Cultural Configuration of Wikipedia: Measuring Autoreferentiality in Different Languages

Marc Miquel Ribé
Universitat Politècnica de Catalunya
mmiquel@lsi.upc.edu

Horacio Rodríguez
Universitat Politècnica de Catalunya
horacio@lsi.upc.edu

Abstract

Among the motivations to write in Wikipedia given by the current literature there is often coincidence, but none of the studies presents the hypothesis of contributing for the visibility of the own national or language related content. Similar to topical coverage studies, we outline a method which allows collecting the articles of this content, to later analyse them in several dimensions. To prove its universality, the tests are repeated for up to twenty language editions of Wikipedia. Finally, through the best indicators from each dimension we obtain an index which represents the degree of *autoreferentiality* of the encyclopedia. Last, we point out the impact of this fact and the risk of not considering its existence in the design of applications based on user generated content.

1 Introduction

"Wikipedia is a free web-based, collaborative, multilingual encyclopedia project supported by the non-profit Wikimedia Foundation", this is the way Wikipedia (WP) is defined in the starting article of the English language edition. What it does not say is that it is the seventh most visited webpage in the Internet and sixteen million articles prove its participation success. It requires a very complex governance system and one of its requisites and rule for achieving the goal of gathering all the human knowledge is maintaining the neutral point of view (NPOV) in its articles.

The repository implements the *wiki technology*, which applies to the ease in creating or modifying text collaboratively as well as the property of linking words to other articles. Due to this differentiated characteristic which enhances the navigation through the content and also for being the focus of attention, WP becomes a highly studied object whose nature is social and technical – textual, relational and quantitative (Ortega et al., 2007) – and is often analyzed by means of disciplines like Data Mining, Information Retrieval or Natural Language Processing.

Although WP maintains its goal and main rules in the almost three hundred language editions in which it is available, the English one is by far the biggest in number of articles. Every WP community decides on which articles are a priority to create, organizes in what is called wikiprojects and ultimately writes the text. Both users and creators of a language edition share a common cultural background and specificities in the writing style. However, when studies approach the community in terms of motivation they coincide they do it for fun, for appeal of the ideology or some sort of altruism (Nov, 2007). However, some informal surveys in Catalan WP association 'Amical Viquipèdia' showed how the national topics were a focus of interest for writing and conflict. Could it not be then that some editors get involved due to some sort of cultural motivation related to their own national or linguistic sphere too?

Yet in WP ideology there is no reason for this to occur, this content exists in any language edition. *Autoreferentiality* concept we propose stands out to describe the interest of a culture on itself, which in WP translates to the interest of editors for their own local content in a WP language edition. Our study makes two contributions: first, we show empirically how by an algorithm using the relations among categories and articles it is possible to retrieve a kind of content which is local to a language; second, how by the use of all kinds of WP features we can understand the importance of this content. We present this theoretical and practical work which will be extended to 20 languages in order to see if its results can be generalized and to give a stronger validity than studies limited to the English language edition.

Proceedings of Recent Advances in Natural Language Processing, pages 316–322,
Hissar, Bulgaria, 12-14 September 2011.

2 Related Work

There has been research on WP regarding many different aspects, but just a few on cultural questions. Pfeil et al. (2006) in their study proved how different behaviors in editing can be related to the culture. Other study from Hecht and Gergle (2010) focused on the differences in concepts common to several languages using Explicit Semantic Analysis by Gabrilovich et al. (2007).

In the context of topical coverage, studies like Kittur et al. (2009) quantify the content and classify the WP articles into general topics. The study showed a big amount of content related to the social sciences sphere and thus more culturally sensitive. However, the closest work on cultural content related problematic has been presented by Hecht (2009), who introduced the concept *self-focus bias* as *"occurring when contributors to a knowledge repository encode information that is important and correct to them and a large proportion of contributors to the same repository, but not important and correct to contributors of similar repositories"*. While he remarked this lack of consensus in theory, his implementation took the geographically located articles shared among languages to see its prominence by the number of incoming links each article had. As such, Hecht's study could make us understand how for each language edition the geographically located articles in their speaking territories were more important to their editors than other geographically labeled articles. However, it is left to be answered the problematic for many other kinds of content which can be included in the definition. Also, it did not compare strictly the existence of a particular content in different language editions since it assumed only those articles which were in available in different languages and then were universal.

In the following pages we want to introduce a different approach to the self-focus or autoreferentiality question. We explain how we relate it closely to the WP object characteristics and how from them we can understand the importance attributed to some information.

3 Approach

We introduce two stages in which we identify and measure autoreferentiality. First, by collecting all the articles which are likely to be included in a local content representative set, then obtaining their features and giving value in relation to the whole language edition articles. For this, we used a tool called wikAPIdia, which counts with multilingual compatibility and is Java and MySQL based. Differently than many systems using WP as knowledge source and limit themselves to the last articles, we used complementary material as history edits for our purpose.

3.1 Measuring Autoreferentiality

Autoreferentiality shows the degree by which a higher interest on local content is manifested in a language edition. An article is the indivisible unit of analysis within its features. We assume that a higher value in some features represents a higher interest, which in different set of articles can be compared by their average values. The features can be considered as interest indicators and grouped in different dimensions which illustrate the WP object. We will divide the analysis in seven dimensions: Semantic, Isolation, Effort, Prominence, Endogamy, Edition and Temporal. The first refers to the selection of articles, *Semantic* (1), takes into account their semantic value and will be extended on the next section.

Following, the other dimensions are about article qualities or the activity by which they are created. *Isolation* (2) explains if an article exists in other language editions and it is checked on the use of Interwiki links[1]. Hence, if there is external interest for a particular concept (which we assume lower for local content), it will be related to the number of this kind of links. *Effort* (3) is quantitative as it is measured by two indicators made out of the amount of bytes and outlinks – links which appear on the text and point to other articles. *Prominence* (4) complements measuring the number of inlinks, IL, the number of category memberships of an article, CM, and the PageRank (PR) value an article has. *Endogamy* (5) wants to know how prominent is the local content within itself, first by measuring the number of inlinks directed to the set which come from the same set, EIL, and second by measuring the number of category memberships of selected articles which already belong to the local content selection, ECM. *Edition* (6) is similar to second but represents a higher interest in number of edits, ED, number of editors and what we call a diversity coefficient. This calculated indicator is the number of editors, EDT, which are necessary to fulfill a high percentage of edits (for instance, we chose 80%) in relation to all the editors which contributed at least once to an article. The higher the coefficient the more diversified is the

[1] Interwiki links are those from one wiki to another.

editing. We assume it lower since there are highly motivated users by editing local content.

Lastly, *Temporal* (7) dimension is defined by the rate of article as indicator. First, comparing the relative values obtained by the rate of articles created in the selected set of articles, RR, and those created in all language edition, IRR. The hypothesis is that the local content will show higher relative rate. Second, looking at the subtraction of these relative values according to the periods and observing if the local content starts to grow or decay earlier than the general trend.

All in all, our end goal is merging the values of the optimal indicators in one single index which helps in comparing WP language editions (l). Therefore we will obtain the indicator from the feature using the next formula which subtracts the average of a feature (f) on the set by the average of all language edition and relates to this last one. The *Isolation* dimension interwiki links and the Edition dimension diversity coefficient will assume the opposite subtraction. It is expected that both average of features will be lower for the selected set of articles than for all language edition articles. The two endogamy indicators will calculate their value by considering the percentage of inlinks/category memberships to the set coming from the set (endo-inlinks and endo-category memberships), then subtracting 50 (minimum for endogamy) and relating it to 50 again as a range of significant data.

$$(1)\quad Ind.Value(l, f) = \frac{avg_{\forall articles \in Selection}(f_i) - avg_{\forall articles \in L.edition}(f_i)}{avg_{\forall articles \in L.edition}(f_i)}$$

Once we have an indicator value for all the language editions we can create an average value of them. This will explain how representative an indicator is and will work as a fair weighting in the index creation.

$$(2)\quad Ind.Weighting(f) = avg_{\forall l \in languages}(Ind.Value(f, l))$$

Then a partial index value is the multiplication of an indicator for a language, the general weighting and the percentage of the local content to all the articles from a language edition. The final index value will be the sum of all the partial values.

$$(3)\quad PartialIdx.Value(l) = \sum_{f=1}^{m}(Ind.Value(l, f) \cdot Ind.Weighting(f) \cdot setpercentage(l))$$

3.2 Selection of Articles

The selection of the twenty languages from all five continents represent a variety in both sociological use, spread in their respective territories and community activity in WP, in number of articles and users' involvement[2]. Hence we consider these factors independent enough from results.

Local content will be heterogeneous in any language. It can include writers and geographic places, music and historical objects. We understand it is relative to the language, to the people who are native writers of the language and to the territory where it is spoken, its legacy and activities. Nastase and Strube (2008) studied the titles of articles and categories and found how relevant they were for propagating semantic relations.

Our method of gathering the local content uses first a retrieval of articles and categories which include certain keywords in their titles, to later crawl the category memberships iteratively. If an article can be reached through two different paths it just appears once. From level zero (the one which includes the keywords) to level three, the content is tightly related to the keywords. Although usually there is seven to ten levels, after the third there appear some interferences with articles which can hardly be considered.

For instance, in a language like Catalan we might use the words which refer to the Catalan speaking territories, their demonym and language names (if the same language has more than one). These would be "catalunya", "català", but also "valencia" or "mallorquí" and would retrieve titles in articles and categories like "escriptors de catalunya" or "dret català", referring to writers and law. Then, any article which hangs from these two categories may specialize in some concepts or aspects and develop the topic.

4 Results

In this study, first we determine whether the scope of the local content in a WP language edition. If the selection process using keywords collected a great amount of articles this may infer later in a great autoreferentiality. In Table 1 we see the number of articles in January 2011 for each language edition and the selected percent-

[2] English has not been considered due its size and difficulties in processing in all dimensions.

age. There is no relation between the size of the language and the scope of local content. Small language editions like Icelandic or Swahili do not have higher percentage than big ones like Italian or Dutch, although these last have more articles in local content. Their values oscillate between 14,08% and 52,06% (mean 24,89%).

Languages	Nº Art. Lang. Edition	Selected. %
Arabic	134253	23,41
Catalan	301304	14,08
Chinese	334175	25,57
Czech	184251	25,65
Danish	141767	31,00
Dutch	650733	14,82
Finnish	261678	21,29
Guarani	1371	38,37
Hebrew	114496	27,73
Hungarian	182467	23,53
Indonesian	149509	12,19
Icelandic	42023	24,83
Italian	777906	14,83
Japanese	737085	52,06
Korean	155256	26,35
Norwegian	290629	19,63
Romanian	155763	31,01
Swahili	21193	23,88
Swedish	382801	28,01
Turkish	155242	19,56

Table 1. Extension of local content

In Table 2, we can see the average of the selected articles is up to three times smaller than that of the whole language edition articles. In the last column, the indicator value is made from the difference between both averages (formula 1), related to the one from all language articles. It is not important the selected set of articles has a low average if the average of all the language edition articles is low too. *Isolation*, measured by the number of interwiki links, wants to prove a smaller external interest. Less interwiki links means the article is no replicated to many other languages. In Table 3, we see in the last row that the standard deviation applied to the average of the set is much higher than on the average of all language editions for interwiki links. This means that there are few articles which have a greater number of interwiki links than the average and these may be those which have interest in other language editions. These could be around emblematic locations, institutions or famous celebrities. The resulting weighting is a high value like 74,4 which proves a good for showing the difference between local content and other kinds.

Languages	Avg. Sel.Set	Avg. Lang.	Diff.	Ind.Val.
Arabic	3,1	7,7	4,6	59,8
Catalan	1,4	6,4	5,0	78,6
Chinese	1,4	5,8	4,4	75,7
Czech	1,7	8,3	6,6	79,1
Danish	2,5	9,0	6,5	71,8
Dutch	1,2	5,5	4,3	78,4
Finnish	1,0	8,0	7,0	87,4
Guarani	10,7	16,9	6,2	36,7
Hebrew	3,0	10,1	7,1	70,2
Hungarian	2,8	8,0	5,2	65,4
Indonesian	0,9	7,1	6,2	87,0
Icelandic	1,3	8,8	7,4	84,7
Italian	2,5	4,9	2,4	49,5
Japanese	0,7	3,7	3,0	80,0
Korean	1,2	8,1	6,9	85,4
Norwegian	1,0	6,3	5,3	84,2
Romanian	1,4	7,9	6,5	82,6
Swahili	2,9	14,6	4,4	80,2
Swedish	1,2	6,4	5,2	81,7
Turkish	2,2	7,5	5,3	70,7

Table 2. Results for Isolation indicator

The procedure is repeated for other dimensions like *Effort*, represented by bytes, B, and Outlinks, OL. Both of them resulted in positive indicator weightings, although they are not fully confirmed as positive indicator for all cases. Our assumption was that a higher interest in local content would be reflected in longer articles and more linked towards other articles, which is just partially confirmed. *Prominence*, shows how only category membership's indicator is positive in all cases. It is proved that articles from the selected set are better socially annotated for all tested language, which results in a good weighting indicator of value 42,73. Other indicators from the dimension like number of inlinks and PageRank are irregular and like those from dimension *Effort* it cannot be concluded the local content represents a relational interest to define the whole encyclopedia. Again, the standard deviation shows us there is more variation in the selected set than in all language edition articles.

Dimensions	Isolat.	Effort		Prominence			Endogamy		Edition			Temporal	
Languages	**IW**	**B**	**OL**	**IL**	**CM**	**PR**	**EIL**	**ECM**	**ED**	**EDT**	**DC**	**RR**	**IRR**
Arabic	59,8	11,3	22,3	21,3	19,6	-22,10	21,20	31,70	-33,10	5,50	11,00	-24,62	-32,31
Catalan	78,6	-18,5	-15	-43,7	52,3	-26,30	35,30	63,10	16,90	1,30	9,00	-18,64	-10,17
Chinese	75,7	-5,8	33,7	5,7	54,2	20,60	40,90	60,30	27,10	19,20	3,70	-9,23	63,08
Czech	79,1	-8,7	-4,1	-33,9	27,5	-10,70	51,90	29,40	-25,00	7,00	5,40	-12,31	-33,85
Danish	71,8	-9,5	11,2	-23,1	36,5	-19,00	47,90	90,10	-8,90	-9,40	0,50	-15,38	-50,77
Dutch	78,4	24,9	36,3	2,4	43,6	85,50	43,00	55,30	-55,80	-35,40	29,00	-20,00	-72,31
Finnish	87,4	-3,6	5,4	-23,1	13	4,50	53,00	37,80	-42,40	-14,90	8,90	-12,31	-49,23
Guarani	36,7	15,5	69,3	34,3	14,3	6,50	51,80	90,80	-37,90	-28,50	-11,10	-41,54	-64,62
Hebrew	70,2	8,8	26	-18	43,9	-21,10	54,10	61,80	-43,10	-25,10	-4,70	-24,62	-40,00
Hungarian	65,4	-4,8	12	-32,6	43,3	31,70	40,00	40,00	-56,60	-2,10	42,00	-16,92	60,00
Indonesian	87	26,2	52,7	52,5	103,6	56,30	11,00	53,80	22,50	65,90	-9,90	-21,54	-15,38
Icelandic	84,7	35,4	10,9	-22,8	61,3	-6,80	50,00	82,40	161,20	275,70	-19,00	-18,46	-38,46
Italian	49,5	55	69,7	23,3	72,5	2,10	25,70	57,80	90,80	64,20	5,80	-15,25	13,56
Japanese	80	-1,7	16,6	-9,5	20,4	69,70	70,50	41,20	-59,40	-45,80	14,10	16,92	-58,46
Korean	85,4	2,1	43,6	-5,7	50,4	-22,80	64,60	34,00	-25,50	23,10	0,00	-15,38	-29,23
Norwegian	84,2	-8,5	6,7	-33,2	47,1	29,30	24,20	11,60	-20,70	24,40	8,20	-20,00	-46,15
Romanian	82,6	-3,1	0,3	-26,5	33,7	-39,90	64,50	40,70	-19,60	-30,80	18,40	-30,77	-67,69
Swahili	80,2	-24,9	9,8	-17,2	20,4	-64,70	76,10	39,00	110,70	289,90	45,80	-23,08	-41,54
Swedish	81,7	-3,9	1,3	-22,1	26,7	108,30	56,40	8,70	-28,90	-15,00	11,90	-10,77	-40,00
Turkish	70,7	-1,4	16	-12,9	70,2	-44,4	23,7	40,1	42,6	37,3	2,40	-27,69	-12,31
Weighting	74,46	4,24	21,24	-9,24	42,73	6,84	45,29	48,48	0,74	30,33	8,57	-18,08	-28,29
S.D.(Ind.Val.)	14,05	17,68	20,79	22,29	23,37	39,73	22,59	26,53	52,48	80,77	14,49	30,15	45,8
S.D.(AvgSet)	1,28	0,35	0,46	0,42	0,29	29,58	0,53	0,24	0,72	0,51	0,14		
S.D(AvgLEdit)	0,41	0,27	0,42	0,42	0,3	24,52			0,32	0,31	0,07		

[3]

Table 3. All indicators values

Those levels which are closer to the zero (containing the keywords in the title) accumulate more effort and are more prominent because they are more general and often inlinked by the specialized ones in the following levels.

In *Endogamy*, both indicators are fulfilled showing how the selected content represents a semantic unity around the keywords. The special procedure for this case implied that endogamy means at least half of the inlinks coming from the same set and then percentage surpassing 50 related to the 50 as a range. With the high value of the indicator tested in inlinks, the local content proved to be defined having a common set of terms which were the core of the selected set. With category memberships it showed how these articles are often classified in several categories which are different but semantically close. For instance, an eminent personality is categorized by his profession but also the city where was born and political positions.

Edition indicators ED or EDT are not positive for all cases. Equally to others, there is almost twice variation in the selected articles than in all articles, which means local content can raise interest in the community but not all the degrees of specialization of the topic receive the same. When the standard deviation is calculated for the indicator values on all languages they give a very high variation which means the communities' responses to this content are very different. The other indicator, diversity coefficient, does not give positive for all cases but it is more stable in its values. It also reflects a tendency of few editors writing the biggest amount of the articles even more emphasized.

From last dimension, *Temporal*, we can conclude the assumption that the article creation in local content would show more interest in time is false. Although the rates show how local content is mostly created while there is a good period of creation for the whole language edition, the relative amount created is not higher for the local content than for the whole language edition. In short, local content is mostly characterized by having few interwiki links and being highly cat-

[3] IW: interwiki links, B: bytes, OL: outlinks, IL:inlinks, CM:category memberships, PR: PageRank, EIL: endogamy inlinks, ECM: endogamy category memberships, ED: edits, EDT: editors, DC: diversity coefficient, RR: relative rate, IRR: increment relative rate. S.D: standard deviation.

egorized. These are the two indicators which can express better the difference of the selected set to all the articles from the language edition. These two represent first an interest not corresponded to other language editions and then a higher will of having it well classified. Endogamy indicators also proved how this content is around the same topic despite it is heterogeneous and can be classified in many other categories like those used by Kittur et al. (2009). When looking at the standard deviation of all the indicator weightings we see how the most stable is diversity coefficient followed by Interwiki links.

With all the indicators already measured and evaluated, the last step is creating the index. Yet, we have another constraint besides having a positive value in the weighting, which is not being correlated among them and therefore avoid redundancy. We checked all the indicators for three different size language editions (Italian, Czech and Romanian) and saw four different correlations: bytes with outlinks, inlinks with endo-inlinks, category memberships with category memberships from set and number of edits with number of editors. Then we select first those which are most independent and from the couples those with higher weighting value. These are interwiki links (*Isolation*), bytes (*Effort*), category memberships (*Prominence*), inlinks from set (*Endogamy*), number of editors and diversity coefficient (*Edition*). In Table 4 we can see the ranking of the overall index.

Languages	Index Value	Position
Icelandic	48,71	1
Japanese	47,41	2
Swahili	46,58	3
Korean	34,43	4
Romanian	30,21	5
Danish	28,01	6
Swedish	26,98	7
Hebrew	25,82	8
Czech	24,60	9
Guarani	23,80	10
Hungarian	21,36	11
Turkish	21,17	12
Norwegian	20,27	13
Finnish	19,60	14
Indonesian	17,59	15
Italian	16,94	16
Arabic	16,33	17
Chinese	16,26	18
Dutch	14,21	19
Catalan	13,35	20

Table 4. Overall results Autoreferentiaity index

5 Discussion

Usually, motivation was approached by classic social sciences methodologies which discuss about where it resides, in the individual by itself or in it while is acting. Further than that, an analysis on the content cannot provide a clear answer on motivation but it can explain what are the cultural preferences and in which degree. While most of the research assumes the results obtained from English language as valid for all language editions, this study remarks how differences exist, they are important to those who create the product, and furthermore they finally shapes the encyclopedia in several dimensions. In the initial selection of articles which represent the local content we found that the extension it covered from the encyclopedia had nothing to do with the sociological characteristics from the community of speakers neither the one involved in WP. But regardless the size of the WP language edition, a non-negligible percentage covered almost a quarter of the total articles.

That said, any of the dimensions we proposed cover different aspects of WP's articles information. What is interesting is that while they vary in number of bytes, they vary less in number of editors and there is a subgroup much more active. This is the confirmation editors change their habits of editing depending on the content they are about to write.

All in all, those indicators which proved more consistent for all languages and their selected articles are the interwiki links and the category memberships, followed by the two from the endogamy inlinks and category memberships. It is paradigmatic that the first, which represented the lack of interest in other languages and was very intrinsic to the definition of autoreferentiality, was also the one with higher value and less variation among the language editions. The second one, showed how in the social annotation process of creating content in articles and structuring it in categories, editors prefer local content to be more precise to all the sorts of content in which can belong. This is important for the future semantic web in which the information must be tagged. And the third, related to endogamy, show how this content shares a sense of unity. No matter how heterogeneous are the articles in discourse or general topic that when they are sorted in categories, on the descendent way from those which include the keywords, they will include some pieces of text (and therefore links) which will tend to refer to themselves. Also, one of the cor-

relations we noticed was that the more endogamy in terms of inlinks, the less interwiki links it had. In other words, the less permeated is a culture by other topics and then diverse, the less connections from abroad.

6 Conclusions and future lines

In this study, first we determined with a simple technique method the scope of the local content in WP language editions, which is in average a 24%. Choosing key words which are very tight to each language like the territories where they are spoken proved right to obtain local content, although a good choice of key words like the territory names and gentilics from the language edition was key to avoid losing content. Most of content comes from the main territory name. While this selection could have been influenced by the noisy category structure, studying after the category memberships as a feature of the content and discovering local content has more categories memberships reinforced the method.

Our results according to our methodology for creating an index showed that autoreferentiality value can increase due to several dimensions. Languages like Japanese and Icelandic gave a high and similar final value but the first relied more on the isolation of their content and their endogamy and the second had a much higher number of editors interested in contributing to local content articles. Since there is no direct relation between features, the extension of the local content and autoreferentiality, every community and its composition must be studied as a different case. For instance, any insight on the general trends the features can show like the length of articles or the very active subgroups of users could be related to a qualitative study which would explain much better motivation works and the social interactions.

To conclude, we want to remark how important understanding autoreferentiality can be when designing applications which retrieve information from WP or another user generated repository. The confirmation of an interest from users in a content in which they identify and develop might not necessarily be considered a bias. While the encyclopedia goal remains in the vague 'collecting all the human knowledge', local content exists part of this collection and because the editors spontaneously created it. Any software which applies to retrieve information from WP or any dataset might be designed aware of giving a better context. Once our best conclusion is the uniqueness of some content in any language, our future work will be on understanding how cultural configuration can be explained by particular topics.

Acknowledgments

This work has been partially funded by KNOW2 (TIN2009-14715-C04-04)

Eduard Aibar, Amical Viquipèdia, Joan Campàs, Marcos Faúndez. Diana Petri, Pere Tuset, Fina Ribé, Jordi Miquel, Joan Ribé, Peius Cotonat.

References

Gabrilovich, E. and Markovitch, S. (2007). *Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis*. 20th Joint Conference for A.I. (IJCAI '07), *1606-16*

Halavais, Alexander and Kaclkaff, Derek. 2008. *An analysis of topical coverage of Wikipedia. Journal of Computer-Mediated Communication.* 13(2)

Hecht, Brent and Gergle, Darren. 2009. *Measuring self-focus bias in community-maintained knowledge repositories.* In C38;T'09: Proc. of the 4th international conf. on Communities and technologies, 11-20, New York, NY, USA, 2009.

Hecht, Brent and Gergle, Darren. 2010. *The Tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context,* 291-300. ACM.

Kittur, Aniket and chi, Ed H. and Suh, Bongwon. 2009. *What's in Wikipedia?: mapping topics and conflict using socially annotated category structure.* CHI'09: Proceedings of the 27th international conference on Human factors in computing systems. pages 1509-1512. ACM. Boston, MA, USA.

Ortega, Felipe and Gonzalez Barahona, Jesus M.. 2007. *Quantitative analysis of the Wikipedia community of users.* WikiSym '07: Proceedings of the 2007 International symposium on Wikis. Pages 75-86. ACM. Montreal, Québec, Canada.

Nastase, Vivi and Strube, Michael. 2008. *Decoding Wikipedia categories for knowledge acquisition.* AAAI'08: Proceedings of the 23rd national conference on Artificial intelligence. Pages 1219-1224. AAI Press. Chicago, Illinois.

Nov, Oded. *What motivates Wikipedians?* 2007. Communic. ACM. 60-64. New York, NY, USA.

Pfeil, Ulrike and Zaphiris, Panayiotis and Ang, Chee S. 2006. *Cultural Differences in Collaborative Authoring of Wikipedia*. Journal of Computer-Mediated Communication. 12(1).

Yang, Heng-Li and Lai, Cheng-Yu. 2010. *Motivations of Wikipedia content contributors.* Computer Human Behaviour. 26(6).