA KW-GL	0.000	31831	48568	0.000	168851	193091	0.000	510806	508201	0.000	555329	519039	0.000	102895	127348
CIRA rest - CIRA GL	1.000	53286	60132	0.000	225275	146937	0.000	868658	522063	0.000	756581	569025	1.000	136479	114674
CIRA rest - CIRA KW-GL	0.037	53286	48568	0.000	225275	193091	0.000	868658	508201	0.000	756581	519039	0.000	136479	127348
CIRA GL - CIRA KW-GL	0.000	60132	48568	0.000	146937	193091	0.000	522063	508201	0.000	756581	519039	0.688	114674	127348

3.6 Proportion of Participation in CIRA

Table 42. Editor proportion of edits in CIRA by edit buckets. *Values correspond to Kruskal-Wallis test mean ranks for each group and p-values for the pairwise comparison among groups according to Dunn's test with Bonferroni correction.*

Edits % in CIRA	**Arabic**			**Basque**			**Catalan**			**English**			**German**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
(1-100) - (101-1000)	0.000	78686	97158	0.000	1467	1989	0.000	17669	22659	0.000	3454520	3534121	0.000	338700	379001
(1-100) - (1001-5000)	0.000	78686	99934	0.000	1467	2069	0.000	17669	23254	0.000	3454520	3512360	0.000	338700	377682
(1-100) - (5001-10000)	0.000	78686	99295	0.231	1467	2043	0.000	17669	23805	1.000	3454520	3461726	0.000	338700	373469
(1-100) - (10001+)	0.000	78686	99588	0.007	1467	2063	0.000	17669	23161	1.000	3454520	3428261	0.000	338700	370382
(101-1000) - (1001-5000)	1.000	97158	99934	1.000	1989	2069	1.000	22659	23254	0.970	3534121	3512360	1.000	379001	377682
(101-1000) - (5001-10000)	1.000	97158	99295	1.000	1989	2043	1.000	22659	23805	0.079	3534121	3461726	1.000	379001	373469
(101-1000) - (10001+)	1.000	97158	99588	1.000	1989	2063	1.000	22659	23161	0.000	3534121	3428261	0.554	379001	370382
(1001-5000) - (5001-10000)	1.000	99934	99295	1.000	2069	2043	1.000	23254	23805	0.857	3512360	3461726	1.000	377682	373469
(1001-5000) - (10001+)	1.000	99934	99588	1.000	2069	2063	1.000	23254	23161	0.016	3512360	3428261	1.000	377682	370382
(5001-10000) - (10001+)	1.000	99295	99588	1.000	2043	2063	1.000	23805	23161	1.000	3461726	3428261	1.000	373469	370382
Edits % in CIRA	**Hebrew**			**Hungarian**			**Icelandic**			**Italian**			**Japanese**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
(1-100) - (101-1000)	0.000	36568	41426	0.000	32681	40990	0.000	1508	1920	0.000	97842	116042	0.000	148086	121343
(1-100) - (1001-5000)	0.000	36568	41521	0.000	32681	43322	0.003	1508	2036	0.000	97842	117010	0.000	148086	116163
(1-100) - (5001-10000)	0.012	36568	41909	0.000	32681	43616	1.000	1508	2032	0.000	97842	117176	0.000	148086	113346
(1-100) - (10001+)	0.000	36568	41850	0.000	32681	43800	0.105	1508	2026	0.000	97842	114259	0.025	148086	111387
(101-1000) - (1001-5000)	1.000	41426	41521	0.094	40990	43322	1.000	1920	2036	1.000	116042	117010	0.000	121343	116163
(101-1000) - (5001-10000)	1.000	41426	41909	0.800	40990	43616	1.000	1920	2032	1.000	116042	117176	0.316	121343	113346
(101-1000) - (10001+)	1.000	41426	41850	0.231	40990	43800	1.000	1920	2026	1.000	116042	114259	0.099	121343	111387
(1001-5000) - (5001-10000)	1.000	41521	41909	1.000	43322	43616	1.000	2036	2032	1.000	117010	117176	1.000	116163	113346
(1001-5000) - (10001+)	1.000	41521	41850	1.000	43322	43800	1.000	2036	2026	1.000	117010	114259	1.000	116163	111387
(5001-10000) - (10001+)	1.000	41909	41850	1.000	43616	43800	1.000	2032	2026	1.000	117176	114259	1.000	113346	111387
Edits % in CIRA	**Macedonian**			**Romanian**			**Russian**			**Spanish**			**Turkish**		
pairwise comparisons	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2	p-value	mean ranks 1	mean ranks 2
(1-100) - (101-1000)	0.000	2309	2983	0.000	148086	121343	0.000	156047	201813	0.000	326968	409700	0.000	70657	83306
(1-100) - (1001-5000)	0.000	2309	3331	0.000	148086	116163	0.000	156047	209502	0.000	326968	417157	0.000	70657	82590
(1-100) - (5001-10000)	0.087	2309	3308	0.000	148086	113346	0.000	156047	209430	0.000	326968	414006	0.064	70657	81462
(1-100) - (10001+)	0.004	2309	3352	0.000	148086	111387	0.000	156047	207998	0.000	326968	409671	0.019	70657	80965
(101-1000) - (1001-5000)	1.000	2983	3331	0.025	121343	116163	0.000	201813	209502	0.472	409700	417157	1.000	83306	82590
(101-1000) - (5001-10000)	1.000	2983	3308	0.316	121343	113346	0.188	201813	209430	1.000	409700	414006	1.000	83306	81462
(101-1000) - (10001+)	1.000	2983	3352	0.099	121343	111387	0.305	201813	207998	1.000	409700	409671	1.000	83306	80965
(1001-5000) - (5001-10000)	1.000	3331	3308	1.000	116163	113346	1.000	209502	209430	1.000	417157	414006	1.000	82590	81462
(1001-5000) - (10001+)	1.000	3331	3352	1.000	116163	111387	1.000	209502	207998	1.000	417157	409671	1.000	82590	80965
(5001-10000) - (10001+)	1.000	3308	3352	1.000	113346	111387	1.000	209430	207998	1.000	414006	409671	1.000	81462	80965