

# Characterizing and comparing Portuguese and English Wikipedia medicine-related articles

Gil Domingues

INESC TEC

Faculty of Engineering of the University of Porto

Porto, Portugal

gil.domingues@fe.up.pt

Carla Teixeira Lopes

INESC TEC

Faculty of Engineering of the University of Porto

Porto, Portugal

ctl@fe.up.pt

## ABSTRACT

Wikipedia is the largest on-line collaborative encyclopedia, containing information from a plethora of fields, including medicine. It has been shown that Wikipedia is one of the top visited sites by readers looking for information on this topic. The large reliance on Wikipedia for this type of information drives research towards the analysis of the quality of its articles. In this work, we evaluate and compare the quality of medicine-related articles in the English and Portuguese Wikipedia. For that we use metrics such as authority, completeness, complexity, informativeness, consistency, currency and volatility, and domain-specific measurements, in order to evaluate and compare the quality of medicine related articles in the English and Portuguese Wikipedia. We were able to conclude that the English articles score better across most metrics than the Portuguese articles.

## CCS CONCEPTS

• **Information systems** → Wikis; • **Applied computing** → Consumer health; *Health informatics*.

## KEYWORDS

Information Quality, Wikipedia, Medicine-related Content, Cross-Language Information Retrieval.

### ACM Reference Format:

Gil Domingues and Carla Teixeira Lopes. 2019. Characterizing and comparing Portuguese and English Wikipedia medicine-related articles. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion)*, May 13–17, 2019, San Francisco, CA, USA. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3308560.3316758>

## 1 INTRODUCTION

Wikipedia is a collaborative, open-source, on-line encyclopedia where users from around the world can freely contribute with information on a wide range of topics, in several languages. As of 2018, it is ranked as the fifth most popular site in the world in the Alexa ranking [2].

This editorial openness characteristic of Wikipedia has its advantages and its drawbacks. While anyone is allowed to edit articles, effectively sharing their knowledge with the rest of the world, this

freedom also opens doors to vandalism and incorrect, unfounded information. However, it has been shown that Wikipedia manages to be comparable in quality with other print encyclopedias [4], and that the speed at which its articles can be updated enables the quick neutralization of vandalism [6].

Although medicine articles may not represent the largest portion of Wikipedia articles (over 33,000 out of a total of over 5,700,000 articles in the English Wikipedia) [1], the relevance of these articles cannot be understated. Wikipedia's medicine articles rank highly in the majority of the popular search engines, managing to surpass other resources, such as MedlinePlus and NHS Direct Online [8], which are domain-specific.

Given the importance of the Wikipedia articles of the medical area to the readers, there is a significant lack of analysis of its quality, more prevalent in languages other than English [5].

The goal of this paper is to use previously suggested information quality metrics applied to the Wikipedia domain, as well as other measurements we considered relevant, to analyze and compare a range of characteristics in Wikipedia's medicine articles, for the English and the Portuguese languages.

## 2 RELATED WORK

As the largest collaborative encyclopedia on-line, Wikipedia has, over time, attracted much attention in the research community, namely to the quality of its contents. According to Blumenstock [3], the quality of the articles is correlated with their word count. By selecting a sample of random articles and another sample of *featured* articles - articles selected by the editors as articles of great quality -, the word count of the articles of each group was determined and was demonstrated to be a good indicator of article quality, as articles with a length above a certain threshold were more likely to be correctly classified as *featured* articles.

On the other hand, Zeng et al. [13] rely on information present in the revision history of an article to predict its quality. It builds a revision history-based trust model using a dynamic Bayesian network.

Stvilia et al. [12] use a different approach to analyze the quality of Wikipedia pages. It adapts a set of previously defined metrics for information quality on the web and adapts them to the context of Wikipedia. These metrics - *Authority*, *Completeness*, *Complexity*, *Informativeness*, *Consistency*, *Currency* and *Volatility* - represent different facets of each article, thus providing a more intricate representation of the articles' quality.

As mentioned earlier, research has shown that Wikipedia is one of the most accessed on-line resources by users looking for medical

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.

ACM ISBN 978-1-4503-6675-5/19/05.

<https://doi.org/10.1145/3308560.3316758>

information, ranking highly in the most popular search engines [8].

Heilman et al. [5] analyze four components of medicine articles in Wikipedia: the amount of medical content in Wikipedia, the citations supporting this content, the user traffic for articles on this topic and characteristics of the users who contribute to these articles. This work did not focus on article quality itself, rather on the state of the medical content on Wikipedia across languages, its quantity, readership and the characteristics of its contributors.

Some other works have analyzed the quality of Wikipedia articles in more specific fields of medicine, such as Thomas et al. [11], which evaluated Wikipedia as an educational tool for patients in the field of nephrology, and Kräenbring et al. [7], which studied the quality of drug related information in Wikipedia. Both these works reached similar conclusions that Wikipedia is a fairly reliable and comprehensive source of information in their specific fields.

### 3 METHODOLOGY

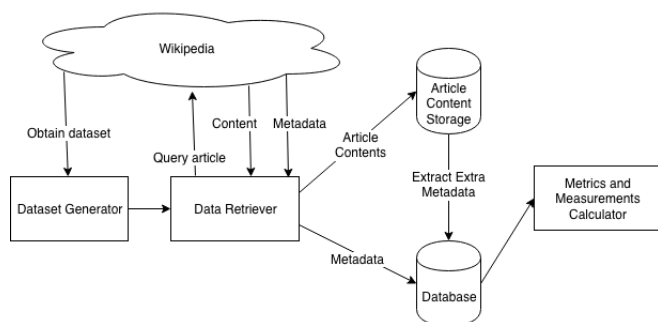
In order to assess the quality of the medicine related articles in both the Portuguese and the English Wikipedia's, an approach consisting of two major steps was followed.

Firstly, we selected a dataset of Portuguese and English medicine related articles and collected the contents of these articles as well as some metadata related to their revision history, editors, content and structure.

Secondly, we chose a set of metrics and measurements that could, on the one hand, help determine the quality of the articles in the dataset and, on the other hand, describe how some of the medical information was structured in these articles.

#### 3.1 Data Collection

The data collection process will be outlined in this section. Figure 1 provides a simplified diagram of the process, from the collection of the dataset and obtaining the article data and metadata from the Wikipedia API, to the process of calculating the metrics and measurements.



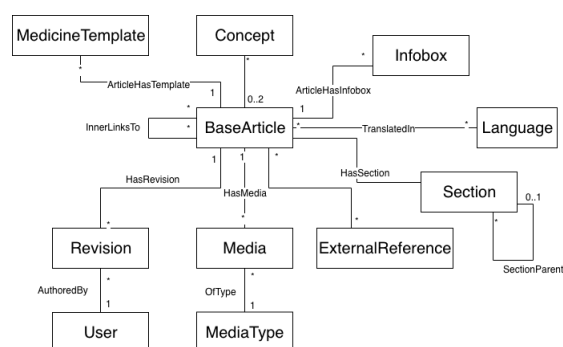
**Figure 1: Outline of the process of data collection, from the creation of the dataset to the calculation of the metrics and measurements.**

To obtain the English dataset we have used the list of top viewed 5,000 articles of the English Wikipedia compiled by user *West.andrew.g*<sup>1</sup>.

<sup>1</sup>[https://en.wikipedia.org/wiki/User:West.andrew.g/Popular\\_medical\\_pages](https://en.wikipedia.org/wiki/User:West.andrew.g/Popular_medical_pages)

The Portuguese dataset was obtained by following the Portuguese language link in each of the articles of the original English dataset. As some of the pages did not have a Portuguese version, this latter dataset is smaller than the former, consisting of 3,496 articles.

The article data and metadata was obtained through the use of the MediaWiki API. Through this API, not only the current state of the article's contents was obtained, but also metadata of the article, information on the revision history, language links, wiki internal links and external links. All the resulting data was stored in a relational database. Figure 2 shows the structure of the database where the data was stored.



**Figure 2: Structure of the database where the data obtained from Wikipedia was stored.**

The remaining data necessary to calculate the metrics was obtained from processing the article's Wikipedia markup. For instance, the media files metadata was gathered through the article's markup as the API's method to obtain the images returns images which are not relevant for the content of the article, such as the lock image in protected pages. When gathering this data from the markup text those images can be filtered out. Templates and infoboxes were also extracted from the markup text, as well as the citations. This information was also added to the database.

To obtain measurements related to text analysis, such as article length, the Flesch reading ease and *InfoNoise* - a function of the number of tokens after stemming and stopping and the article length -, all the markup was removed from the article's content and a plain text file with said content was generated as a result. These measurements were then collected from this plain text file.

In order to only gather data on medicine related infoboxes and templates, two lists of relevant templates were generated. One was composed of all the templates belonging to the category *Medicine Templates* or to its subcategories, while the other was composed of all the infoboxes belonging to the category *Medicine infobox templates* or to its subcategories. Later, when analyzing every article, every template which appeared on the article and on the template list, as well as every infobox which appeared on the article and on the infobox list, were inserted into the database. To obtain the infoboxes and templates in Portuguese, the language links were followed for each entry.

A similar approach was followed to determine which users were administrators. The list of administrators from the English

Wikipedia<sup>2</sup> and the Portuguese Wikipedia<sup>3</sup> were obtained from the pages listing these users. As mentioned before in the metric formulas, this list was required to calculate the consistency metric.

3.2 Metrics

In order to compare the quality between articles, metrics on information quality had to be selected. We resorted to IQ - information quality - metrics previously defined in literature. We used the metrics as defined by Stvilia et al. [12], as these metrics accurately represented distinct components of the articles, being more comprehensive than using a single measurement to evaluate information quality, such as article length.

For the purpose of convenience, the following list describes how each metric is calculated, as described by Stvilia et al [12]. [12]:

- **Authority** = 0.2 \* Num. Unique Editors + 0.2 \* Total Num. Edits + 0.1 \* Connectivity + 0.3 \* Num. of Reverts + 0.2 \* Num. External Links + 0.1 \* Num. Registered User Edits + 0.2 \* Num. Anonymous User Edits.
- **Completeness** = 0.4 \* Num. Internal Broken Links + 0.4 \* Num. Internal Links + 0.2 \* Article Length
- **Complexity** = 0.5 \* Flesch reading ease - 0.5 \* Kincaid grade level.
- **Informativeness** = 0.6 \* InfoNoise - 0.6 \* Diversity + 0.3 \* Num. Images.
- **Consistency** = 0.6 \* Admin. Edit Share + 0.5 \* Age.
- **Currency** = Currency
- **Volatility** = Median Revert Time

These metrics are calculated from a set of measurements obtained from each article. Some of these measurements are simple to obtain, while others have to be computed and are not as straightforward.

When calculating the connectivity measurement, which consists of the number of articles connected to an article via common editors, we faced an implementation obstacle. As we only collected the revisions' history from the articles included in the datasets, the calculation of connectivity was limited to the articles in the dataset. Given the large dimension of the datasets, we assume this approximation will not deteriorate the metric. From an optimistic perspective, it will strengthen the concept of authority as the connected articles will necessarily be related to the topic of medicine.

Analyzing texts in two languages may cause issues with the *Complexity* IQ metric, as the process of calculating the Flesch reading ease and the Flesch-Kincaid grade level are not applicable to Portuguese as they are to English. There is a counterpart to the Flesch reading ease in Portuguese, consisting of adding 42 to the result of the text's Flesch reading ease [10], which was used. However, as far as we are aware, there is no equivalent in Portuguese to the Kincaid grade level. To maintain a *Complexity* metric, we consider its value equal to the article's Flesch reading ease value. While this modification may seem arbitrary, this formula for *Complexity* was present in an earlier version of Stvilia et al. [12] and as such we consider it a reasonable approximation.

The metrics were defined in order to be generic enough to be applicable to an article of any topic in Wikipedia. However, in this paper we intend to analyze articles from the medicine domain,

which motivated the use of a complementary set of measurements, specifically relevant to this domain, which we used to compare the specificity and detail of the medical content of the articles.

One of the elements used to structure information in Wikipedia are templates. Templates are embedded pages that allow the repetition of information. These are examples of very useful components, which can improve the manner in which the readers obtain information and have not been considered in the metrics previously described. Figure 3 shows an example of a template, taken from the influenza page on the English Wikipedia. This particular template is used to store links to external medical resources and classifications.

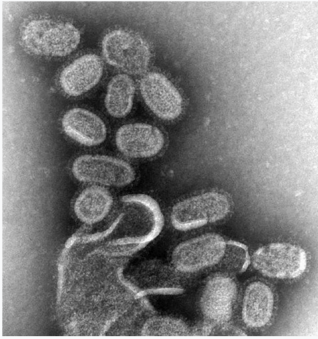
Classification	ICD-10: J10, J11, ICD-9-CM: 487, OMIM: 614680, MeSH: D007251, DiseasesDB: 6791
External resources	MedlinePlus: 000080, eMedicine: med/1170, ped/3006, Patient UK: influenza

Figure 3: Template present in the English wikipedia page for influenza.

One of the most recurring types of template are infoboxes, which provide the user with a way to read critical information in a compact form. These structures are usually positioned on the top of the page, containing a summary of important information and sometimes figures. In figure 4 we can see an example of an infobox, also taken from the influenza page on the English Wikipedia. This infobox contains critical information such as symptoms, causes, usual onset, among others.

Influenza

SynonymsFlu, the flu



Influenza virus, magnified approximately 100,000 times

Specialty

Infectious disease

Symptoms

Fever, runny nose, sore throat, muscle pains, headache, coughing, sneezing, feeling tired<sup>[1]</sup>

Usual onset

Two days after exposure<sup>[1]</sup>

Duration

~1 week<sup>[1]</sup>

Causes

Influenza viruses<sup>[2]</sup>

Prevention

Handwashing, surgical mask, influenza vaccine<sup>[1][3]</sup>

Medication

Antiviral drugs such as oseltamivir<sup>[1]</sup>

Frequency

3–5 million per year<sup>[1]</sup>

Deaths

~375,000 per year<sup>[1]</sup>

Figure 4: Infobox present in the English wikipedia page for influenza.

<sup>2</sup>[https://en.wikipedia.org/wiki/Wikipedia:List\\_of\\_administrators](https://en.wikipedia.org/wiki/Wikipedia:List_of_administrators)

<sup>3</sup><https://pt.wikipedia.org/wiki/Wikip%C3%A9dia:Administradores/Lista>

**Table 1: Median values for each metric by language, along with the p-value resulting from the Mann-Whitney test. Bold values represent the best values in each metric.**

	EN	PT	p-value
Authority	<b>951.65</b>	708.5	<2.2e-16
Completeness	<b>2423.1</b>	576.8	<2.2e-16
Complexity	40	<b>72</b>	<2.2e-16
Informativeness	<b>0.53</b>	0.11	<2.2e-16
Consistency	<b>2582.04</b>	2073.63	<2.2e-16
Currency	<b>21</b>	245	<2.2e-16
Volatility	0.5	<b>0</b>	0.13

The goal of this extra set of measurements is, therefore, to compare the articles through the scope of some structural and content-related meta-data, in combination with the metrics which are primarily reliant on edit history rather than on content. These new measurements are: number of medicine templates, number of medicine infoboxes and number of citations.

## 4 RESULTS AND DISCUSSION

A statistical analysis was performed on each of the IQ metrics. As the samples of the several metrics in both languages do not follow a normal distribution, we have used the Mann-Whitney test to compare the means between the English and Portuguese articles.

The results of the test for each metric are shown in Table 1. The best results for each metric are displayed in bold. *Currency* and *Volatility* are the exceptions where a lower value is better, as lower *Currency* means more up-to-date articles and lower *Volatility* means faster recovery from vandalism.

According to the results presented in Table 1, we conclude that there are significant differences across most of the metrics between English and Portuguese medical articles in Wikipedia. The exception to this being *Volatility*, which means both Portuguese and English articles recover from vandalism in a similar time frame, and *Complexity*, which means Portuguese articles are easier to read than English articles.

One could argue that the *Complexity* metric is not adequate for this comparison. Martins et al. [10] states that the Portuguese Flesch adaptation was not researched extensively enough to guarantee complete accuracy, which can lead us to believe that the comparison between Portuguese and English article *Complexity* may not be reliable.

In comparison to the results obtained by Stvilja et al. [12], our results follow the same order of magnitude, with some inflation in metrics which involve a number of events in edit history, such as *Authority* and *Consistency*, which is to be expected given the date difference between the two analysis.

In Figure 5 we can see the number of medicine templates, medicine infoboxes and total citations in the English and Portuguese articles in our samples. Each data point in these graphics represents an article of a given language. These measurements were considered as particularly relevant in the context of medicine related articles. As one can observe from the figure, the English articles make more use

of medicine related templates and infoboxes, as well as including a greater number of citations, when compared to Portuguese articles.

It should be noted that, from the English set, 129 articles were classified as *good* articles - articles which meet a set of editorial standards, but are not *featured* article quality - and 47 as *featured*, whereas in the Portuguese set only 15 articles were *good* articles and 27 were *featured*. English articles appear to use more templates and infoboxes, which may indicate a better effort in terms of content organization and reutilization of information, as well as more citations.

The benefits of English content found in this work are corroborated by others studies. Previously, Lopes et al. [9] have concluded that, when compared to Portuguese, English web content is more adequate for the distribution of health information.

## 5 CONCLUSION

We performed a quality comparison between English and Portuguese Wikipedia medicine articles, across a set of descriptive generic metrics, as well as some structural metadata and medicine-specific article elements.

We concluded that the medicine articles vary significantly in terms of quality, with the English Wikipedia outperforming the Portuguese Wikipedia across the majority the metrics used to describe the articles. Our results corroborate the importance of the investment in cross-language information retrieval strategies in the health domain.

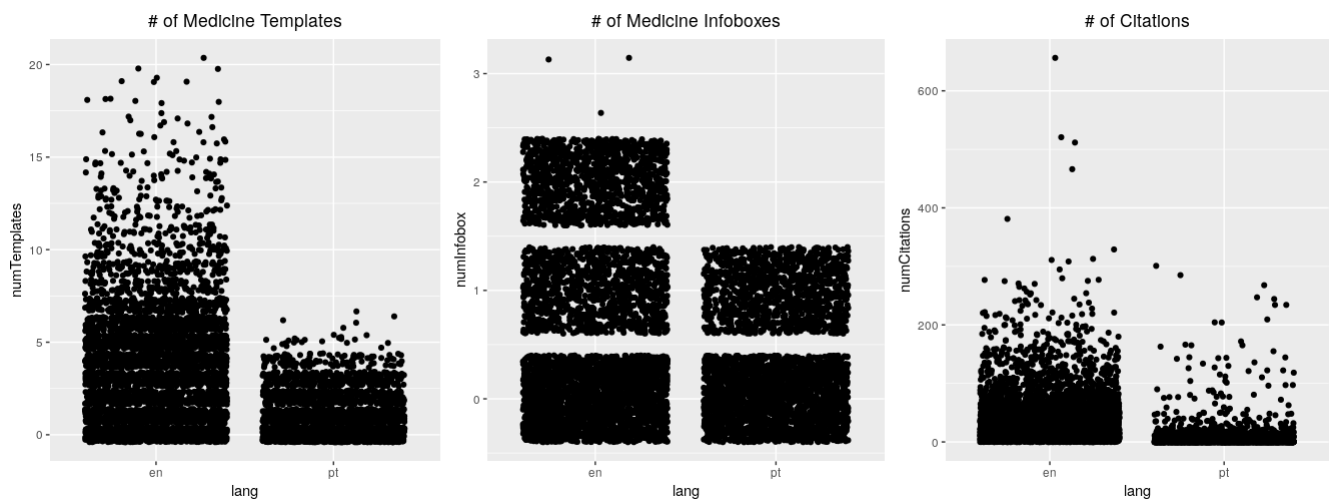
As future work, we intend to investigate and develop a set of quality metrics specific to the medicine domain on Wikipedia, most likely using some of the extra measurements we used in this analysis.

## ACKNOWLEDGMENTS

Project “NORTE-01-0145-FEDER-000016” (NanoSTIMA), financed by the North Portugal Regional Operational Programme (NORTE2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

## REFERENCES

- [1] 2018. WikiProject Medicine. [https://en.wikipedia.org/wiki/Wikipedia:WikiProject\\_Medicine](https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine)
- [2] 2019. The top 500 sites on the web.
- [3] Joshua E. Blumentstock. 2008. Size matters: Word Count as a Measure of Quality on Wikipedia. *Proceeding of the 17th international conference on World Wide Web - WWW 08* (2008). <https://doi.org/10.1145/1367497.1367673>
- [4] Jim Giles. 2005. Internet encyclopaedias go head to head. *Nature* 438, 7070 (2005), 900–901. <https://doi.org/10.1038/438900a>
- [5] James M Heilman and Andrew G West. 2015. Wikipedia and Medicine: Quantifying Readership, Editors, and the Significance of Natural Language. *Journal of Medical Internet Research* 17, 3 (Apr 2015). <https://doi.org/10.2196/jmir.4069>
- [6] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. He Says, She Says: Conflict and Coordination in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, New York, NY, USA, 453–462. <https://doi.org/10.1145/1240624.1240698>
- [7] Jona KrÄČÄdenbring, Tika Monzon Penza, Joanna Gutmann, Susanne Muehlich, Oliver Zolk, Leszek Wojnowski, Renke Maas, Stefan Engelhardt, and Antonio Sarikas. 2014. Accuracy and Completeness of Drug Information in Wikipedia: A Comparison with Standard Textbooks of Pharmacology. *PLoS one* 9 (09 2014), e106930. <https://doi.org/10.1371/journal.pone.0106930>
- [8] M. R. Laurent and T. J. Vickers. 2009. Seeking Health Information Online: Does Wikipedia Matter? *Journal of the American Medical Informatics Association* 16, 4 (Jan 2009), 471–479. <https://doi.org/10.1197/jamia.m3059>



**Figure 5: Number of medicine templates (left), medicine infoboxes (center) and citations (right) in the English and Portuguese articles.**

- [9] Carla Teixeira Lopes and Cristina Ribeiro. 2013. Measuring the value of health query translation: An analysis by user language proficiency. *Journal of the American Society for Information Science and Technology* 64, 5 (2013).
- [10] Teresa B.F. Martins, Claudete M. Ghiraldelo, M. Graças V. Nunes, and O.N. Oliveira Jr. 1996. Readability Formulas Applied to Textbooks in Brazilian Portuguese. (Jun 1996).
- [11] Garry R Thomas, Lawson Eng, Jacob De Wolff, and Samir Grover. 2013. An Evaluation of Wikipedia as a Resource for Patient Education in Nephrology. *Seminars in dialysis* 26 (02 2013). <https://doi.org/10.1111/sdi.12059>
- [12] Besiki Stvilia, Michael B. Twidale, Les Gasser, and Linda C. Smith. 2005. Information Quality In A Community-Based Encyclopedia. *Knowledge Management* (2005). [https://doi.org/10.1142/9789812701527\\_0009](https://doi.org/10.1142/9789812701527_0009)
- [13] Honglei Zeng, Maher A. Alhossaini, Li Ding, Richard Fikes, and Deborah L. McGuinness. 2006. Computing Trust from Revision History. (Jan 2006). <https://doi.org/10.21236/ada454704>