Field Lab for the Attainment of the Degree

Master of Science
at NOVA SBE Lisboa

Convergence of Truth through Language Links in Historical Data

Convergence Beyond Battles

Advisor: Michael Kummer

Study program: Master of Science in Business Analytics

Submitted by: Jona Weishaupt

61374

Submitted on: 20.01.2025

Abstract

Our research examines whether numerical data on historical battles, such as casualty figures or troop strengths, converge over time across multiple languages on Wikipedia. Our analysis of annual revisions, reveals no trend of convergence, in most cases the numbers rarely change, with discrepancies persisting or even increasing over time.

The study also tests these patterns n simpler data categories outside of the battle context, like bridge length. Findings include persistent barriers, like editorial biases and resource disparities, that limit numerical consistency and reflect cultural and editing dynamics in global knowledge.

Collectively this thesis finds barriers that limit data alignment across languages and ultimately reveals that Wikipedia's goal of unifying knowledge globally faces unaddressed challenges.

1. Introduction

In today's globalized environment, Wikipedia stands as both a crucial knowledge resource and a reflection of global information practices. Available in over 300 language editions, it constitutes an unprecedented collective knowledge venture shaped by a multitude of cultural, linguistic, and historical backgrounds. Although this openness offers significant opportunities for democratizing information access, it also introduces complexities. Divergent cultural narratives, linguistic differences, and varying historical interpretations can create inconsistencies, particularly when comparing content across languages. This thesis contributes to broader discussions on the reliability of global knowledge-sharing platforms and the impact of cultural factors on collective memory.

The research within this work project focuses on numerical data related to historical battles, such as troop strengths, casualty figures, and outcomes, as an illustrative case for examining how contested historical information evolves on Wikipedia. By analyzing data from six major language editions – English, French, German, Portuguese, Spanish, and Italian – this study seeks to identify and understand patterns of convergence and divergence in reported figures over time.

To provide additional context, the thesis also investigates numerical data in less contentious domains. This includes metrics related to infrastructure (e.g., bridge lengths), scientific facts (e.g., melting points of elements), and geographic data (e.g., mountain heights), as outlined in the work of Hecht and Gergle (2010).

Methodologically, the project employs a multi-layered approach to data collection and analysis. A custom-built pipeline integrates Wikipedia's API to retrieve historical article revisions, an HTML

parser to extract structured data from infoboxes, and the Gemini AI model to preprocess unstructured textual content. This framework supports the systematic gathering of metrics (troop strengths, casualty counts, and battle outcomes) over extended temporal ranges for approximately 40-50 battles across six language editions.

Methodology

To accurately measure convergence and reduce the chance of statistical errors, we needed a solid and organized approach. This chapter walks through the main steps of our analysis, including how we gathered data, prepared it, selected the right metrics, and the statistical methods we used to assess convergence. Our goal was to create a clear and repeatable framework for the study, which we will also apply to the different avenues we explore outside of the battle context.

2.1. Data Collection

We aimed to build a dataset that tracks how battle-related information has changed over time across different languages. Specifically, for each battle, we collected one version of its Wikipedia page per year, starting from the first mention. This allowed us to follow how key metrics like troop strength and casualties evolved over time. The data we chose to track, composed of: troop numbers for each side, giving us an idea of the conflict's scale, and detailed casualty figures broken down into deaths, injuries, missing persons, captures, and total casualties. We chose these metrics because they are consistently reported and are the main numeric indicators for battles, that rely on different sources and could therefore converge over time with the introduction of language links. Additionally, having this detailed information lets us explore different ways convergence might occur in our analysis.

To ensure our analysis was thorough, we gathered data in six languages: English, French, German, Portuguese, Spanish, and Italian. These languages were selected because they represent major European powers that played significant roles in battles before World War I, which is the focus of our study. Furthermore, they are written in the Latin alphabet, which eased the process of collecting the data and reduced the risk of errors due to wrong translation.

Our main goal was to see if adding interlanguage links on Wikipedia led to more consistent numerical data across different language editions. By collecting yearly revisions in multiple languages, we could examine whether these links helped standardize battle-related metrics through shared sources and cross-referencing. The data we collected forms the basis for assessing this trend toward convergence.

Manually collecting this information was not feasible because of the large amount of data. Analyzing 40 to 50 battles, each with about 20 yearly revisions and six language versions, would result in thousands of data points. Manually extracting data from each infobox would take one to two hours per battle, making the total workload too high. Additionally, Wikipedia's dynamic nature means there are inconsistencies in formatting, language-specific conventions, or naming schemes, which would complicate manual efforts due to, for example, translation requirements.

To handle this, we decided to automate the process, allowing us to scale up in the future and also give us the opportunity to expand our information base beyond the battle context. However, directly scraping structured data from infoboxes was challenging. As is shown by Exhibit 1, the numerical metrics for troop strength and casualties were often stored as unstructured text rather than in separate columns, making automated parsing difficult. There were also many data inconsistencies: some infoboxes only included partial metrics (like the number of deaths but not injuries), while others provided ranges (e.g., "50,000–100,000"). When multiple nations were involved on one side,

their contributions were listed separately instead of being combined, adding another layer of complexity. These issues were made worse by changes in formatting and naming conventions across different languages and over time.

To overcome these issues and build our automated Wikipedia scraper, we adopted a three-step approach:

As a first step, we used the Wikipedia API to collect yearly revisions of each battle's Wikipedia page. Utilizing the Wikipedia API, we were able to automatically find and store the yearly revisions for each language. These revisions were then saved by their URL in a structured dataframe that would be filled at a later stage with the numeric information scraped from the infobox.

Once the yearly revisions were obtained, we used an HTML parser to extract the raw infobox content. The raw data is then cleaned from any formatting and saved to be further processed in the next step. As explained previously, due to the dynamic nature of the infobox, we cannot directly process the raw text in a structured and scalable fashion.

In the final step, to overcome this challenge, the extracted raw data is processed using an AI processor. We chose the Gemini AI processor because it is freely available and accessible via an API. As large language models like Gemini or ChatGPT, are more powerful with English data, any non-English text is translated to English using Google's free translation API. The usage of an AI processor allows us to directly establish the same guideline as our manual scraping efforts through prompting, without having to investigate the data ourselves. The output of this processing step is a structured dictionary format, organizing the data into categories for each side of the conflict (e.g., Strength_A, Strength_B). The data returned can then easily be inserted into the columns of our

previously empty data frame. The same procedure is then repeated for revision in the data frame and for each language specified.

During this process, we noticed that a modification to our approach was needed. While our initial plan was to collect as much data as possible (separating into Injured, Captured, Missing, etc), we realized that the inconsistency in reporting of these metrics made it impossible to analysis at a later stage. Specifically, in many cases these metrics were not reported at all or just summed up as a total number (for example "4000 Missing, Injured or Captured"). This left us with a large amount of NaN fields, which made comparisons across languages unfeasible. Hence, we decided to modify our approach by changing the data we wish to collect. We focused on Strength, which is always recorded consistently and the total number of casualties summed up, without separation in different categories. Additionally, we improved our handling of ranges (for example "6000-7000 dead") by recording both numbers as Lower- and Upper bound. This way we minimized the number of NaN values and handled ranges without modifying the original data, by for example calculating an average of the two numbers. As a result, our final data was stored in the columns: Strength_Lower_A, Strength_Upper_A, Total_Lower_A, Total_Upper_A, with the same repeated for the other side, as shown in *Table 1*.

Battle Name	**Date**	**lang**	**Strength Lower A**	**Strength Upper A**	**Total Upper A**	**…**	**Total Lower B**
Siege Paris	23.01.2009	de	6000	7000	1000		500

Table 1: Example Row of our Dataframe

To ensure the accuracy of the AI-processed data, we compared Gemini's results with data that had been manually scraped in previous tests. This comparison confirmed that the AI generally performed well, closely matching our manual extraction results. However, we did encounter some discrepancies due to variations in infobox formatting that Gemini was not able to handle. The handling of these inconsistencies are addressed in detail in the Data Processing chapter of this thesis. A detailed workflow of our automated bot is illustrated in *Figure 1* which outlines the interactions between the Wikipedia API, the HTML parser, and the AI processor.

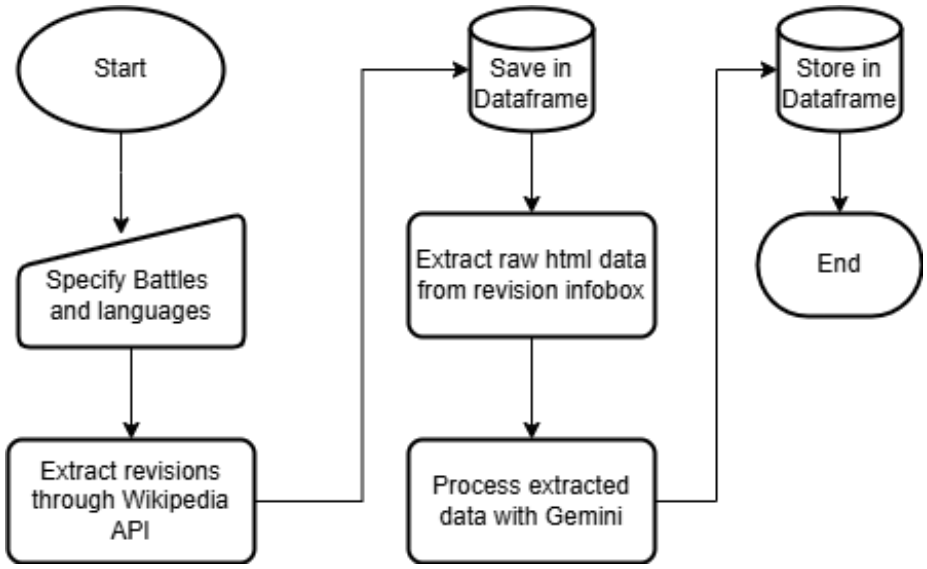

Figure 1: AI Processor Workflow

2.2. Data Cleaning

As we touched upon earlier, simply relying on AI-driven automated scraping doesn't guarantee perfect accuracy. That's why processing and thoroughly cleaning our data before moving forward is very important. We wanted to be sure our dataset was as accurate as possible, to achieve this we implemented the following techniques:

First, we made sure all the data was in the same format. Different languages handle numbers differently—many European languages, like French or German, use a period to separate thousands instead of a comma. Python interprets this as a decimal point, which could cause the

numbers to be largely skewed. To prevent that, we changed the separators so the numerical data would be stored correctly in our dataframe.

We also tackled sudden spikes in the numbers. Sometimes, due to how the data was processed, values could suddenly jump to very high levels. Given the historical context of battles, it's not impossible for certain figures (due to, for example, injuries now being included in total casualties) to surge over time. Still, we set a threshold of 1000%. If any number suddenly increased from one revision to another above that threshold, it was inspected and if needed, we corrected it with our backward fill method. We picked such a high threshold to make sure we weren't changing numbers that were actually correct.

2.3. Metrics for Convergence

To statistically measure if our data converges, we developed and selected an approach that can easily be replicated in future expansions of our research.

The main metric we calculated was the relative difference of each datapoint to the reference point, see *Equation 1*. Specifically, for each recorded datapoint—such as troop strength or casualties—we calculated the relative difference between a given language's value and the corresponding English value. This was achieved using the formula:

$$Percentage\ Difference = \frac{Value - Reference\ Value}{Reference\ Value} \times 100$$

Equation 1: Relative Percentage Difference Equation

Choosing the relative difference and not the absolute was the obvious choice for us given the context. Absolute differences would have caused comparisons to be dominated by battles featuring

extraordinarily large numbers, overshadowing the convergence patterns in smaller or more moderate battles. By normalizing the data in terms of proportional change rather than raw magnitude, we obtained a more balanced perspective. Adopting this approach also allowed us to skip the normalization steps we would have otherwise had to take, as this is our main metric which is already normalized against the reference value in its' calculation. As a result, every data point, whether from a large battle or a smaller skirmish, could be assessed on an equal term. Furthermore, after having computed the relative differences for each data point, we averaged these values per year for each metric and language. Visualizing these values over time allowed us to get a first feel for the data developments, see any spikes that we might have missed and estimate if convergence is happening or not. If the slopes of our trendlines decrease over time, we would be able to assume convergence, however without statistical assurance.

The slope of this regression line served as a "convergence coefficient." A negative slope indicates that, as time progresses, the values reported in the given language edition are becoming more similar to those in English. On the other hand, a positive slope suggests that the differences are growing over time, implying divergence rather than convergence.

Just computing slopes is not enough; it is crucial to establish whether these slopes are statistically meaningful. To achieve this, we conducted one-sample t-tests on the slopes. The t-tests determine if the average convergence coefficient significantly differs from zero. A statistically significant negative slope would confirm that the observed trend toward convergence is unlikely to be the product of random variation. Similarly, a non-significant or positive slope would prompt us to reconsider the patterns or investigate potential sources of divergence. To gain a more complete understanding of convergence in each language, we then averaged these convergence coefficients across all metrics for each language.

3. Findings

4.1. Results

Following the extensive data cleaning, we were left with 40 battles that were processed during our analysis. As stated, for each revision we compared the scraped data to the data of the reference language English. As a first insight we computed the average percentage difference in the various categories across languages, the results can be seen in *Table 2*.

Strength Lower A	Strength Upper A	Total Lower A	Total Upper A	Strength Lower B	Strength Upper B	Total Lower B	Total Upper B
14,6%	13,8%	29%	30%	16%	17%	26%	22%

Table 2: Average percentage difference per category across languages

At first glance, we can see that the lower and upper bounds of each category are fairly identical in the average percentage difference between them across languages. Only for the Side B values for the total casualties we see a 4% difference, which could be due to a few battles that are influencing these numbers, which isn't abnormal due to our smaller sample size. Furthermore, we can see that the average difference is much higher for both sides for the casualty values. This is to be expected due to the inconsistent reporting for this category. As mentioned before, it is common that there are large differences between the revisions in this category, some may include captured in the total, while some may not. For strength this is not the case, which is why we have much smaller differences on average in this category. As a next insight, we investigated each language separately by computing the average difference to English across all categories for each language. The results can be seen in *Table 3*.

German	Spanish	French	Italian	Portuguese	English
24%	23%	33%	28%	27%	0%

Table 3: Average percentage difference per language

The overall are fairly similar across all languages, with Spanish and German having the lowest average difference and French the highest. Of course, as English is the reference language, the difference to itself is 0%. This gives us an indication there might be a similar theme across languages and that there are a good number of differences in reported numbers between the English data. However, while the numbers show that reported values differ on average, it gives us no insights about convergence over time. For this we computed average percentages per year across all languages and all languages. The results can be seen in *Figure 2.*

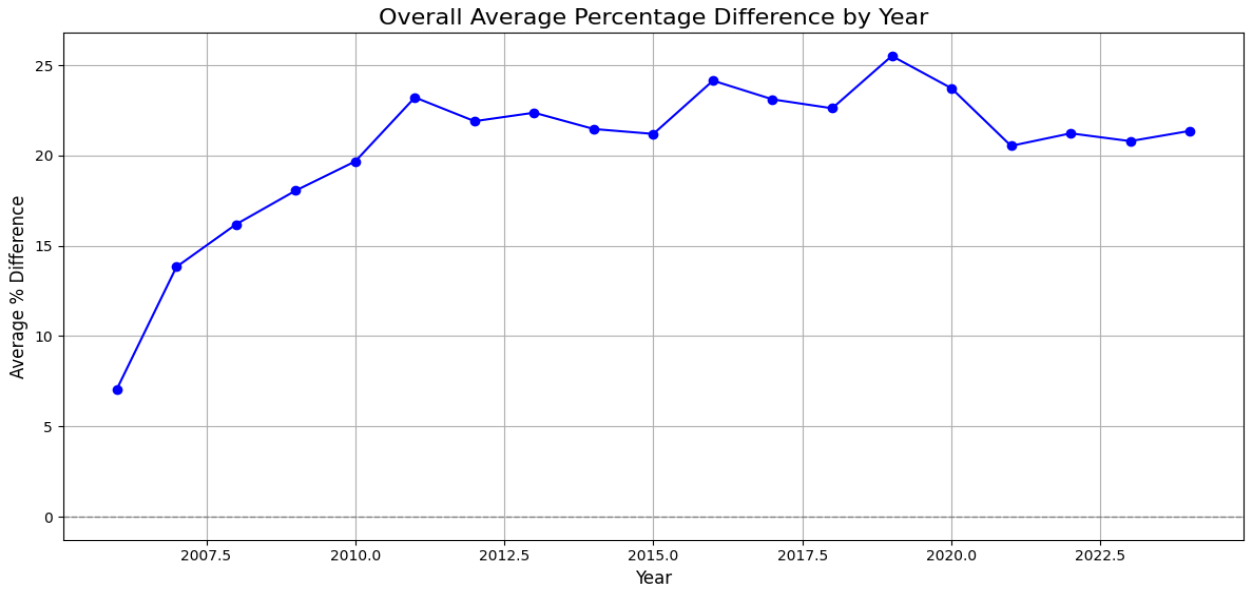

Figure 2: Overall Average Percentage Difference per Year

As evident, the average difference for all metrics and languages does not seem to decrease over time, which would signal convergence. It also does not seem to increase over time either, rather it stagnates over time. The lower difference at the beginning in earlier years can be attributed to lack

of data available for many battles in the early years of Wikipedia or that early versions might have just been a direct one to one translation from another language. From the developments over time, we could draw the conclusion that instead of converging or diverging, the differences seem to stagnate, meaning that numbers are drawn from a source at one point of time and rarely change after that. This is interesting as it implies that languages seem to stick with their sources and do not revise or compare them with other languages. To gain a better understanding of these developments, we split the computation above into the various language groups, to see if the above holds across all languages. The results can be seen in *Figure 3*.

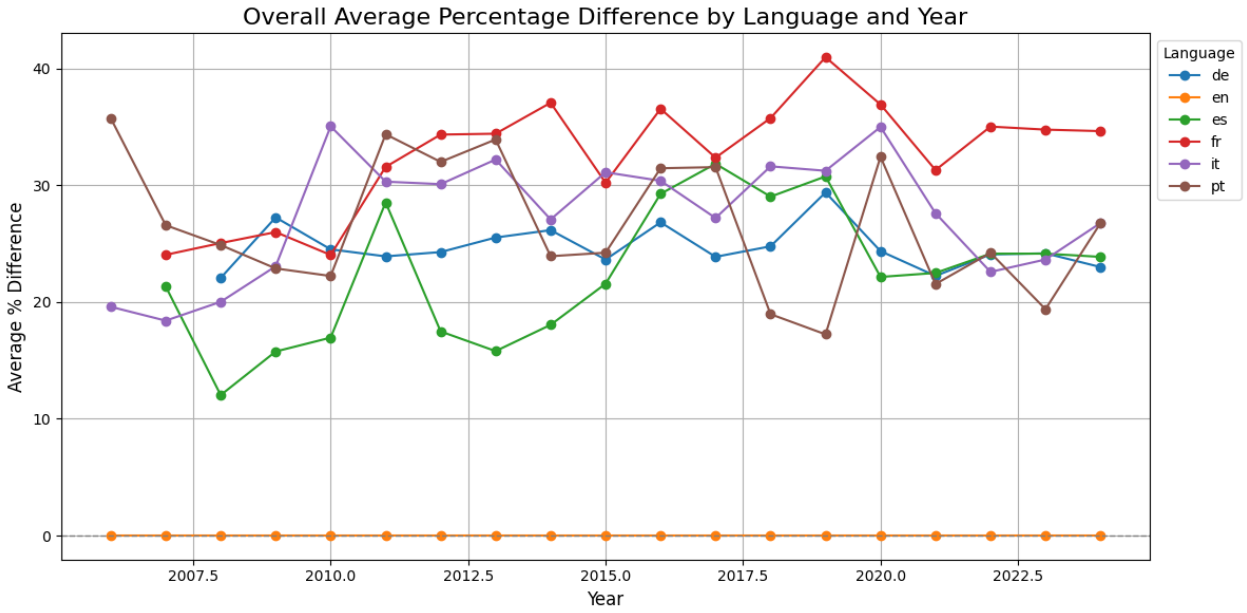

Figure 3: Overall Average Percentage Difference by Language and Year

The data confirms our previous suspicions, the trend for all languages moves fairly identically, with French having the highest average differences and German the lowest. Again, we are not able to see convergence or divergence for any language over time. The trends remain that the differences seem to stagnate around a certain point, indicating that languages stick to their sources and data.

To validate these results, as explained in our methodology, we conducted a one sample t-test. The results of our t-test per language can be seen in *Table 4*.

Language	t-statistic	p-value	Significant
de	1.403276	0.203303	False
es	4.492156	0.002825	True
fr	2.688328	0.031161	True
it	2.214359	0.062395	False
pt	1.376840	0.210981	False

Table 4: T-test results of the original battle context

The results of our tests confirmed some of our suspicions but also revealed some surprises. First of all, for German, Italian and Portuguese we confirmed our theory that the trend of time is stagnation. The convergent coefficient (slopes of the trendlines) is statistically not different from zero, meaning we have no evidence to reject the null hypothesis that the coefficient is different from zero. On the other hand, for French and Spanish, we see a p-value <5%, meaning that we can have enough evidence to reject the null hypothesis, showing that statistically the convergence coefficient differs from zero. However, after looking at the coefficients resulting from our linear regression in *Table 5*, we see a positive slope for French and Spanish. This means that for these two languages, the data seems to diverge over time, contradicting our theory of convergence over time. Seeing these results, we found it unnecessary to include language links in our analysis, as they would only be relevant and interesting to analyze if we actually have convergence. These results do not surprise us, as during our manual scraping attempts, which were done to validate the results of our automatic bot, we saw rare movements in the battle metrics over time. On the contrary, we witnessed just

small changes and rather as proven by our statistical tests and visualizations, the language's stayed with their sources over most of the revisions.

Language	Average Slope
de	0.414533
es	1.067906
fr	1.269892
it	1.021357
pt	0.614562

Table 5: Average Slope per language in the original battle context

4.2. Limitations

While these results provide valuable insights, that we will also explore and expand in the following parts of our thesis, they do come with certain limitations that we need to acknowledge. First of all, with the limitations we put on the languages and the battle timeframe, the sample size is relatively small. While we did conduct extensive cleaning, investigated outliers and corrected wrong data, some singular battles with this large of a sample size can still possibly skew our results. A larger sample size with 300+ battles would balance these out and allow for more generalizable results. Adding to this, there are still possible results from the battle scraper that are incorrect, which would also skew our results. A more robust scraping method with a stronger AI processor developed in the future, could strengthen the results. Furthermore, for some languages there is just more extensive work put into Wikipedia articles than others. Larger language communities like French, Spanish or English, are likely to have more revisions and data available than Italian for example. Finally in our scraping methodology, we are focusing only on the Infobox, if this not available or no casualty data is recorded in it, we have no data. The entire text offers a more complete picture

and also offers more data on certain metrics. An expansion of the scraper that can include information from the entire text, would strengthen our results.

Convergence beyond battles

5.1. Motivation

While the battle context provides valuable and informative insights into data convergence on Wikipedia, it represents a single, extensive category. This focus limits understanding as patterns observed within this context may not necessarily apply to other types of data. Additionally, confounding variables tied to battles could influence my conclusions about convergence, making it difficult to generalize findings across the broader scale of data available on Wikipedia. I therefore find a sequential analysis of a broader scope of data categories an important measure to achieve a more comprehensive and accurate picture of convergence trends.

This expansion was achieved by including, at a first glance, simple datapoints, such as the length of bridges, the height of mountains, or the size of lakes, into my analysis. These additional categories differ in complexity and provide a more balanced picture of the convergence of information across different languages and revisions of Wikipedia articles. Tracing the convergence trends over time across these varied categories allows me to check if the patterns found in the context of battle are similar or not when it comes to other forms of data.

This broader approach serves multiple purposes. It not only affirms my initial findings through testing against my battle dataset but also adds reinforcement for concluding information consistency throughout Wikipedia, given that divergence trends in battle-related data are also observed within other kinds of datasets. On the other hand, significant differences indicate that convergence relies on specific properties of individual categories, justifying the need for

categorical treatments. Widening the range of data points therefore strengthens and reinforces the reliability of my study because the conclusions drawn will no longer be biased and will not portray the characteristics of a particular group of data but rather present a more authentic representation of what the convergence dynamics on Wikipedia are. Future research can then be expanded into all kinds of categories to connect to the results of this thesis.

The expansion will also serve a practical purpose. The automatic Wikipedia page scraper will be modified and able to handle any data group on Wikipedia, which will allow information extraction from any Wikipedia page, not just specifically in the battle context.

To conclude, this expansion beyond the battle context will provide a generalizable view on convergence trends. By including various new categories, I can verify my findings, uncover new patterns, and validate that my prior conclusions are applicable across different types of information. Adding to my work in this regard will strengthen my thesis and facilitate future research.

5.2. Methodology

Building on our original research, I've chosen to maintain the same methodology to ensure comparability with prior results and consistency in the variables used. This continuity allows me to draw meaningful comparisons and build upon my earlier findings. However, as I am now working with a new dataset that includes additional categories and complexities, my previous scraper—designed specifically for the original context—no longer fully meets my needs. To address these challenges, I must adapt my approach and implement different data collection methods tailored to the unique requirements of the new data.

The focus of my expanded research involves scraping specific data points across various categories to enhance my understanding of convergence on Wikipedia. Each category was chosen

because it offered a wide range of examples to fill the dataset and, most importantly, has a numeric value that is consistently available to scrape. Hence, I decided on the following categories:

First, I am collecting length data for bridges, as it serves as their primary datapoint and is present in all infoboxes. Similarly, for mountains, the height is the key measure consistently available. Moving on, I decided on tunnels, specifically scraping for their length, similar to bridges. The same metric I also scraped for rivers, where length also serves as the primary datapoint. For lakes, the surface area was taken to be stored. Finally, I expanded to chemical elements, for which I am scraping the melting points. While there are many numeric datapoints that can be extracted for elements, melting points were proven to be consistently stored in the infobox and, more importantly, always stored in the same scale: Kelvin.

These specific data points are essential for a detailed and nuanced analysis within their respective contexts, but their varied nature and unique formats necessitate adjustments to my data collection methods. By tailoring my approach to accommodate these differences, I aim to ensure the accuracy and reliability of the data collected for this expanded study.

5.2.1. Data Collection

The automatic scraping of the data also followed a similar approach to the battle context. While I first tried to modify the script, to skip the AI processing step and directly store the raw number from the html, I quickly noticed that this would be unfeasible. Even though most of the time, as we are dealing with a single number only, it is possible to directly scrape the number without having to process it, the formatting of the Wikipedia Infoboxes once again proved unreliable. The data is often not presented in the same metric scale, for example 0.14km or 140m, which required a

processing step to convert the data to a standard format. In addition, on the English pages, the data was usually not shown in the metric system, but rather in the imperial system. Bringing the data to a common scale in post processing with python proved challenging and time consuming, which opted the way for an AI processing step like in our prior scraper. The only task of this step was to convert the extracted data to meters format, so that all data could be compared evenly. Another difference, was that as I am only extracting one variable per category, I was able to reduce the complexity of the dataframe to just one column for the extracted numeric value. This simplification was done in such a way, that if I wanted to expand to even more categories, only the name of the category and the name of the variable that we want to scrape needs to be added to the script. The final dataframe structure can be seen in *Table 6*.

Landmark Name	Revision ID	Revision Date	lang	Category	Field Name	Value
Mount Everest	675541	2004-08-05	fr	Mountain	Elevation	8844m

Table 6: Context Expansion dataframe structure

5.2.2. Data Cleaning

Once again data cleaning was necessary, however as I am only investigating one value per category, this process was much simpler as before during our battle context. My approach followed the same standards as before during the battle context, unexpected NaN values were filled through a backwards fill approach, sudden spikes of over 100%, were flagged as an error and filled with the previous values (I do not expect the height of a mountain to suddenly increase by 100% due to a new source like in the battle context) and increases over 50% were investigated and changed if needed. The thresholds were set lower as in this context, the metrics are more “factual” in

comparison to battles, were two sources might disagree in a substantial amount. These techniques smoothed out any errors done by the processor, due to bad conversion for example and increased the accuracy of the dataset substantially.

5.2.3. Metrics for Convergence

To actually measure convergence, I followed the same approach as from our original battle context. The only difference being, as I am only dealing with one numeric value scraped per category, I am only measuring convergence for this singular metric, in comparison to the 2 sides and 2 metrics for battles. However, the main convergence metric, being the relative difference between the value and reference value in English stays the same.

$$Percentage\ Difference = \frac{Value - Reference\ Value}{Reference\ Value} \times 100$$

Equation 2: Relative Percentage Difference Calculation

After computing this percentage difference, I then once again investigated the various summary statistics, like the mean difference per category, language and year. Giving a first insight into how each category overall differentiate from each other when it comes to convergence.

To better understand the convergence dynamics, I visualized these trends for each category and language over time. These visualizations made it easier to observe how differences evolved, revealing both consistent patterns and any notable outliers. I also performed linear regression on the data for each category and language to compute the slopes of the trendlines. These slopes quantified the rate and direction of convergence, providing a precise measure of how differences were decreasing or increasing over time.

As a final step, I performed a one-sample t-test on the slopes across all categories to determine whether they were statistically significant. This provided the confirmation needed, ensuring that the observed trends were not due to chance. A significant result with a negative slope, indicates that the trends reflect meaningful convergence dynamics.

6.3 Results

Following our data cleaning and processing, I was left with 168 unique entries across my 6 categories, with 25 bridges, 51 elements, 23 lakes, 25 mountains, 25 rivers and 19 tunnels respectfully. A first inspection of each categories average percentage difference across the entire timeframe, can be seen in *Table 7*.

Bridge	Element	Lake	Mountain	River	Tunnel
1.25%	0.37%	1.92%	0.05%	1.9%	0.65%

Table 7: Average percentage difference across category expansion

As evident from the numbers, the average percentage difference to the reference language value is very low, especially compared to the battle context in which we saw differences ranging above 25%. These results are not surprising at all given the context, in my expansion I am dealing with categories that are fact based, without much room for interpretations or estimations. The height of a mountain can be measured and checked constantly, while the reported casualties of a battle 200 years ago cannot. Therefore, only minor differences between the reported numbers are expected, as larger deviations would not be logically consistent. Especially for Elements, I expect a common scientific consensus, that does not allow room for interpretation. Interesting to see is that Rivers and Lakes have (even though only slightly) the highest percentage difference across the categories. While this may seem counterintuitive at first, as for examples we don't expect Rivers to grow, this

may be due to different measurements being used, that may or may not include certain side rivers into the equation. For example, by exploring the data and looking at the most recent revisions of the world's second largest River, the Amazonas, we would still get different lengths depending on language we are searching in. This dynamic can be seen in *Table 8*.

Language	de	en	es	fr	it	pt
Length	6,400,000	6,400,000	7,062,000	6,625,500	6,992,000	6,992,000

Table 8: Amazons example

The comparison of average differences between languages in our expanded dataset aligns closely with our previous findings. Across all languages, the reported differences remain consistently low. This mirrors the results observed in our battles context, where the average differences between languages were similarly minimal, showing only slight variations. Interesting to see is that again, French, Italian and Portuguese have the highest average difference of all languages we investigate. An overview of these numbers can be found in *Table 9*.

de	**en**	**es**	**fr**	**it**	**pt**
0.94%	0%	0.956%	1.22%	1.31%	1.001%

Table 9: Average percentage difference across languages in category expansion

Even though the differences are small, it's still important to investigate if there are convergence trends. As mentioned before, the captured data in theory leaves no room for interpretation, as it is scientifically measurable, giving the question of convergence overtime a much larger sense of relevance.

To do this, I need to look at how the data evolves over time—first as a whole, then for each category individually. I'll start by analyzing all categories together to see if there's a general downward trend in differences, which will help us understand whether the data is converging overall. As can

be seen by *Figure 4*, the opposite seems to be the case, from its lowest point in 2011 at around 0.5% difference, the average difference across all categories increase to its maximum of 1% in 2024. This is surprising to see, but it could be driven by the high percentage categories Lakes and Rivers, which makes it imperative to dive into a per category differentiation. However again, I am able to see the low overall differences, with only 1% being the maximum. This confirms the previous hypothesis that in our category expansion, we are dealing with much more factual data, that can confirmed or rejected through scientific means. The casualties of battles from 200 years ago do not offer this advantage.

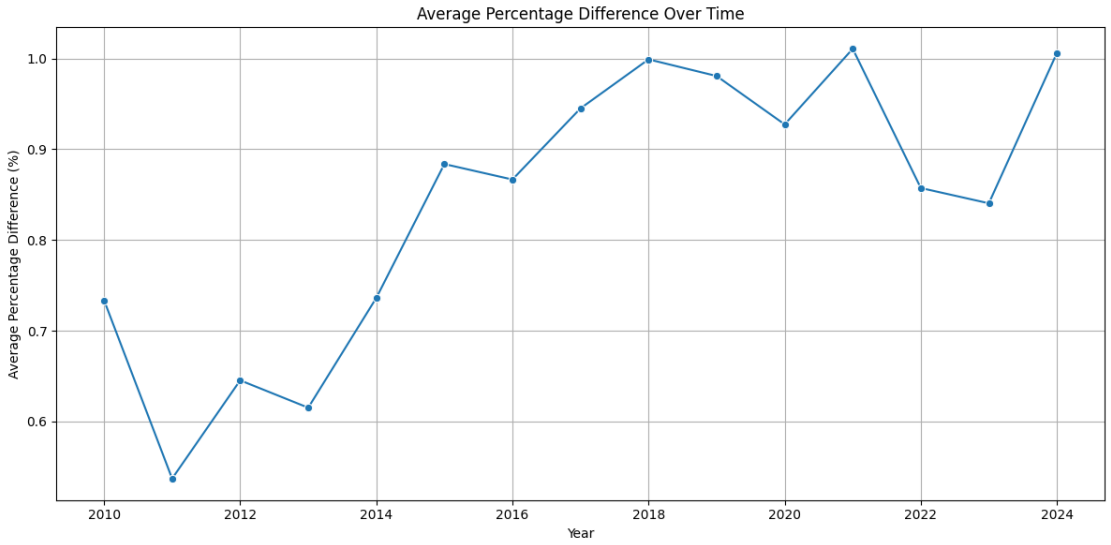

Figure 4: Average Percentage Difference over Time in category expansion

Looking at categories split, represented by *Figure 5* I get a better understanding of the overall trend of divergence I saw before. As theorized, the three categories that showed the highest percentage differences, also push the overall trend upwards. Bridges, Rivers and Lakes all seem to display a trend of divergence, while Tunnels, Elements and Mountain stay relatively stable. However these numbers are still very small, with the highest average peak of any category being only 3%. The overall conclusion that I can draw, is that we again do not see any sign of convergence, rather the

opposite or a constant level of difference in the categories, similar to battles. It is also interesting to see how low the category of Mountain remains, with almost no divergence at all from the reference language English. This could be due to the fact that the height of a mountain leaves little to no room for different calculations, while bridges could count the length from different starting points, lakes could include small adjacent lakes in their calculations and the same holding for rivers. Mountains are measured from sea level consistently. I would expect the same to hold for Elements, however the discrepancy observed could be due to rounding differences when converting to Kelvin.

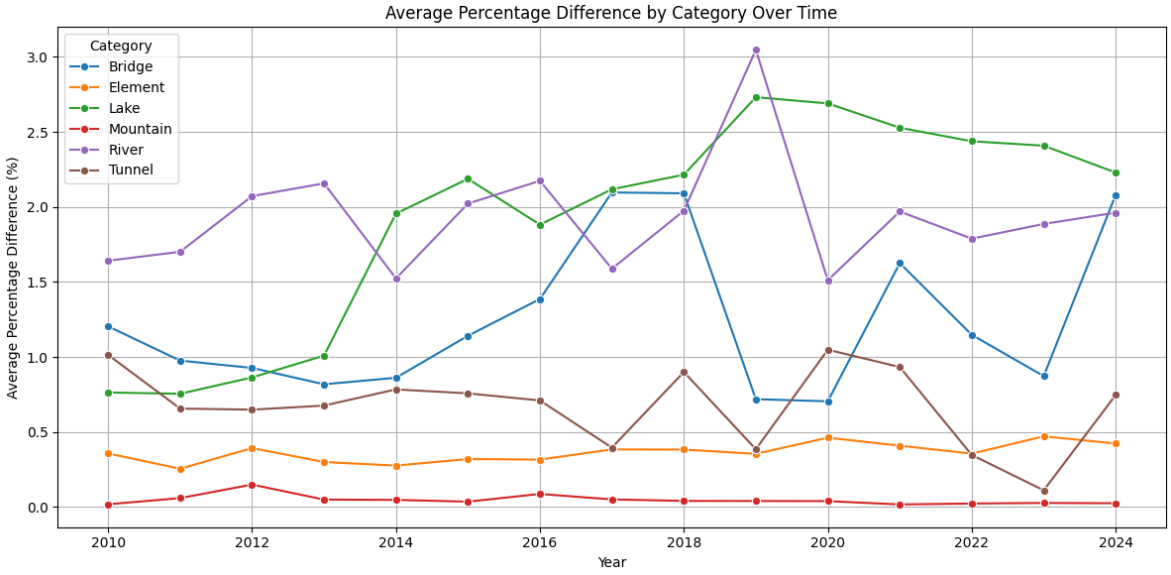

Figure 5: Average Percentage Difference by Category over time

Quickly investigating these trends on a language basis, also reveals a similar trend to the battle context. As can be seen by *Figure 6*, all languages share a similar overall trend, moving upwards and downwards but generally staying inside the same range. The similarity of the movements of all languages is striking, which gives me confidence that our data is accurate and no single language is moving the overall average, that is indeed the differences across the various categories. This finding moves to confirm our previous statements from the battle context, that we cannot see any trend of convergence, hinting that languages tend to stick with their sources and numbers instead

of changing them to convergence with the reference English article. This trend is worrying as it shows the challenges Wikipedia faces to create a unified source of information that is available to everyone. If editors are not comparing and converging their sources overtime, there could still be large information discrepancies for some languages and cultures.

Detailed graphs splitting the categories further into the separate languages over time can be found in the Appendix.

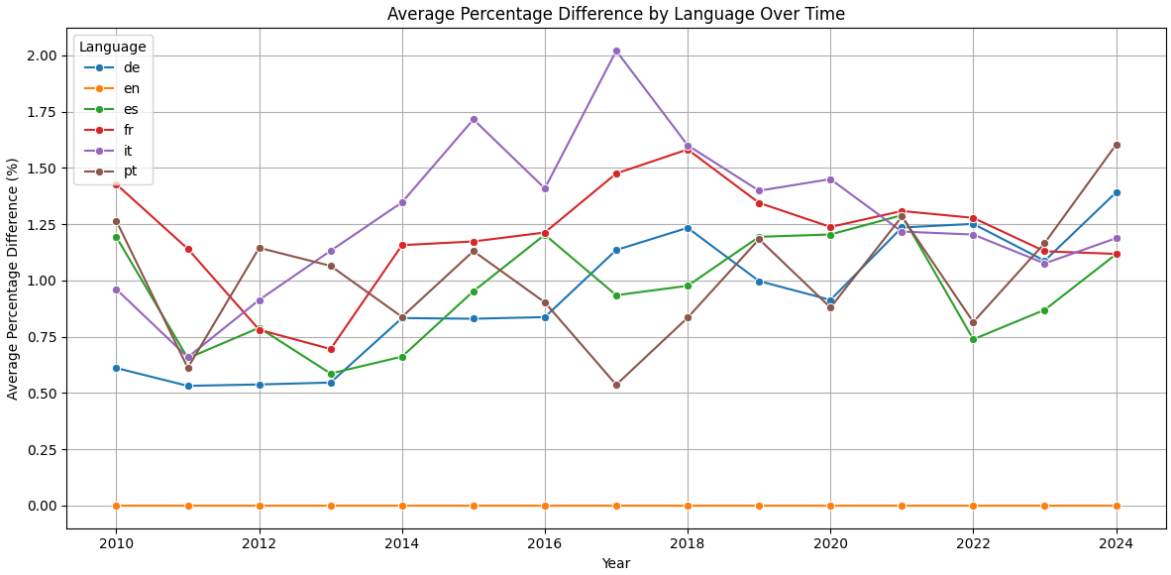

Figure 6: Average Percentage Difference by Language over time in the category expansion

To validate these observations as explained in our methodology, we conducted a one sample t-test. The results can be seen in *Table 10*.

Category	**Bridge**	**Element**	**Lake**	**Mountain**	**River**	**Tunnel**

t_stat	0.490486	-0.241384	5.666641	-4.241933	0.460072	0.061661
p_value	0.649479	0.821127	0.004783	0.013243	0.669362	0.953791
Significant	False	False	True	True	False	False

Table 10: T-Test results category expansion

The t-test results provide further insights into convergence and divergence trends based on the sign of the t-statistic. For bridges, elements, rivers, and tunnels, the non-significant t-statistics confirm our earlier suspicions that these categories exhibit stable differences over time, without any clear convergence or divergence.

However, for mountains and lakes, the results are more revealing. Mountains, with a significant negative t-statistic (-4.24), show a clear convergence trend, indicating that differences are decreasing over time despite being the smallest among all categories. In contrast, lakes, with a significant positive t-statistic (5.67), display a divergence trend, where differences are increasing rather than aligning.

The results overall are similar to the battle context, while in most cases we were not able to see convergence nor divergence, in some cases were. Consequentially even with this context expansion, our conclusion remains the same, information in different languages on Wikipedia does not converge overtime.

6.4 Limitations

While the results align with my previous findings, there are several limitations that need to be acknowledged. Although gathering information across multiple categories provides a broad overview of convergence trends, it inherently limits the sample size within each specific category. For instance, further investigation focusing on a larger dataset of bridges could yield more comprehensive insights.

Additionally, the landmarks analyzed were primarily famous and well-known, which might inherently feature stable and consistent data. It is possible that convergence dynamics are more pronounced in less well-documented or less familiar landmarks, which were not included in this study.

Another limitation is that I only scraped one metric per category. Expanding the analysis to include multiple metrics for each category could provide a more detailed and nuanced understanding of convergence patterns. Finally, as previously mentioned, some of the data extracted might contain errors due to limitations in the scraping process, which could have affected the results.

Acknowledging these limitations highlights opportunities for future research to expand and refine the analysis.