View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Servicio de Difusión de la Creación Intelectual

On the Assessment of Information Quality in Spanish Wikipedia

Guido Urquiza[1], Matías Soria[1], Sebastián Perez Casseignau[1], Edgardo Ferretti[1,2], Sergio A. Gómez[3], and Marcelo Errecalde[1,2]

[1] Universidad Nacional de San Luis (UNSL), San Luis - Argentina
[2] Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (UNSL)
[3] Laboratorio de Investigación y Desarrollo en Inteligencia Artificial, Universidad Nacional del Sur (UNS), Bahía Blanca - Argentina
e-mails: {ferretti,merreca}@unsl.edu.ar, sag@cs.uns.edu.ar

Abstract. Featured Articles (FA) are considered to be the best articles that Wikipedia has to offer and in the last years, researchers have found interesting to analyze whether and how they can be distinguished from "ordinary" articles. Likewise, identifying what issues have to be enhanced or fixed in ordinary articles in order to improve their quality is a recent key research trend. Most of the approaches developed in these research trends have been proposed for the English Wikipedia. However, few efforts have been accomplished in Spanish Wikipedia, despite being Spanish, one of the most spoken languages in the world by native speakers. In this respect, we present a first breakdown of Spanish Wikipedia's quality flaw structure. Besides, we carry out a study to automatically assess information quality in Spanish Wikipedia, where FA identification is evaluated as a binary classification task. The results obtained show that FA identification can be performed with an F1 score of 0.81, using a document model consisting of only twenty six features and AdaBoosted C4.5 decision trees as classification algorithm.

Keywords: Wikipedia, Information Quality, Featured Article Identification, Quality Flaws Prediction

1 Introduction

The online encyclopedia Wikipedia is one of the largest and most popular user-generated knowledge sources on the Web. Considering the size and the dynamic nature of Wikipedia, a comprehensive manual quality assurance of information is infeasible. Information Quality (IQ) is a multi-dimensional concept and combines criteria such as accuracy, reliability and relevance. A widely accepted interpretation of IQ is the "fitness for use in a practical application" [1], i.e. the assessment of IQ requires the consideration of context and use case. Particularly, in Wikipedia the context is well-defined by the encyclopedic genre, that forms the ground for Wikipedia's IQ ideal, within the so-called *featured article criteria*.[4] Having a formal definition of what constitutes a high-quality article, i.e. a

[4] `http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria`

featured article (FA), is a key issue; however, as indicated in [2], in 2012 less than 0.1% of the English Wikipedia articles were labeled as featured. At present, this ratio still remains, since there are 4 785 featured articles out of 5 187 923 articles on the English Wikipedia.[5]

Information quality assessment in Wikipedia has become an ever-growing research line in the last years [3–10]. A variety of approaches to automatically assess quality in Wikipedia has been proposed in the relevant literature. According to our literature review, there are three main research lines related to IQ assessment in Wikipedia, namely: (i) featured articles identification [5, 6, 10]; (ii) quality flaws detection [7–9]; and (iii) development of quality measurement metrics [3, 4]. In this paper we will concentrate on the first two research trends mentioned above.

All the above-mentioned approaches have been proposed for the English Wikipedia, which ranks among the top ten most visited Web sites in the world.[6] With 1 265 961 articles, Spanish Wikipedia ranks ninth in the list after English, Swedish, Cebuano, German, Dutch, French, Russian and Italian languages. In spite of being one of the thirteen versions containing more than 1 000 000 articles,[7] and despite being Spanish one of the most spoken languages in the world by native speakers, few efforts have been made to assess IQ on Spanish Wikipedia. To the best of our knowledge, [11] and [12] are the most relevant works related to IQ in Spanish Wikipedia, and [12] can be characterized as belonging to the third main research trend mentioned above.

In [11], Pohn et al. presented the first study to automatically assess information quality in Spanish Wikipedia, where FA identification was evaluated as a binary classification task. The research question which guided their experiments was to verify if successful approaches for the English version, like word count [5] and style writing [6], also work for the Spanish version, and if not, what changes were needed to accomplish a successful identification. Results showed that when the discrimination threshold is properly set, the word count discrimination rule performs well for corpora where average lengths of FA and non-FA are dissimilar. Moreover, it was concluded that character tri-grams vectors are not as effective for the Spanish version as they are for FA discrimination in the English Wikipedia; but Bag-of-Words (BOW) and character n-grams with $n > 3$ performed better in general. This may be because in Spanish many kind of adverbs are fully encompassed in 4-grams or 5-grams. The best F1 scores achieved were 0.8 and 0.81, when SVM is used as classification algorithm, documents are represented with a binary codification, and 4-grams and BOW are used as features, respectively.

The contribution of our work is twofold. On one hand, we report results on FA identification evaluated as a binary classification task, like in [11], but where the document model used is composed of static features rather than dynamic features. On the other hand, this paper also targets the investigation of quality

[5] `https://en.wikipedia.org/wiki/Wikipedia:Featured_articles`
[6] Alexa Internet, Inc., `http://www.alexa.com/siteinfo/wikipedia.org`
[7] `http://meta.wikimedia.org/wiki/List_of_Wikipedias`

flaws. We have conducted an exploratory analysis similar to the original one proposed in [13], to reveal both the quality flaws that actually exist and the distribution of flaws in Spanish Wikipedia articles.

With this aim, in Sect. 2, we describe the experimental design and results obtained in the FA identification task. Then, Sect. 3 introduces the problem of predicting quality flaws in Wikipedia based on cleanup tags. Also, our findings are presented and discussed. Finally, Sect. 4 offers the conclusions.

2 Featured Articles Identification

Given the question: *is an article featured or not?* we have followed a binary classification approach where articles are modeled using a vector composed by twenty six features. All article features correspond to *content* and *structure* dimensions, as characterized by Anderka et al. [7]. We decided to implement these features based on the experimental results provided by Dalip et al. [14], which showed that the most important quality indicators are the easiest ones to extract, namely, textual features related to length, structure and style. The dataset used, was the one compiled in [11], which consists of two corpora, namely: a balanced corpus and an unbalanced corpus. It is worth noting that "balanced" means that FA and non-FA articles were selected with almost similar document lengths. In a similar manner, "unbalanced" refers to the fact that non-FA articles were randomly selected without considering their average lengths. Both corpora are balanced in the traditional sense, i.e. the positive (FA) and negative (non-FA) classes contain the same number of documents. In particular, the balanced corpus contains 714 articles in each category and the unbalanced one has 942 articles in each category as well. It is ensured that non-FA articles belonging to the balanced corpus have more than 800 words. The articles belong to the snapshot of the Spanish Wikipedia from 8th, July 2013.

Formally, given a set $A = \{a_1, a_2, \ldots, a_n\}$ of n articles, each article is represented by twenty six features $F = \{f_1, f_2, \ldots, f_{26}\}$. A vector representation for each article a_i in A is defined as $a_i = (v_1, v_2, \ldots, v_{26})$, where v_j is the value of feature f_j. A feature generally describes some quality indicator associated with an article. A few differ slightly from one another, e.g., counts divided by the number of characters instead of words or ratios instead of a pure count. Table 1 shows the features composing our document model; for specific implementations details cf. [15].

Given the characteristics of these features, content-based features were implemented with AWK and shell-script programming using as input the plain texts extracted from the Wikipedia articles. By using the same programming languages, but using as input the wikitexts of Wikipedia articles, structure-based features were calculated. It is worth mentioning that wikitexts are not provided in the corpora of Pohn et al. and they were extracted from the corresponding Wikipedia dump.[8] To perform the experiments we have used the WEKA Data

[8] The updated corpora, including the wikitexts, can be downloaded from: `https://dl.dropboxusercontent.com/u/35037977/Corpus.tar.gz`

Mining Software [16], including its SVM-wrapper for LIBSVM [17]. Notice that all the results discussed below are average values obtained by applying tenfold cross-validation.

2.1 Results

In the first place, we replicated the experimental setting of Pohn et al. [11], where Naive Bayes (NB) and Support Vector Machine (SVM) classification approaches were evaluated, for both corpora. For the unbalanced corpus, the best F1 scores achieved by Pohn et al. were 0.91 and 0.94, for NB and SVM, respectively. In both cases, character 4-grams were used as features (with full vocabulary size) in a binary document model (*bnn* codification from the SMART nomenclature [18]). In particular, the C parameter of SVM was set to 32, after experimentally deriving its value ranging in the set $\{2^{-5}, 2^{-3}, 2^{-1}, \ldots, 2^{13}, 2^{15}\}$. For a linear kernel like the one used by Pohn et al., we achieved an F1 score of 0.92, and F1 = 0.94 was achieved by using an RBF kernel with $C = 2^9$ and $\gamma = 2^{-3}$; a configuration similar to the linear kernel given that γ value is close to zero. For the NB classifier, the F1 score obtained was 0.9. Hence, as it can be observed, the performance of both proposals are similar.

For the balanced corpus, a more challenging setting, the F1 scores reported by Pohn et al. for NB classifier were below 0.78 and the best F1 scores achieved were 0.8 and 0.81, for the SVM classifier with full and reduced vocabulary, respectively, using a binary document model. In our experiments, NB performed notably worse than in [11], given that this classifier achieved an F1 = 0.62. For SVM, the best F1 score achieved was 0.78, with an RBF kernel with parameters set to $C = 2^{11}$ and $\gamma = 2^{-3}$, respectively. As usual, these parameters were experimentally derived by a grid-search in the ranges $C \in \{2^{-5}, 2^{-3}, 2^{-1}, \ldots, 2^{13}, 2^{15}\}$ and $\gamma \in \{2^{-15}, 2^{-13}, 2^{-11}, \ldots, 2^{1}, 2^{3}\}$. Different configurations of polynomial kernels were also evaluated (with $d \in \{2, 3, 4, 5\}$ and $r \in \{0, 1\}$) but no better results were obtained than 0.78. It is well known that increasing γ and d parameters from the RBF and polynomial kernels allow for a more flexible decision boundary, but if they are increased too much, this might yield in principle an over-fitting of the model and hence obtaining a poor capability of generalization of the classifier.

Besides, we also evaluated other classification approach — Ada-boosted C4.5 decision trees. This approach has been used before in the context of Wikipedia IQ, but for quality flaws prediction task [19]. Using unpruned trees and one hundred boosting rounds achieved an F1 score of 0.81. This meta-algorithm, was also run with Pohn et al. document models and the performance achieved was F1 = 0.8. As it can be observed, both approaches have quite alike performances, with both classification methods. We believe that the advantage of our feature-engineering approach relies on the fact of having a fixed-size document model, that with only 26 features has a performance comparable to dynamic document models with thousand of features. This is not a minor issue, since having a classifier in a productive environment (like a Wikipedia bot[9]), also implies being able of computing document models efficiently, as in our case.

[9] `https://en.wikipedia.org/wiki/Wikipedia:Bots`

Table 1. Features which comprise the document model.

Feature	Description
Content-based	
Character count	Number of characters in the plain text, without spaces
Word count	Number of words in the plain text
Sentence count	Number of sentences in the plain text
Word length	Average word length in characters
Sentence length	Average sentence length in words
Paragraph count	Number of paragraphs
Paragraph length	Average paragraph length in sentences
Longest word length	Length in characters of the longest word
Longest sentence length	Number of words in the longest sentence
Shortest sentence length	Number of words in the shortest sentence
Long sentence rate	Percentage of long sentences*
Short sentence rate	Percentage of short sentences⋆
Structure-based	
Section count	Number of sections
Subsection count	Number of subsections
Heading count	Number of sections, subsections and subsubsections
Section nesting	Average number of subsections per section
Subsection nesting	Average number of subsubsections per subsection
Lead length	Number of words in the lead section†
Lead rate	Percentage of words in the lead section
Image count	Number of images
Image rate	Ratio of image count to section count
Link rate	Percentage of links‡
Table count	Number of tables
Reference count	Number of all references using the `<ref>...</ref>` syntax (including citations and footnotes)
Reference section rate	Ratio of reference count to the accumulated section subsection and subsubsection count
Reference word rate	Ratio of reference count to word count

* A long sentence is defined as containing at least 30 words.
⋆ A short sentence is defined as containing at most 15 words.
† A lead section is defined as the text before the first heading. Without a heading there is no lead section.
‡ Every occurrence of a link (introduced with two open square brackets) in the unfiltered article text is considered when computing the ratio of link count to word count in the plain text.

3 A Preliminary Breakdown of Quality Flaws

Despite the fact that FA identification is a useful task, assessing what kind of shortcomings of an article must be enhanced, would help writers to improve the article's quality. In this respect, cleanup tags are a means to tag flaws in Wikipedia. As shown in Fig. 1, they are used to inform readers and editors of specific problems with articles, sections, or certain text fragments. However, there is no single strategy to spot the entire set of all cleanup tags. Cleanup tags are realized based on templates, which are special Wikipedia pages that can be included into other pages.

Quality flaws prediction in Wikipedia was a research line started in 2011 by Anderka et al. [13] and evolved in seminal works like [2, 20, 21]. Particularly, in [2] an extensive exploratory analysis on Wikipedia's quality flaw structure is presented for the English version, whose approach consisted in creating a local copy of the Wikipedia database. Their results revealed that tagging work in Wikipedia mostly targets the encyclopedic content rather than pages used for content organization and user discussions. Based on this, we decided to use an alternative method, viz. a query retrieving approach on indexed documents with *Elasticsearch*, a search engine which provides scalable and real-time search.[10]

We hence introduced an extraction approach that consists of automatically querying the search engine with patterns representing maintenance templates.[11] These templates are organized into categories depending on the maintenance

Fig. 1. The Wikipedia article "Salto Base" (Base Jumping) with a cleanup tag indicating that certified references need to be included.

[10] https://www.elastic.co/

[11] https://es.wikipedia.org/wiki/Wikipedia:Plantillas_de_mantenimiento

task required, but not all maintenance templates necessarily imply a quality flaw. For example, *notification* templates are used to inform Wikipedians to proceed in agreement with the policies and conventions of Wikipedia. Similarly, *protection* templates warn Wikipedians that a particular working space has been blocked for its proper restoration by a librarian due to violations on the policies and conventions of Wikipedia. Likewise, according to our analysis, the remaining categories, namely: *critic maintenance*, *content*, *style*, *fusion* and *development*, do contain templates which can be associated with quality flaws, as shown in Table 2. It is worth noting that, as stated in this table, this preliminary breakdown of quality flaws has been carried out on a recent Wikipedia snapshot, the dump corresponding to April 2016.[12]

The first column of Table 2 specifies the category where templates (second column) associated with a particular flaw type (third column) are organized. The fourth column presents the number of articles that were found containing these particular templates. The flaw type scheme used corresponds to the one proposed by Anderka et al. [2, 13]. Figure 2 shows, from among the tagged articles, how flaw types are distributed. As it can be observed, *verifiability* is by far the most extended flaw type, corresponding to approximately 70% of the tagged content. This finding agrees with the results reported in [2, 13].

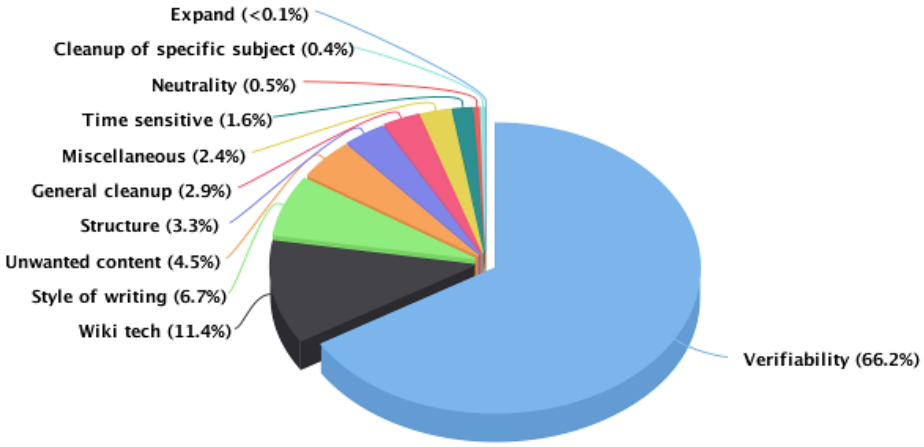

Fig. 2. Flaw types found and their distribution in the Spanish Wikipedia snapshot from April 2016. The percentages relate to the set of 111 072 tagged articles.

Besides, from Table 2, we can notice that the template *Referencias*[13] represents 90% of the articles tagged with the flaw concerning verifiability. That means that most of the articles suffer from this flaw because they contain neither references nor footnotes. From the remainder 10%, template *Referencias*

[12] https://dumps.wikimedia.org/eswiki/20160407/
[13] https://es.wikipedia.org/wiki/Plantilla:Referencias

Table 2. Flaw types breakdown for the Wikipedia snapshot corresponding to April 2016.

Category	Template	Flaw type	# articles
Content	Actualizar	Time sensitive	275
	CDI	Neutrality	4
	Complejo	Style of writing	107
	Desactualizado	Time sensitive	1515
	Discutido	Verifiability	610
	Documentación deficiente	Expand	18
	Ficticio	Cleanup of specific subject	464
	Fuentes no fiables	Verifiability	23
	Globalizar	Neutrality	92
	No neutralidad	Neutrality	341
	Problemas artículo	General cleanup	3218
	PVfan	Neutrality	60
	Referencias	Verifiability	66616
Critic maintenance	Artículo indirecto/esbozo	Wiki tech	7
	Bulo	Verifiability	701
	Contextualizar	Style of writing	162
	Fuente primaria	Verifiability	54
	Infraesbozo	Wiki tech	93
	Plagio	Unwanted content	27
	Posible copyvio	Verifiability	11
	Promocional	Unwanted content	141
	Sin relevancia	Verifiability	320
Development	Traducción	Miscellaneous	2612
Fusion	Fusión historiales	Unwanted content	19
	Fusionar	Unwanted content	2002
	Fusionar desde	Unwanted content	470
	Fusionar en	Unwanted content	740
	Posible fusionar	Unwanted content	73
Style	Categorizar	Wiki tech	99
	Copyedit	Style of writing	3641
	Excesivamente detallado	General cleanup	18
	Formato de cita	Wiki tech	546
	Huérfano	Wiki tech	223
	Identificador	Verifiability	1344
	Largo	Structure	3701
	Mal traducido	Style of writing	3082
	Mejorar redacción	Style of writing	113
	Publicidad	Unwanted content	1528
	Recentismo	Neutrality	5
	Referencias adicionales	Verifiability	3850
	Revisar traducción	Style of writing	342
	Traducción incompleta	Miscellaneous	8
	Wikificar	Wiki tech	11731
		Total over all types	**111072**

adicionales comprise almost 5% and template *Identificador* represents almost 2%. This means that existing references are not enough or are difficult to be found since particular key features are missing in the references, like the ISBN in a book. From Fig. 2, we can also see that *Wiki tech* flaw type, ranks second with 11.4%. In [2, 13], this flaw type also ranked second with approximately 19% and 16%, respectively. In a similar manner as occur with verifiability flaw and the *Referencias* template, in this case, 92% of the articles tagged with the *Wiki tech* flaw type, correspond to template *Wikificar*; indicating that these articles notoriously do not comply to Wikipedia's style manual. The remaining flaw types and their orderings, differ in [2, 13], as well as in our case; nonetheless, flaw types *Unwanted content*, *Style of writing* and *General cleanup*, are those having in general higher percentages after *Verifiability* and *Wiki tech*.

4 Conclusions

In this work, we have presented a first breakdown of Wikipedia's quality flaw structure for the Spanish language, following the pioneering approach of Anderka et al. [2, 13]. As reported in these works, *verifiability* related flaws comprise approximately 70% of tagged articles, like found in our study. Without doubts, this preliminary report paves the way for the development and evaluation of existing approaches to predict quality flaws by means of machine learning techniques, like in [8, 9, 19].

Besides, we carried out a study to automatically assess information quality, where FA identification was evaluated as a binary classification task. The results obtained showed that FA identification can be performed with an F1 score of 0.81, using a document model consisting of only twenty six features and AdaBoosted C4.5 decision trees as classification algorithm. These results were compared to previous results reported by Pohn et al. [11], who used dynamic document models with thousand of features, and both approaches have quite alike performances. In our view, the advantage of our feature-engineering approach relies on the fact of having a fixed-size document model which can be efficiently computed in a productive environment, like a Wikipedia bot.

Acknowledgments

This work has been partially founded by PROICO 30312, Universidad Nacional de San Luis, Argentina. Sergio A. Gómez is supported by Secretaría General de Ciencia y Técnica, Universidad Nacional del Sur, Argentina. The authors also thank to PROMINF (*Sub-proyecto "Desarrollo conjunto de sistema inteligente para la Web, con alumnos y docentes de las Licenciaturas en Cs. de la Computación de la UNS y la UNSL"*), Plan Plurianual 2013-2016, SPU.

References

1. Wang, R., Strong, D.: Beyond accuracy: what data quality means to data consumers. Journal of management information systems **12**(4) (1996) 5–33

2. Anderka, M., Stein, B.: A breakdown of quality flaws in Wikipedia. In: 2nd joint WICOW/AIRWeb workshop on Web quality (WebQuality'12), ACM (2012) 11–18
3. Lih, A.: Wikipedia as participatory journalism: reliable sources? Metrics for evaluating collaborative media as a news resource. In: Proceedings of the 5th international symposium on online journalism. (2004) 16–17
4. Stvilia, B., Twidale, M., Smith, L., Gasser, L.: Assessing information quality of a community-based encyclopedia. In: 10th Intl. Conf. on Information Quality. (2005)
5. Blumenstock, J.: Size matters: word count as a measure of quality on Wikipedia. In: 17th international conference on World Wide Web, ACM (2008) 1095–1096
6. Lipka, N., Stein, B.: Identifying featured articles in Wikipedia: writing style matters. In: 19th international conference on World Wide Web, ACM (2010) 1147–1148
7. Anderka, M., Stein, B., Lipka, N.: Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. In: 35rd annual international ACM SIGIR conference on research and development in information retrieval, ACM (2012)
8. Ferretti, E., Fusilier, D.H., Guzmán-Cabrera, R., y Gómez, M.M., Errecalde, M., Rosso, P.: On the use of PU learning for quality flaw prediction in wikipedia. In: CLEF (Online Working Notes/Labs/Workshop). (2012)
9. Ferretti, E., Errecalde, M., Anderka, M., Stein, B.: On the use of reliable-negatives selection strategies in the pu learning approach for quality flaws prediction in wikipedia. In: 11th Intl. Workshop on Text-based Information Retrieval. (2014)
10. Lex, E., Völske, M., Errecalde, M., Ferretti, E., Cagnina, L., Horn, C., Stein, B., Granitzer, M.: Measuring the quality of web content using factual information. In: 2nd joint WICOW/AIRWeb workshop on Web quality (WebQuality), ACM (2012)
11. Pohn, L., Ferretti, E., Errecalde, M.: Identifying featured articles in Spanish Wikipedia. In: Computer Science & Technology Series: XX Argentine Congress of Computer Science - selected papers. EDULP (2015) 171–182
12. Druck, G., Miklau, G., McCallum, A.: Learning to predict the quality of contributions to wikipedia. WikiAI **8** (2008) 7–12
13. Anderka, M., Stein, B., Lipka, N.: Towards Automatic Quality Assurance in Wikipedia. In: 20th intl. conference on World Wide Web, ACM (2011) 5–6
14. Dalip, D.H., Gonçalves, M.A., Cristo, M., Calado, P.: Automatic assessment of document quality in web collaborative digital libraries. Journal of Data and Information Quality **2**(3) (December 2011) 1–30
15. Fricke, C.: Featured article identification in wikipedia. Bachelor Thesis, Bauhaus-Universität Weimar (2012)
16. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explorations **11**(1) (2009)
17. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2** (2011) 27:1–27:27
18. Salton, G.: The Smart Retrieval System: Experiments in Automatic Document Processing. Prentice Hall (1971)
19. Ferschke, O., Gurevych, I., Rittberger., M.: FlawFinder: a modular system for predicting quality flaws in Wikipedia. In: Notebook papers of CLEF 2012 labs and workshops. (2012)
20. Anderka, M., Stein, B., Busse, M.: On the Evolution of Quality Flaws and the Effectiveness of Cleanup Tags in the English Wikipedia. In: Wikipedia Academy 2012, Wikipedia (July 2012)
21. Anderka, M.: Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. PhD thesis, Bauhaus-Universität Weimar (June 2013)