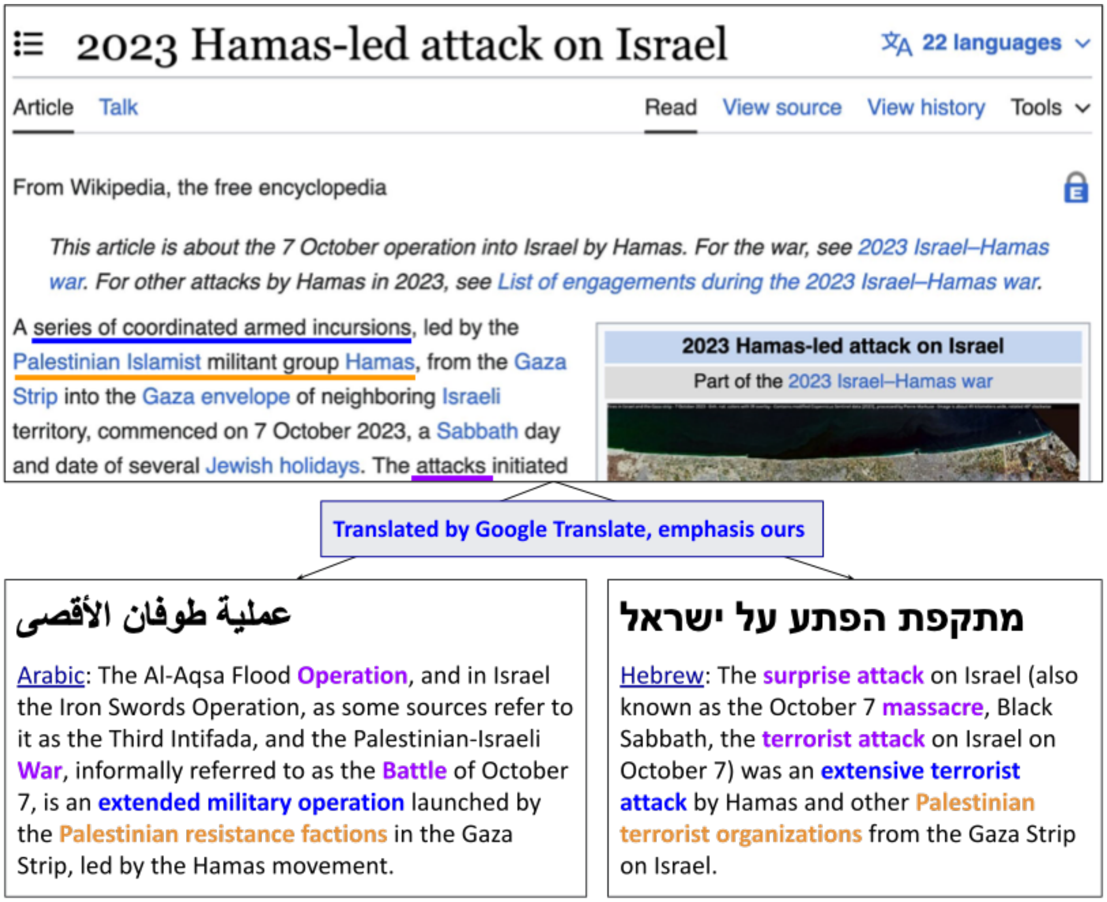
Wikimedia Research Fund 2024 - Stage II proposal

Investigating cross-lingual disparities in representations of conflicts on Wikipedia

Dr Christine de Kock
University of Melbourne



2023 Hamas-led attack on Israel

From Wikipedia, the free encyclopedia

This article is about the 7 October operation into Israel by Hamas. For the war, see 2023 Israel–Hamas war. For other attacks by Hamas in 2023, see List of engagements during the 2023 Israel–Hamas war.

A series of coordinated armed incursions, led by the Palestinian Islamist militant group Hamas, from the Gaza Strip into the Gaza envelope of neighboring Israeli territory, commenced on 7 October 2023, a Sabbath day and date of several Jewish holidays. The attacks initiated

2023 Hamas-led attack on Israel
Part of the 2023 Israel–Hamas war

عملية طوفان الأقصى

Arabic: The Al-Aqsa Flood **Operation**, and in Israel the Iron Swords Operation, as some sources refer to it as the Third Intifada, and the Palestinian-Israeli **War**, informally referred to as the **Battle** of October 7, is an **extended military operation** launched by the **Palestinian resistance factions** in the Gaza Strip, led by the Hamas movement.

מתקפת הפתע על ישראל

Hebrew: The **surprise attack** on Israel (also known as the October 7 **massacre**, Black Sabbath, the **terrorist attack** on Israel on October 7) was an **extensive terrorist attack** by Hamas and other **Palestinian terrorist organizations** from the Gaza Strip on Israel.

Figure 1: The Hebrew, Arabic and English articles on the 2023 Hamas-led attack on Israel.

Abstract

The proposed research explores disparities in the articles describing conflicts across different language versions of Wikipedia using natural language processing. We aim to contribute to research on biases and unfair representations on the platform and produce publicly accessible datasets and software that can be used to discover and mitigate unwarranted variations.

Introduction

Wikipedia, as one of the largest online repositories of information, plays a crucial role in shaping public perceptions of historical and contemporary conflicts. The Wikipedia guideline on translation discourages automated translations, stating that articles on a given subject in different languages are typically edited independently and need not correspond closely in form, style or content. This can result in divergences like those shown in Figure 1 for the October 7 attacks in Israel.

Such differences are a natural conclusion of the No original research policy, which states that “In many cases, there are multiple established views of any given topic. In such cases, no single position, no matter how well researched, is authoritative. It is not the responsibility of any individual editor to research *all* points of view.” While it makes sense to remove a need for “truth arbitration” among editors, it can result in a situation where readers are unaware of the views espoused in other Wikipedia versions, which has implications in terms of bias, fairness and misinformation.

The goal of this work is to develop a metric and computational tool to describe how aligned an article is with equivalent articles in other languages. We address the following research questions:

- **RQ1**: Which metric(s) best capture disparities in cross-lingual representations on conflicts as compared to human impressions?
- **RQ2**: Are there significant differences in the disparities observed between certain pairs of languages or on certain topics?
- **RQ3**: How do these disparities vary over time (during and after a conflict)?’

This project addresses the Wikimedia 2030 Strategic Direction priority of Knowledge Equity by breaking down barriers preventing people from accessing knowledge: specifically, the knowledge and awareness of divergent perspectives on a topic across languages.

Date: July 1, 2024 - June 30, 2025

Related work

Characterising the relationship between pairs of text has been studied from several perspectives in NLP. Articles may differ or align along different axes; as such, it is important to identify exactly what signals are being targeted. We are particularly interested in inter-text differences. This section describes prior work on cross-lingual disparities on Wikipedia and approaches to comparing texts.

Disparities on Wikipedia

Cross-lingual differences on Wikipedia have been studied in prior work. *Callahan and Herring (2011)* manually coded 60 articles on famous individuals from Poland and the USA. Their results pointed to systematic biases in the English articles of famous people from the USA. *Rajcic (2017)* compared articles about famous individuals for availability in different languages and number of views per article, also finding significant divergences. *Field et al. (2022)*, studying biases, propose a matching algorithm to identify a comparison biography page for a given target page, such that the pair are aligned on all but one attribute (e.g., gender) to compare representations on related topics.

Although these studies are able to identify divergences between representations, they concentrate predominantly on language-agnostic evaluations such as citation count or article length. Focusing on language itself, we propose to develop text-based techniques for automatic analysis and quantification of cross-lingual divergences. However, the language-agnostic features can be included in a comparative framework or tool as well.

Modelling text-based disparities

In Figure 1, it can be argued that the same (or similar) basic facts are conveyed: a Palestinian group attacked Israel on 7 October. However, the word choices (“terrorist organisations” vs "resistance factions"; "Operation" vs "massacre") lead to vastly different impressions in the reader. Within NLP, two types of effects are identified, classically referred to as the denotative and connotative dimensions of meaning. Denotation refers to the literal or explicitly stated "surface" meaning and connotation refers to emotions or ideas that are invoked in addition to the literal meaning (*Feng et al., 2013*). These dimensions relate directly to the Wikipedia Neutral Point of View policy (NPOV) that highlights the aim for "an **impartial tone** that document[s] and explain[s] **major points of view**, giving **due weight** for their prominence". In the remainder of this section, we review prior work on modelling connotation and denotation.

Modelling denotation

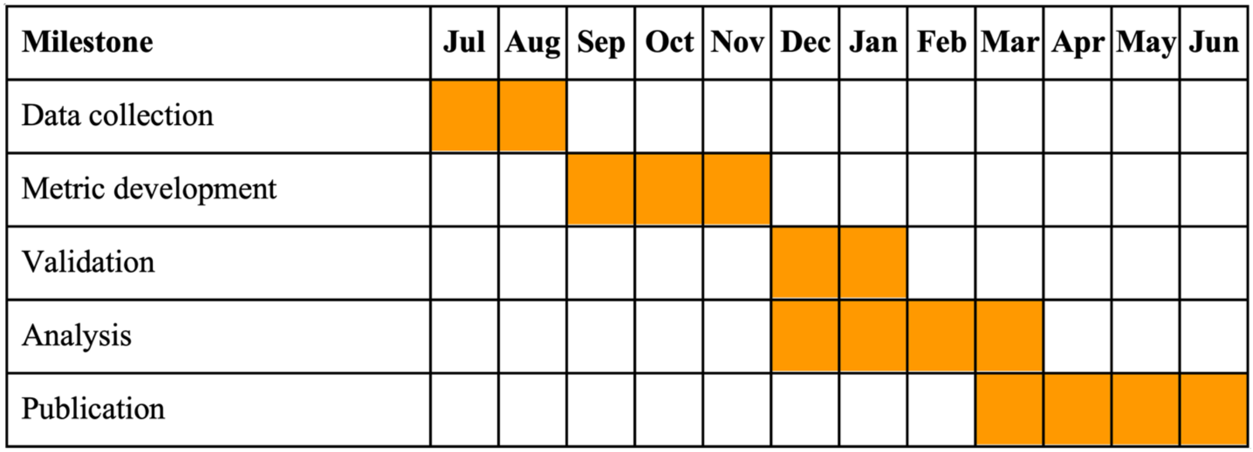
Identifying disparities in denotative information corresponds to identifying where there are differences in the facts or the focus of an article. To study the article focus, one option is to look at the proportional representation of different entities (i.e., mentions of persons, locations and organisations). Neural network transformer-based pretrained models are available to automatically recognise named entities (NER), and methods have been proposed to use the Wikipedia link structure to perform crosslingual mapping of entities (*Tedeschi et al., 2021; Nothman et al., 2013).* The relative frequency of mentions per entity can provide a simple heuristic for the focus of an article. This aligns with the NPOV aim to assign due weight to different topics in an article based on their importance.

Modelling differences in factual statements is more challenging, since Wikipedia articles are not guaranteed (nor intended) to be word-for-word translations; as such, this would require a method for a preprocessing step to align word- or sentence-level text spans that *should* contain the same facts across different documents and languages. This in itself is a highly challenging task and an ongoing area of research (see e.g., *Dixit & Al-Onaizan, 2019*). Given these complexities, article-level analyses may be more suitable in the scope of this project.

However, if a reliable alignment method can be identified, techniques like textual entailment and paraphrasing (*Androutsopoulos & Malakasiotis, 2010*) and semantic textual similarity (*Agirre et al., 2013*) provide relevant denotational information. Relevant systems have also been developed in the fact-checking domain; for example, a recent system by *Cole et al. (2023)* compare paired Wikipedia passages that correspond to factual edits by generating a discriminating question for a given answer span. Factual differences are then defined as pairs where the question is answerable by one of the passages but not the other, or yields different answers. A challenge to using this approach is that article length on Wikipedia can vary substantially between different languages, which may have a confounding effect on the set of overlapping facts.

Modelling connotation

Connotation is by definition a more vague concept than denotation, as it refers to implied or evoked information. It is nevertheless a popular research topic within NLP. Recent systems have been proposed for subjectivity detection (e.g., CLEF-2024 Task 2; *Antici et al., 2024*), promotional tone detection (e.g., our own work: *De Kock and Vlachos, 2022)* and target-dependent sentiment classification (*Hamborg & Donnay, 2021)*. These systems all produce related but specialised perspectives, with the prior two operating at the document-level and the latter at the entity-level. *Field and Tsvetkov (2019)* also focus on entity-centric connotation. Drawing on research in social psychology, they model an entity's potency (i.e., their inferred



Milestone	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun
Data collection	🟧	🟧										
Metric development			🟧	🟧	🟧							
Validation						🟧	🟧					
Analysis						🟧	🟧	🟧	🟧			
Publication									🟧	🟧	🟧	🟧

Table 1: Provisional timeline

strength or weakness in a text), valence (their “goodness”/“badness”), and activity (activeness/passiveness), based on contextualised word embeddings. As these systems are pretrained, we should be able to use them with minimal adaptation to produce useful signals for a connotation comparison, with a key benefit being that many of these systems are indeed pretrained on Wikipedia data. If an entity-targeted metric is used, the need for alignment of content spans (as described for denotation modelling) is mitigated.

Finally, a recent study by *Chen et al. (2022)* combines connotative and denotative features to predict whether a pair of news articles are covering the same news story. They propose 6 similarity indicators that partially overlap with ours: geographical, temporal, narrative, entities, style and tone. However, our task is notably different from theirs in that the articles that we will study are known to be on the same topic.

Methods

Our proposed milestones and timeline are laid out in Table 1, with descriptions of the different milestones provided below. We intend to fund a research assistant for 6 months (full-time equivalent), and the PI will also work on this part-time throughout the year (0.1 FTE, given in-kind).

Data collection

An initial investigation into potential data sources will be conducted, with the objective of identifying articles for comparison. We will use the Wikipedia List of Military Conflicts as a starting point. We refer to the set of different language representations on the same topic as the *article set* and to individual (language-specific) articles in this set as *renderings*. Different historical versions of an article are referred to as *revisions*.

A key question at this stage is which languages to include in the study. We expect to see the largest disparities in the language communities of the parties involved in the conflict; however, it may not be straightforward to identify them analytically. Furthermore, for conflicts where the involved parties speak the same language (e.g. civil wars), cross-lingual divergences cannot arise. Another consideration is the article length: for smaller language communities, renderings may not be sufficiently developed. Finally, certain revisions of these articles may have been deleted, which would necessitate special data access arrangements.

Given the above considerations, we will proceed as follows:

1. Collect conflicts from all subcategories in the List of Military Conflicts.
2. If possible, identify the language groups involved for each conflict. If two or more languages are involved, retrieve their renderings.
3. If it is not possible to identify the involved language groups reliably, or not enough cross-lingual conflicts can be found, retrieve all available article renderings (that is, the full set of available versions of the article in different languages).

We will likely also filter based on article lengths to ensure that there is sufficient text per rendering.

Developing metrics

Given multiple renderings of an article, our aim is to identify and characterise outliers. Our approach will consist of inferring a prototype representation based on an article set using a number of text-based signals. For each rendering, its divergence from the prototype will be given by the distance of its representation from the relevant prototype. We aim to use language-specific technologies where possible, but may use translation-based approaches where there are not adequate resources available for a given language.

As discussed, we propose to distinguish between two dimensions: a denotative perspective (denoted D) and a connotative perspective (denoted C). Given an article rendering a, we define a tuple of vector representations $V_a = (D_a, C_a)$. Each dimension can include multiple signals, such that N different signals s_i would be represented as the concatenation $\{s_1, s_2, \ldots, s_N\}$. In this work, we will focus on signals related to entities (that is, people, places, events and things); however, the framework could be extended to include other types of features, for instance article-level sentiment, article length, and number of citations.

Denotative representation

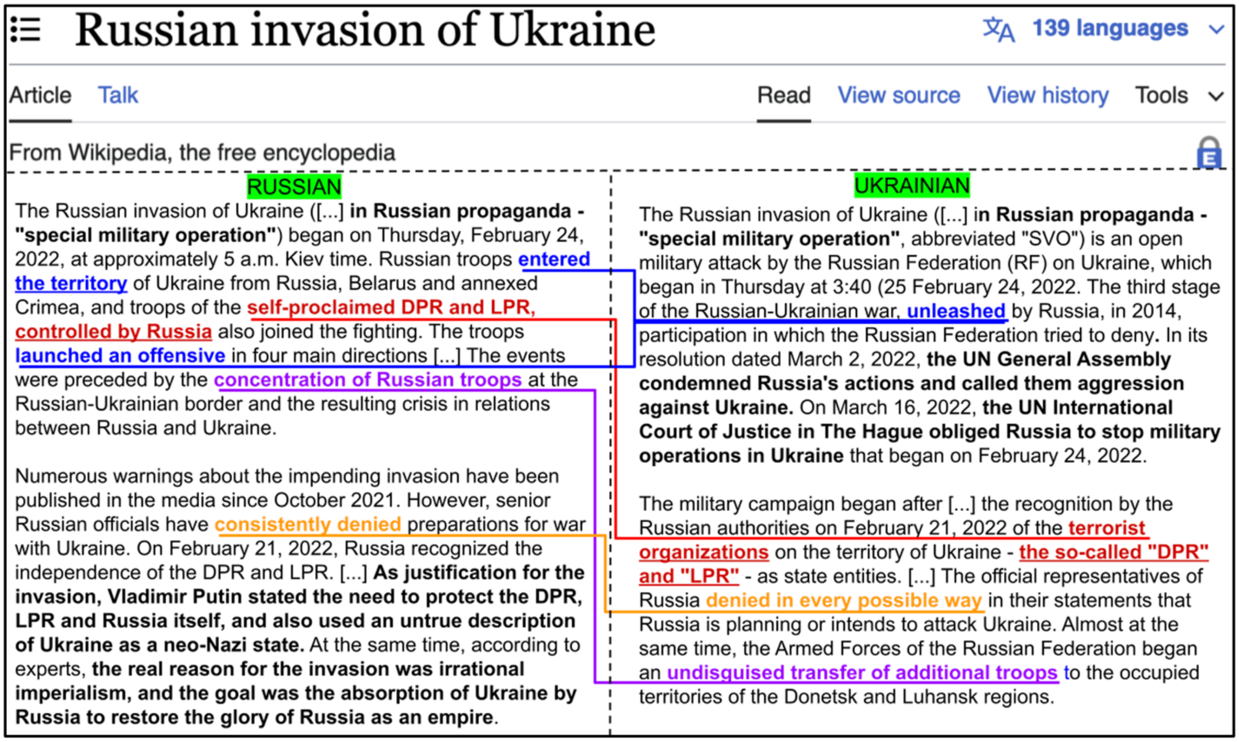
This dimension is intended to capture how much an article attends to various entities and events. For each article set A, we will use NER to find the set of featured entities, applying filters to remove infrequent entities. Wikipedia link structure can be used to map mentions of the same entity across different renderings. If time permits, we may experiment with coreference resolution (*Lee et al., 2017*) to consolidate different mentions of the same entity within a rendering. The resulting set of entities is denoted by E_A. Let $\boldsymbol{c_a}$ represent a vector of entity counts in a for each entity in E_A. Then, we define the denotative representation as $D_a = \frac{c_a}{||c_a||}$.

Connotative representation

Here, we intend to capture differences in the author's attitude towards the entities being discussed, focusing only on the entities that are common in all renderings (denoted u_A). Minimally, the set will contain the topic of the article.

To compute signals for this dimension, we will use the entity-centric connotation framework of Field and Tsvetkov (2019) as described in Section 2. For each rendering a, the power, sentiment and agency scores towards each entity e_i are represented as a 3-tuple $c_{e_i,a} = (P, S, A)$. The full connotative representation C_a would consist of the concatenation of the three factors for all N common entities in the article set: $C_a = (c_{e_1,a}, c_{e_{2\ldots,a}} c_{e_N,a})$.

Should the connotation framework prove challenging to implement or validate, target-dependent sentiment analysis (e.g., the approach



The Russian invasion of Ukraine ([...] **in Russian propaganda - "special military operation"**) began on Thursday, February 24, 2022, at approximately 5 a.m. Kiev time. Russian troops **entered the territory** of Ukraine from Russia, Belarus and annexed Crimea, and troops of the **self-proclaimed DPR and LPR, controlled by Russia** also joined the fighting. The troops **launched an offensive** in four main directions [...] The events were preceded by the **concentration of Russian troops** at the Russian-Ukrainian border and the resulting crisis in relations between Russia and Ukraine.

Numerous warnings about the impending invasion have been published in the media since October 2021. However, senior Russian officials have consistently denied preparations for war with Ukraine. On February 21, 2022, Russia recognized the independence of the DPR and LPR. [...] **As justification for the invasion, Vladimir Putin stated the need to protect the DPR, LPR and Russia itself, and also used an untrue description of Ukraine as a neo-Nazi state.** At the same time, according to experts, **the real reason for the invasion was irrational imperialism, and the goal was the absorption of Ukraine by Russia to restore the glory of Russia as an empire**.

The Russian invasion of Ukraine ([...] in **Russian propaganda - "special military operation"**, abbreviated "SVO") is an open military attack by the Russian Federation (RF) on Ukraine, which began in Thursday at 3:40 (25 February 24, 2022. The third stage of the Russian-Ukrainian war, **unleashed** by Russia, in 2014, participation in which the Russian Federation tried to deny**.** In its resolution dated March 2, 2022, **the UN General Assembly condemned Russia's actions and called them aggression against Ukraine.** On March 16, 2022, **the UN International Court of Justice in The Hague obliged Russia to stop military operations in Ukraine** that began on February 24, 2022.

The military campaign began after [...] the recognition by the Russian authorities on February 21, 2022 of the **terrorist organizations** on the territory of Ukraine - **the so-called "DPR" and "LPR"** - as state entities. [...] The official representatives of Russia **denied in every possible way** in their statements that Russia is planning or intends to attack Ukraine. Almost at the same time, the Armed Forces of the Russian Federation began an **undisguised transfer of additional troops** to the occupied territories of the Donetsk and Luhansk regions.

Figure 2: Wikipedia page of the Russian-Ukrainian war in Russian (left) versus Ukrainian (right), translated using Google Translate and aligned by us.

of *Hamborg et al., 2019*) would be a viable alternative.

Prototyping

Given the denotative and connotative representations of a set of article renderings, we can find a prototype representation P for each dimension by calculating the centroids (the average of the different representations; e.g., for the denotative dimension, $P_{A,D} = \frac{1}{|A|} \sum_{a \in A} D_a$) or the mediod (i.e., the representation with the minimum distance to all other representations in the set, using a vector distance metric like cosine similarity). A benefit of the medoid is that it provides an interpretable output, as we can inspect the prototypical article. The divergence of each rendering from the prototype can then be expressed as its distance to the prototype along each dimension: $Div_a = [d(D_a, P_{A,D}), d(C_a, P_{A,C})]$, where d represents a vector distance metric. We may explore combining these scores as a (possibly weighted) sum over the two dimensions.

Validation

We will use human validation to judge the effectiveness of our divergence metrics. We intend to use a crowd-sourced annotation platform for this purpose. Given a pair of texts, participants will be tasked with judging their similarity along the two dimensions on a Likert scale. Since the task is quite subjective, we would like to have three annotators assess each sample.

Once the human judgments are collected, we can investigate how much the automated dimensions and the individual signals correspond to the intuitive assessments, in answer to RQ1 (Which metric(s) best capture disparities in cross-lingual representations on conflicts as compared to human impressions?).

Analysis
Assuming the metrics can be validated, we can use them to evaluate the magnitude and nature of cross-lingual divergences on Wikipedia. In particular, we are interested in whether any patterns can be observed for different pairs of languages or topics (RQ2) and whether the divergences change over time (RQ3). We hypothesise that the level of divergence would stabilise as the conflict becomes resolved, but that some level of divergence would remain indefinitely in some cases. For example, as we can see from the example of the Russian Ukrainian war in Figure 2, substantial differences still exist in this article, despite the conflict being older than the Israel-Gaza war.

If time permits, we will explore weighing the signals relative to their importance. As mentioned, we expect to see a maximum divergence between the two opposing sides of a conflict, whereas smaller differences are expected between neutral parties. This can serve as noisy labels for a learning algorithm. Since there is a certain level of noise expected in these labels, there is some uncertainty in the quality of outputs this will produce; as such, we would validate the weighted metrics along with the unweighted variants.

We would further like to assess the capability of large language models (LLMs, e.g., GPT-4) to perform this task. As such, we will also report the agreement with human annotations and our proposed metrics. Using LLMs is likely to be prohibitively expensive as a tool for this task; however, it may be useful as a downstream analysis on articles that have been ranked by our system.

Expected output

The following outputs will be produced:

1. Software and dataset: The software and data produced in this project will be made available to the larger research community. If the analysis indicates that large divergences exist, this can form the basis of an editor support tool, an API, and/or banners for readers based on our system. However, further funding and a longer timeline would need to be secured to develop this.

2. Sharing of findings: We will produce two forms of output to communicate our findings. Firstly, we expect to publish an article to an NLP conference, as well as one workshop paper (possibly to WikiConf). Furthermore, we will publish a blog post with the aim of reaching a broader audience. We believe that many readers are not aware of these disparities and that it is important to communicate this.

3. Further grant applications: The broader problem of divergences in cross-lingual representations is of great importance in the media more generally. If we can prove through this project that our approach is viable, we intend to apply for grants to support a project of larger scope. For example, the PI is eligible for funding through Australia's DECRA scheme.

Risks

We have identified the following risks in this project:

1. Our proposal is based on empirical observations, which may not necessarily reflect a broader phenomenon. Though this is a risk, a negative result would also be informative as we could state that examples like Figure 1 are the exception and larger interventions are not required. However, prior work on the topic would suggest that it is indeed a broader phenomenon.

2. If it does exist, we may be unable to quantify the intended discrepancies accurately using existing NLP tools. The topic of bias and fairness is difficult to capture computationally, in part because it is subjective. We mitigate these risks by *(i)* using multiple dimensions (connotation and denotation), and *(ii)* avoiding human annotation except for validation.

3. Political pushback. The topic of conflicts and their representations should naturally be treated as sensitive. For this reason, we avoid any framing that suggests that there is a “true” representation of an event, but rather focus on identifying similarities and differences based on specific characteristics.

Community impact plan

We plan to use two channels to reach audiences beyond academics. Firstly, we will publish a more accessible form of our study as a blog. Secondly, we will present the work at the Wiki Workshop, which attracts the broader Wikipedia community as well as researchers.

Systems developed in this work can form the basis of tools for readers and editors. For example, an article-level banner indicating its level of cross-lingual divergence would inform readers’ perspective while avoiding the need to arbitrate over the truth. For editors, an interface that highlights an article’s divergent parts may be useful. This would require further development and funding, but the system to be produced is a strong starting point.

Finally, we hope that publishing these findings will inform policies and strategic directions. At a governance level, knowing the extent of cross-lingual divergence is important for promoting fair representation.

Evaluation

We plan to use human evaluation to measure the accuracy of our cross-lingual divergence metrics. Regardless of the outcome, the findings will be shared with the broader community. Project success will be indicated by the delivery of the outputs as described above.

Budget

Please see the budget here.

Response to reviewers and meta-reviewers

We thank the reviewers for their constructive feedback. In particular:

1. Reviewers 1 and 2 mention that the project is ambitious for its timeline. We have narrowed its scope by using a narrow but flexible model for disparities, which can be extended if time permits, and by reducing the number of papers to one full paper and one workshop presentation. In addition, a Masters student supervised by the PI has started working on a related project, which may serve as a baseline for this work. Finally, the research assistant who we have in mind for this project (Gisela Vallejo) has been working on cross-lingual divergences in news articles for two years; as such, she will be able to work on this in an effective and efficient manner. For these reasons, we believe the current scope is achievable with the current budget. We recognise that this is not an easy problem to solve, but believe that we can make some progress in this project which can lead to larger efforts and grants.

2. Based on the chairs’ recommendation, we have added Professor Andreas Vlachos as an advisor. He has a strong research record in topics related to misinformation and Wikipedia,

and has published relevant papers with investigators De Kock and Vallejo.

3. We have provided specific details on how we intend to capture the observed differences. We have also reviewed related work on textual relatedness and biases in text.

4. We added political pushback as a potential risk per the recommendation of Reviewer 4.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., & Guo, W. (2013, June). * SEM 2013 shared task: Semantic textual similarity. In *Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity* (pp. 32-43).

Androutsopoulos, I., & Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, *38*, 135-187.

Arkhipov, M., Trofimova, M., Kuratov, Y., & Sorokin, A. (2019, August). Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing* (pp. 89-93).

Callahan, E. and Herring, S. "Cultural bias in Wikipedia content on famous persons," J. Am. Soc. Inf. Sci., vol. 62, no. 10, pp. 1899–1915, Oct. 2011, doi: 10.1002/asi.21577.

Cole, J. R., Jain, P., Eisenschlos, J. M., Zhang, M. J., Choi, E., & Dhingra, B. (2023, May). DiffQG: Generating Questions to Summarize Factual Changes. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (pp. 3088-3101)

Chen, X., Zeynali, A., Camargo, C., Flöck, F., Gaffney, D., Grabowicz, P., ... & Samory, M. (2022, July). SemEval-2022 Task 8: Multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 1094-1106).

De Kock, C., & Vlachos, A. (2022, May). Leveraging Wikipedia article evolution for promotional tone detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 5601-5613).

Dixit, K., & Al-Onaizan, Y. (2019, July). Span-level model for relation extraction. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 5308-5314).

Feng, S., Kang, J. S., Kuznetsova, P., & Choi, Y. (2013, August). Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1774-1784).

Field, A., Park, C. Y., Lin, K. Z., & Tsvetkov, Y. (2022, April). Controlled analyses of social biases in Wikipedia bios. In *Proceedings of the ACM Web Conference 2022* (pp. 2624-2635).

Francesco Antici, Andrea Galassi, Federico Ruggeri, Katerina Korre, Arianna Muti, Alessandra Bardi, Alice Fedotova, Alberto Barrón-Cedeño, *A Corpus for Sentence-level Subjectivity Detection on English News Articles*, in: Proceedings of Joint International Conference on Computational Linguistics, Language Resources and Evaluation (COLING-LREC), 2024

Hamborg, F., Donnay, K., & Merlo, P. (2021, April). NewsMTSC: a dataset for (multi-) target-dependent sentiment classification in political news articles. Association for Computational Linguistics (ACL).

Hecht, B. and Gergle, D. “The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, in CHI ’10. New York, NY, USA: Association for Computing Machinery, Apr. 2010, pp. 291–300. doi: 10.1145/1753326.1753370.

Lee, K., He, L., Lewis, M., & Zettlemoyer, L. (2017, September). End-to-end Neural Coreference Resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 188-197).

Scialom, T., Dray, P. A., Lamprier, S., Piwowarski, B., Staiano, J., Wang, A., & Gallinari, P. (2021, November). QuestEval: Summarization Asks for Fact-based Evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 6594-6604).

Tedeschi, S., Maiorca, V., Campolungo, N., Cecconi, F., & Navigli, R. (2021, November). WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 2521-2533).

Rajcic, N. “Comparison of Wikipedia articles in different languages,” p. 98 pages, 2017, doi: 10.34726/HSS.2017.35937.