Locating Information Gaps and Narrative Inconsistencies Across Languages: A Case Study of LGBT People Portrayals on Wikipedia

Farhan Samir[1,4*] **Chan Young Park**[2] **Anjalie Field**[3]

Vered Shwartz[1,4] **Yulia Tsvetkov**[2]

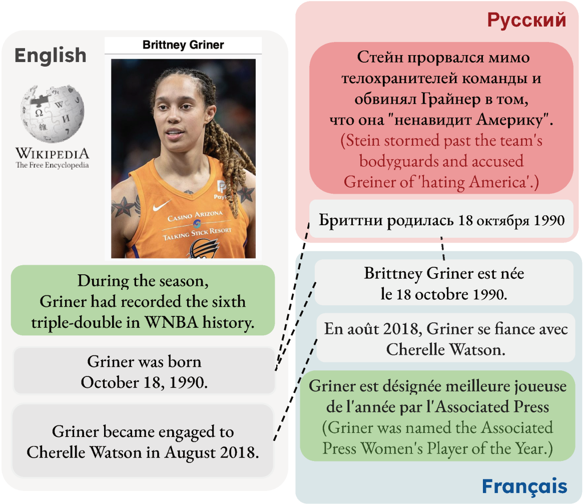
[1] University of British Columbia [2] University of Washington

[3] Johns Hopkins University [4] Vector Institute for AI

fsamir@cs.ubc.ca

Abstract

To explain social phenomena and identify systematic biases, much research in computational social science focuses on comparative text analyses. These studies often rely on coarse corpus-level statistics or local word-level analyses, mainly in English. We introduce the INFOGAP method—an efficient and reliable approach to *locating information gaps and inconsistencies in articles at the fact level, across languages.* We evaluate INFOGAP by analyzing LGBT people's portrayals, across 2.7K biography pages on English, Russian, and French Wikipedias. We find large discrepancies in factual coverage across the languages. Moreover, our analysis reveals that biographical facts carrying negative connotations are more likely to be highlighted in Russian Wikipedia. Crucially, INFOGAP both facilitates large scale analyses, and pinpoints local document- and fact-level information gaps, laying a new foundation for targeted and nuanced comparative language analysis at scale.[1]



During the season, Griner had recorded the sixth triple-double in WNBA history.

Griner was born October 18, 1990.

Griner became engaged to Cherelle Watson in August 2018.

Стейн прорвался мимо телохранителей команды и обвинял Грайнер в том, что она "ненавидит Америку". (Stein stormed past the team's bodyguards and accused Greiner of 'hating America'.)

Бриттни родилась 18 октября 1990

Brittney Griner est née le 18 octobre 1990.

En août 2018, Griner se fiance avec Cherelle Watson.

Griner est désignée meilleure joueuse de l'année par l'Associated Press (Griner was named the Associated Press Women's Player of the Year.)

Figure 1: We propose a method, INFOGAP, to locate fact (mis)alignments in Wikipedia biographies in different language versions. INFOGAP identifies facts that are common to a pair of articles ("Griner was born on October 18, 1990"), and facts unique to one language version ("Griner had recorded the sixth triple-double"; En only) enabling further analysis of information gaps, editors' selective preferences within articles, and analyses at scale across languages, cultures, and demographics.

1 Introduction

Wikipedia has several hundred language editions, a sizeable number of which have more than 100K articles. Despite its "neutral point of view" policy, abundant evidence of content discrepancies across language editions has been well-documented on the platform (e.g., Hecht and Gergle, 2010; Callahan and Herring, 2011; Eom et al., 2015; Wagner et al., 2015; Park et al., 2021). There are numerous motivations for identifying and studying these variations, e.g., identifying content variations and gaps can aid editors in removing social and cultural biases (Field et al., 2022). Alternatively, from a social science perspective, comparative analyses of prominent topics across Wikipedia language editions provides a window into studying cross-cultural differences at scale (Callahan and Herring, 2011).

Existing methods for examining cross-language differences and gaps across Wikipedias rely on aggregate statistics, such as the number of languages an article is available in (Wagner et al., 2015), summary metrics of positive and negative connotations (Park et al., 2021), text complexity measures (Kim et al., 2016; Field et al., 2022), or differences in hyperlink graph structures (Hecht and Gergle, 2010; Laufer et al., 2015). While these metrics are useful for understanding broad trends, they do not facilitate nuanced comparative analysis, failing to inform readers and editors *how* the content they engage with varies across language versions. At the same time, manual fine-grained comparative analyses (e.g., Callahan and Herring, 2011) do not

[*] Work done while visiting the University of Washington.

[1] https://github.com/smfsamir/infogap

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 6747–6762
November 12-16, 2024 ©2024 Association for Computational Linguistics

scale and run the risk of incorporating researchers' biases.

In this work, we propose InfoGap, a highly reliable method for identifying overlaps and gaps across different language articles on the same topic. Our method is composed of two steps: an *alignment* step aimed at aligning facts across different language versions, followed by a *validation* step aimed at determining fact equivalence. InfoGap allows us to automatically identify exact content differences, as illustrated in Fig. 1, enabling both aggregate and fine-grained comparative analyses (§2).

After verifying the accuracy of InfoGap against our manual annotations, we demonstrate its usefulness through a comparative analysis on thousands of multilingual Wikipedia biographies from the LGBTBioCorpus (Park et al., 2021). We find that the coverage of public figures differs substantially across languages (§3). For example, when comparing Russian and English biographies, we find that on average 34% of the content in Russian biographies is not present in their English counterparts.

Critically, as suggested in manual analyses by Park et al. (2021), our automatic analyses at scale identify that many of the bios carry a significantly different implied sentiment towards the figure, depending on the language version that is accessed.[2] Aggregating these sentiment imbalances across 2.7K biographies, we contribute the insight that Russian LGBT biographies share disproportionately more negative sentiment facts with English biographies than positive ones. Overall, InfoGap enables the pinpointing of fine-grained factual and framing distinctions between narratives across languages, aggregates these insights across thousands of articles, and offers tools to identify the specific documents that most clearly highlight these nuances.

2 InfoGap: Identifying Information Asymmetry in Wikipedia Articles

Consider a pair of articles on a topic written in different languages. We call one article E and the other in the pair F. Moreover, we represent E by a series of facts $e_1, \ldots, e_n$ and F similarly as $f_1, \ldots, f_m$. Our method determines for a given fact $e_i \in E$ whether it appears in F ($F \Vdash e_i$) or not ($F \nVdash e_i$). The pipeline is directional, so we can compute both $F \Vdash e_i$ and $E \Vdash f_i$. Without loss of generality, we will describe the procedure for obtaining the labels $F \vdash e_i$, for all e_i. We refer to this as the $E \to F$ direction.

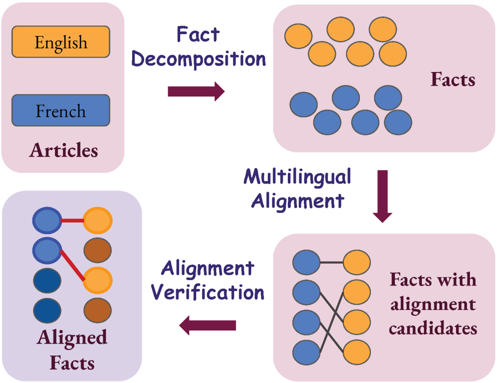


Figure 2: Schematic of the InfoGap procedure. We describe the Fact Decomposition and Multilingual Alignment steps in §2.1, and the Alignment Verification step in §2.2.

Fig. 2 presents an overview of InfoGap. We primarily focus on two steps. First, following Min et al. (2023), we narrow the search space of equivalent facts by aligning a fact in E to facts in F that may convey the same information (Sec 2.1). This allows us to efficiently assess the equivalence between aligned facts (Sec 2.2). We determine the reliability of InfoGap in Section 2.3.

2.1 X-FactAlign: Cross-Lingual Fact Alignment

Fact Decomposition. As a first step, we need to represent an article (e.g. E) as a series of facts ($e_1, \ldots, e_n$). Sentences are suboptimal for this purpose, as they can be overly complex. Instead, following Kamoi et al. (2023), we use GPT-4 (Achiam et al., 2023) for fact decomposition. Differently from Kamoi et al. (2023), who decompose sentences, we decompose entire paragraphs, to provide more context to the model and allow it to resolve co-references. See Appendix A for the prompt.

Fact Representation. In order to determine whether e_i is also conveyed in F, we embed each fact in E and in F using multilingual LaBSE embeddings (Feng et al., 2020). The straightforward way to align facts is by computing the cosine similarity between e_i and each fact $f_j \in F$, aligning e_i to the most similar fact: $\arg\min_j d(\mathbf{e_i}, \mathbf{f_j})$. We find this approach can further be improved by considering the context of the surrounding facts. In the

[2] We focus on the LGBT subset of LGBTQIA+ people on Wikipedia, due to data scarcity for other groups.

following paragraphs, we describe two improvements we made in X-FACTALIGN. First, we restrict the pool of paragraphs in F from which f_j can be retrieved. Second, we apply an adjustment to the computation of $d(\cdot, \mathbf{f_i})$, accounting for the hubness of $\mathbf{f_j}$ (Lazaridou et al., 2015).

Paragraph Alignment. We can partition the facts in E into their paragraphs: $P_E^1, ..., P_E^N$. Similarly for F: $P_F^1, ..., P_F^M$. We represent each paragraph by the set of its facts' embeddings. We then construct a bipartite graph between paragraphs in E and paragraphs in F, adding a directed edge from each paragraph in E, P_E^i, to a paragraph in F, P_F^j such that $j = \texttt{MaxSim}_j \, d(P_E^i, P_F^j)$ (Khattab and Zaharia, 2020). We do the same in the other direction, going from F to E. Removing the direction from the edges, we obtain an adjacency matrix A between the paragraphs so that each paragraph in E is connected to at least one paragraph in F. For a given fact e_i, we can now limit the pool of alignment candidates in F to f_js where the paragraphs of e_i and f_j are adjacent in the graph.

Correcting for Hubness. Given that we are comparing facts from articles on the same topic, directly computing $d(\mathbf{e_i}, \mathbf{f_j})$ can lead to aligning unrelated facts that discuss the same common named entities. In particular, some facts f_j are similar to many other facts e_i, causing a "hubness problem" (Lazaridou et al., 2015; Conneau et al., 2017). To mitigate this, we follow Artetxe and Schwenk (2019) and normalize $d(\cdot, \mathbf{f_j})$ so that it is a function of the semantic density of $\mathbf{f_j}$. The density-normalized distance $D(\mathbf{e_i}, \mathbf{f_i}) = d(\mathbf{e_i}, \mathbf{f_j}) - \text{hubness}(\mathbf{f_j})$. We compute the hubness of f_j by computing the average nearest neighbor distance ($k_{NN} = 5$) between f_j and 50 other facts drawn from paragraphs that are *not* in the adjacency list of the paragraph containing e_i. Overall, this process enables us to retrieve $k = 2$ facts from F that may convey the same information as e_i.

2.2 X-FACTMATCH: Cross-Lingual Fact Matching

With e_i and its aligned facts f_j, we can now answer the question whether a given fact $e_i \in E$ appears in F ($F \Vdash e_i$) or not ($F \nVdash e_i$). We assume that if $F \Vdash e_i$, there exist facts in F that entail e_i. In particular, we can expect these facts to be aligned with e_i. We thus relax the problem of judging whether $F \Vdash e_i$ to whether any of the facts f_j retrieved by X-FACTALIGN entail e_i, i.e. whether $\text{any}(\{\ f_j \Vdash e_i \mid j \in [k]\ \})$.[3]

Language Pair	**#Labeled**	**#Annotated**
En → Fr.	2,213	80
Fr → En.	2,165	
En → Ru	2,832	80
Ru → En	2,435	

Table 1: **Number of facts labeled using INFOGAP** for each language pair and direction, and number of manually annotated facts.

We use entailment as a shorthand for "conveying the same information as" despite a minor deviation from the definition of entailment in linguistics as a strict logical entailment (Heim and Kratzer, 1998), and in NLP as "a human [reading the premise] would typically think that the hypothesis is likely true" (Dagan et al., 2005; Bowman et al., 2015). Our definition is a bit more relaxed and we also consider partial entailment (Levy et al., 2013), i.e., when the most important information in e_i is conveyed by F, allowing the omission of peripheral information. To that end, we don't use existing NLI models. Furthermore, research on cross-language entailment detection is limited (Negri et al., 2012; Rodriguez et al., 2023), and to our knowledge there are no publicly available models that can determine the entailment between a premise and a hypothesis in different languages.

Inspired by Min et al. (2023) and Shafayat et al. (2024) who used GPT-4 to assess the truthfulness of a model-generated fact against a trusted knowledge base, we prompt GPT-4 to compare an English fact to its aligned facts in F. Concretely, we prompt the model with the hypothesis fact e_i and the two immediately preceding facts for context (e_{i-1} and e_{i-2}), along with all of the premise facts f_j and their contexts (f_{j-1} and f_{j-2}). We instruct the model to determine whether e_i can be inferred from any of the f_j ($j \in [k]$). Appendix B presents the prompt that we use for all language pairs. The model's prediction serves as the final label for whether $F \Vdash e_i$.

2.3 Assessing the Reliability of INFOGAP

To assess the reliability of INFOGAP, we evaluate its final results with human annotations. We apply INFOGAP to Wikipedia biographies in English and

[3] We use $[k]$ for $\{1, \ldots, k\}$; see Harvey (2022, p. 11).

Language pair	InfoGap	NLI	Random
En → Fr	**0.81**	0.28	0.62
Fr → En	**0.90**	0.27	0.61
En → Ru	**0.78**	0.52	0.43
Ru → En	**0.88**	0.50	0.35

Table 2: Performance of InfoGap with respect to the manual annotations ($n = 80$ for each language pair), in terms of F_1 score.

French of 10 people, and in English and Russian for 12 people, comprising nearly 10K facts altogether. We draw on biographies from the LGBTBioCorpus (Park et al., 2021), a corpus we analyze in §3 at a larger scale. See Table 1 for a breakdown of the number of facts and Appendix C for the biographies.

We annotated a subset of the facts in each language pair and direction. Given a hypothesis e_i, the retrieved candidate facts f_j from X-FactAlign, and their contexts, we ask the annotator to choose between three options: (1) the hypothesis fact e_i can be inferred from one of the retrieved f_j; (2) the hypothesis fact can be inferred from the article F, but not from the f_j (indicating that X-FactAlign failed to retrieve the correct fact); (3) e_i cannot be inferred from F. We also provide relaxed versions of options (1) and (2), where e_i can be partially inferred from one of the retrieved f_j or partially inferred from F. Concretely, our annotation task closely resembles the X-FactMatch step, with two key differences. First, we provide the annotators with English translations of non-English facts and their contexts, using the NLLB model (Costa-jussà et al., 2022). Second, if a hypothesis fact e_i cannot be inferred from the k facts retrieved by X-FactAlign, we ask the annotator to read the full Wikipedia article F to determine whether e_i can be inferred from it.

One author annotated 80 facts in each language pair and for both directions within each language pair. Another author annotated 40 of those 80 for each language pair. We obtained substantial inter-annotator agreements, with Cohen's $\kappa = 0.71$ for En/Fr and $\kappa = 0.78$ for En/Ru. We thus conclude that the task is relatively unambiguous.

In order to determine the reliability of InfoGap, we compute the predictions against the annotated 80 facts for each language pair. Table 2 presents the F_1 scores that range from 0.78 to 0.9, indicating that the InfoGap pipeline is highly reliable in identifying whether a fact in E is present in F (and vice-versa). Substituting X-FactMatch with a RoBERTa NLI baseline (Liu, 2019) performs significantly worse.[4] The RoBERTa NLI model rarely predicts an entailment label on our dataset, resulting in its poor performance. With the exception of En → Ru, the NLI baseline is outperformed by a classifier that randomly predicts whether the target fact is entailed. InfoGap significantly outperforms both ($p < 0.05$, with a bootstrap percentile test; Efron and Tibshirani, 1994).

3 Using InfoGap to Analyze Asymmetries in LGBT Wikipedia Bios

Having validated the effectiveness of InfoGap, we move onto applying it to answer questions about information gaps in Wikipedia. We focus on identifying content differences between language versions' articles on LGBT public figures. Prior work by Park et al. (2021) identified that English articles on average portrayed these figures with more positive sentiment, as well as greater power and agency (Sap et al., 2017), relative to articles in Russian and Spanish.

To gain further insight into cross-linguistic variation towards LGBT people portrayals, we draw on the LGBTBioCorpus corpus (Park et al., 2021). The corpus comprises $1,350$ biographies of LGBT people, each paired with biographies of non-LGBT people matched on most social attributes except sexual orientation using the matching method introduced in Field et al. (2022). Given that InfoGap enables us to directly compare the content between different language versions of a biography, we contend that our analysis can provide a more direct characterization of differences in LGBT bios. Specifically, we look at En, Fr, and Ru Wikipedias. We consider the following research questions:

RQ$_1$: To what extent does factual knowledge differ across language versions of the same bio (Sec 3.2)?

RQ$_2$: Does a person's affiliation with the LGBT community have an effect on the information gap in their bios (Sec 3.3)?

RQ$_3$: Can we use InfoGap to identify sections to remediate (Sec 3.4)?

These questions are intentionally ordered from high-level (language-level) to low-level (individual- and fact-level) to demonstrate that InfoGap enables both high-level quantitative analyses and ef-

[4] The baseline is available on HuggingFace as cross-encoder/nli-roberta-base.

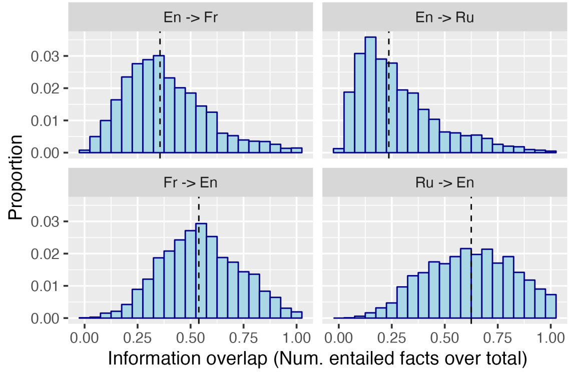


Figure 3: **Distribution of information overlaps for LGBTBioCorpus**. Top: Distribution over the percentage of facts in `En` biographies also found in their `Fr` and `Ru` counterparts. Bottom: Distribution over the percentage of facts in `Fr` and `Ru` biographies also found in their English counterparts. $N = 2,700$ biographies. In general, `En` biographies contain more facts that are exclusive to `En`.

ficient low-level descriptive analyses. We start by providing the implementation details in Section 3.1 before answering each of the research questions.

3.1 Implementation Details

The LGBTBIOCORPUS is significantly larger than the small set of 22 biographies from Section 2, leading to high cost and runtime when applying INFOGAP. Parsing Alan Turing's biography with INFOGAP alone, for example, can require more than 100K GPT-4 tokens.[5] We thus use the GPT-4 predictions to finetune smaller models that are more efficient. Specifically, we use `flan-t5-large` (Chung et al., 2024) for both directions of the `En`/`Fr` pair and `mt5-large` (Xue et al., 2020) for both directions of the `En`/`Ru` pair. We find that the T5 variants perform well at modeling the annotations from §2, obtaining macro-averaged F1 scores of 0.90 (`En` → `Ru`, `Ru` → `En`) and 0.87 (`En` → `Fr`, `Fr` → `En`). We provide fine-tuning hyperparameters and validation set performances in Appendix D.

3.2 RQ$_1$: Information Gaps in Bios

Previous smaller-scale manual qualitative analyses showed that people portrayals differ systematically across language versions (Callahan and Herring, 2011). However, this was challenging to quantify as it would be unreasonably laborious to manually count the number of overlapping facts between language versions of an article. Equipped with INFOGAP, we can for the first-time quantify variance in information overlap between language versions of Wikipedia biographies at scale. Specifically, we consider the INFOGAP predictions for the entire corpus (LGBT and non-LGBT bios).

Fig. 3 visualizes the distribution of the number of facts that can be found in both language versions of the same bio. In the top-left subfigure, we show a histogram of the amount of information in the `En` article that can also be found in the `Fr` article (`En` → `Fr`). The median of the distribution is 0.35, indicating that for half of the biographies, only 35% of the information in the `En` article can be found in the `Fr` article. By comparison, the median of the `Fr` → `En` distribution is 0.55, much higher than the median of the `En` → `Fr` distribution, indicating that `En` biographies contain more unique information than their `Fr` counterparts.

Considering `En`/`Ru`, we find that `En` articles contain significantly more unique information than `Ru` counterparts, with the median `En` → `Ru` overlap being 0.23. Much of the information in the `Ru` articles meanwhile can be found in the `En` articles, with a median overlap of 0.66 for `Ru` → `En`.

We also note that the INFOGAP ratios reflect the well known "local heros" effect, where biographies of individuals whose nationality matches the language of the article tend to have greater coverage, length, and visibility (Callahan and Herring, 2011; Field et al., 2022; Hecht and Gergle, 2010; Oeberst and Ridderbecks, 2024). When the nationality of the person is Russian (66 people), the median `En` → `Ru` overlap increases to 0.29 (+5%) while the `Ru` → `En` overlap decreases to 0.44 (−22%). Similarly for French (148 people), the median `En` → `Fr` overlap increases to 0.52 (+17%), while the `Fr` → `En` overlap decreases to 0.29 (−26%). Overall, this result indicates that there are large scale disparities in information overlap ratios across language versions, building on Callahan and Herring's (2011) early analysis.

3.3 RQ$_2$: Effect of LGBT Affiliation on Information Gaps

Given the large scale differences in content between language versions, we turn to the question of whether LGBT people biographies exhibit different patterns of information overlap compared to non-LGBT people. For example, do Russian biographies tend to include or exclude certain types of information depending on whether the biography is about an LGBT person? To investigate this question, we fit a binomial regression model to de-

[5] We used `gpt-4-1106-preview`. At the time of writing, this API cost $30.00 USD/1M tokens.

Language pair	Factor	Coefficient
En → Ru	conn_pos**	-0.16
	conn_neg**	-0.18
	is_lgbt**	0.10
	conn_pos:is_lgbt	0.04
	conn_neg:is_lgbt	0.03
En → Fr	conn_pos**	-0.07
	conn_neg	0.01
	is_lgbt**	-0.05
	conn_pos:is_lgbt	0.01
	conn_neg:is_lgbt**	0.06
Ru → En	conn_pos**	-0.14
	conn_neg**	-0.51
	is_lgbt**	0.26
	conn_pos:is_lgbt	0.04
	conn_neg:is_lgbt**	0.25
Fr → En	conn_pos**	-0.07
	conn_neg**	-0.14
	is_lgbt	0.00
	conn_pos:is_lgbt	0.03
	conn_neg:is_lgbt**	0.09

Table 3: **Mean of posterior distribution of regression coefficients**. ** indicates that 95% posterior credible interval for the coefficient does *not* contain zero.

termine which factors contribute to the inclusion of En facts in the corresponding Fr or Ru bios, and vice versa.

Features. Table 3 displays the features we use, along with their coeffcient estimates.[6] Naturally, we include a binary feature is_lgbt indicating whether the bio is of an LGBT person. Crucially, we also need to consider the connotation of facts in the English article. Park et al. (2021) found that English LGBT bios were portrayed with greater sentiment, power, and agency than Russian bios. However, this prior work cannot shed light on whether the difference in sentiment is due to Russian bios including negative sentiment facts that are not in the English bios, excluding positive sentiment facts from the English bios, or both. We directly address this question using INFOGAP.

To determine the connotation of a fact e_i, we have to consider the context in its original sentence, which requires mapping between a fact and a sentence. To map facts to their original sentences (e.g., "*Cook is on the board of directors of Nike*" → "*Cook is also on the boards of directors of Nike, Inc. and the National Football Foundation*"), we use forced alignment; see Appendix E.

We obtain connotation predictions at the sentence level by prompting a language model to determine whether a given sentence (in the context of the two prior sentences) portrays the subject of the biography in a positive, negative, or neutral light. Similar to the distillation of the INFOGAP process to a smaller model (§3.1), we first obtain connotation labels using GPT-4 for a smaller set of bios, and use those labels to finetune a smaller model for scaling to the full LGBTBIOCORPUS (see Appendix E for details, including human annotation of the connotation labels). We use the sentence-level connotation label as the label for its constituent facts. Table 4 presents this label distribution.

Language	Positive	Neutral	Negative
English	0.434	0.488	0.077
French	0.442	0.455	0.102
Russian	0.327	0.658	0.014

Table 4: **Distribution of implied sentiment** about biography subjects for En, Fr, and Ru articles.

Regression Model. Without loss of generality, consider modeling the amount of information in the En bios that is also present in the Fr bios, i.e., the En → Fr direction. To perform our binomial regression, we first partition each bio into three sets – positive, negative, and neutral facts. Each partition represents one datapoint for fitting the regression model, so each bio contributes three datapoints. Within each of these three partitions, some facts will also be present in F, while others will be exclusive to E. We model this using a bayesian binomial regression model (McElreath, 2018):

$$\texttt{overlap} \mid N_p \sim 1 + \texttt{conn} + \texttt{is\_lgbt} + \texttt{is\_lgbt:conn}$$

where conn gets the value of either conn_pos, conn_neg, or conn_neutral, depending on the input partition, N_p is the number of facts in the current partition of E, and overlap is the number of facts that are also in F (at most N_p). is_lgbt:conn is an interaction between the two categorical variables. See Appendix F for model-fitting details.

Connotation is a predictive factor. Listed in Table 3, our results indicate that connotation is a predictive factor in nearly all language pairs and directions considered, except conn_neg in En → Fr. Further, the polarity of the conn_pos and conn_neg factors is always negative, suggesting that polarized facts tend to be included in lower rates than neutral facts, which are more agreeable across language versions. To ground the effect

[6] We also fit models with the covariates of gender, nationality, and ethnicity. Including these covariates did not change the estimates for the features in Table 3, so we omit them for clarity.

size of the coefficients, we can simulate predictions from the regression model. For example, a value of -0.07 for `conn_pos` in the `En` → `Fr` model indicates that 34.4% of the positive facts in `En` are included in the `Fr` bios, compared to 36.6% of the neutral facts.

Negative connotation facts are disproportionately included in Russian LGBT bios. Considering Russian biographies, we draw from the large coefficient value of the `is_lgbt` feature that facts from the English article are more likely to be referenced when the article is about an LGBT public figure. Moreover, from the `is_lgbt:conn_neg` interaction, we find that negative facts are more likely to be referenced than positive ones. To quantify the size of this effect, we simulate posterior predictions from the binomial regression model. We find an average 50.87% of negative Russian facts are shared with the English biographies when they describe an LGBT public figure, whereas only 38.53% of negative facts are shared with English bios when they are non-LGBT.

3.4 RQ$_3$: Identifying Sections to Remediate

Our analysis in Sec 3.3 revealed that facts carrying a more polarizing (non-neutral) connotations are less likely to be shared across language versions. This suggests that many biographies may carry a significantly different overall connotation, depending on the language version in which they are read. Unlike the manual analysis performed in prior works (e.g., Park et al., 2021; Callahan and Herring, 2011, among others) to identify such language-version imbalanced content, InfoGap can automatically locate imbalanced content. Park et al. (2021) in particular focused on bios where the subject was portrayed with a more negative implied sentiment. Here, we focus on a different aspect of sentiment differences: the *omission* of content with positive implied sentiment from one language version.

Specifically, we follow these steps to identify imbalanced content: First, we identify bios from the LGBTBioCorpus where a high rate of positive facts are excluded from one language version compared to another.[7] Next, we introduce a method to identify positive life events that are missing in that language version. We provide a formal argument demonstrating that our InfoGap based method for identifying missing events is highly accurate. Finally, we conclude with examples of findings.

Step 1. Identifying biographies with imbalanced implied sentiment. Consider a pair of articles E and F written in different languages, and suppose we wanted to find bios where F omitted positive content at a high rate. We conduct a hypothesis test to determine whether the number of positive facts included in both languages is significantly lower than expected based on the overlap rate of neutral facts. Concretely, we perform a bayesian hypothesis test based on the `BetaBinomial` distribution; we provide complete details in Appendix G.

Our test identifies 274 imbalanced LGBT biographies when considering `En` → `Ru` and 236 when considering `En` → `Fr`. We can follow the same procedure for finding English biographies that comparatively lack positive information, when compared to their French and Russian counterparts. We find 105 and 199 biographies in the `Ru` → `En` and `Fr` → `En` direction, respectively.

Step 2. Identifying events that are unique to a language version. Having identified biographies that could benefit from remediation, we next focus on finding the positive-connotation carrying content that is missing from one language version. By comparison to Park et al. (2021), who could only analyze 10 biographies for identifying imbalanced content, we can leverage InfoGap to identify imbalanced content at scale within the subset of biographies we identified. We focus on finding positive connotation *events* – longer collections of facts that are thematically related – rather than individual isolated facts since the omission of a whole event is more egregious. Practically speaking, we search for paragraphs $\mathcal{V} = e_1, \ldots, e_{N_V}$ where all facts in the paragraph are missing from F:

$$M = \{\mathcal{V} \in E | \, \text{all}(\{ \, F \nVdash e_i \mid i \in [N_V] \, \})\} \quad (1)$$

We then select a subset of M: paragraphs containing at least one positive connotation fact.

InfoGap is highly effective at identifying missing events. Consider an event $\mathcal{V} = e_1, \ldots, e_{N_V}$ that is described in article E. Suppose that InfoGap predicted that $\mathcal{V}$ is not covered by F, that is that F does not entail any of the events in $\mathcal{V}$: $F \nvDash e_i, i \in [N_V]$. For InfoGap to be wrong, i.e., $\mathcal{V}$ is actually present in F, there needs to exist a subset of facts $e_{i(1)}, \ldots, e_{i(k)} \in \mathcal{V}$, for

[7] It is also effective at identifying individual facts present in one language version but absent in another (see §2), but for this analysis we consider collections of multiple facts.

$k = p \cdot N_V$ $(0 < p < 1)$, that are entailed by F: $F \models e_{i(j)}, j \in [k]$. We can bound the probability of this error.

Proposition 1 (Error Bound of Event Identification through InfoGap)**.** *The probability of* INFOGAP *making k errors is $\leq \exp(-2(1-\epsilon)^2 k)$, where ϵ is the error rate of the classifier when it predicts $F \not\models e_i$.*

Proof. Given that the error rate of the classifier is ϵ, the expected number of errors for k predictions is $\epsilon \cdot k$. However, the classifier made k mistakes, so we have made $\epsilon \cdot k + (1-\epsilon) \cdot k$ errors, an additive factor of $t = (1-\epsilon) \cdot k$ more mistakes than expected. By Hoeffding's inequality (Appendix H), where we supply our expected value $\mu = \epsilon \cdot k$ and the deviation from the expected value of $t = (1-\epsilon) \cdot k$, we obtain an upper bound of:

$$\leq \exp\left(-2(1-\epsilon)^2 k^2 / k\right) = \exp(-2(1-\epsilon)^2 k).$$

□

The significance of this claim is that it is rare for the INFOGAP classifier to make a large number (k) of mistakes when the error rate is ϵ (where $\epsilon << 1$). Moreover, the probability of mistakes decreases very quickly in the accuracy of the classifier and the number of facts in $\mathcal{V}$ that were predicted to not be entailed by F. As we showed empirically in Section 2, the INFOGAP classifier is reliable (low ϵ) and thus it has a strong capacity to find events that are only described in one language version.[8]

Findings. In Table 5, we demonstrate positive events that are unique to one language version when compared to another. We find that Chelsea Manning's Fr page describes praise for her whistleblowing during the Afghanistan war. The Fr page also discusses her whistleblowing on the Abu Ghraib prison conditions (Hersh, 2004). Conspicuously, both events are omitted from the En page, despite the En page being otherwise longer. American perception of this instance of whistleblowing skewed negative (Pew Research Center, 2010), which may have played a role in the disparities between the En and Fr pages.

We also find Tim Cook's Ru page – but not his En page – makes note of his fundraising initiative to defend Ukraine in the current Russo-Ukranian war. It is unsurprising that it appears in the Ru page, as it directly pertains to Russia. However, the omission of this fact from the En page is remarkable, since it had received some media attention from American outlets (Clark and Schiffer, 2022). One reason for this omission may be that there is a partisan divide on US involvement in the war (Pew Research Center, 2024). This fact may not have been included in En to maintain a veneer of neutrality.

It is important to note however that Wikipedia's Neutral Point of View policy advocates for a balanced representation of views (Matei and Dobrescu, 2011), rather than outright filtering or censorship. Our findings raise questions about the degree to which a cross-linguistically consistent "Neutral Point of View" is realizable. INFOGAP enables studying these cross-linguistic differences in portrayals of public figures at scale.

4 Related Work

Automated comparison of multilingual Wikipedia articles. We contribute to a large body of work on understanding differences between language versions of Wikipedia. Hecht and Gergle (2010) also compare Wikipedia language versions and consider their information gaps, and later develop a web tool to bridge these multilingual gaps (Bao et al., 2012). However, their evaluation is at a higher level of abstraction: they look at whether or not two language versions have on a topic. By comparison, we compare content differences between two language versions on the same topic. Duh et al. (2013) considered a pipeline similar to INFOGAP for the task of keeping multilingual Wikipedia documents consistent. However, their pipeline used embedding similarity; in early experiments, we found that using embedding similarity for identifying potential entailments performed very poorly (relative to X-FACTMATCH). Massa and Scrinzi (2012) created a web tool that permits visual comparison of Wikipedia articles in two different languages. Rodriguez et al. (2023) also perform comparative analyses across language versions in Wikipedia. However, they consider more fine-grained content differences between pairs of the most closely related paragraphs between different language versions' article on a topic. Their method was not designed for computing the overall article level overlaps and differences of the form we demonstrate in Fig. 3.

[8] One shortcoming of this argument is if F discusses a completely different aspect of the event $\mathcal{V}$ than E. We conjecture that this is unlikely since both articles should at least contain the central propositions about the event.

Pair	Person	Events
En ✗, Ru ✓	Tim Cook	In 2022, following Russia's invasion of Ukraine, **Tim Cook** called on the company's employees to donate to help Ukraine. **Apple's CEO** announced the decision to suspend sales of equipment in Russia and also said that the company would triple the amount of donations made by employees to support Ukraine, and this would be retroactive to February 25, 2022.
En ✗, Fr ✓	Chelsea Manning	"Ron Paul, a leader of the libertarian movement within the Republican Party, endorsed **Manning** on April 12, 2013, stating that **Manning** had done more for peace than Obama—referring to Obama's 2009 Nobel Peace Prize win: "While President Obama was initiating and expanding unconstitutional wars abroad, **Manning**, whose actions caused exactly zero deaths, was shining a light on the truth behind those wars. Which of the two has done more for peace is clear."
En ✓, Fr ✗	Caster Semenya	In 2010, the British magazine *New Statesman* included **Semenya** in its annual list of "50 People That Matter" for unintentionally instigating "an international and often ill-tempered debate on gender politics, feminism, and race, becoming an inspiration to gender campaigners around the world"
En ✓, Ru ✗	Ada Colau	During her period as mayor of Barcelona, **Colau** has maintained a political stance against activities that are susceptible of contributing to greenhouse gas emissions and air pollution. She has repeatedly opposed the expansion of El Prat airport and the use of private cars in the city, and has pushed regional authorities to restrict the number of cruise ships arrivals in Barcelona. In 2020 she declared a "climate emergency", advocating limiting the consumption of meat at schools and forbidding councillors from using the Barcelona-Madrid air shuttle.

Table 5: Examples of events from biographies that contain a large number of positive facts that are only contained in one language version of the article relative to another. We provide translations (Google Translate) for the first two rows, rather than the original French and Russian content.

Case studies on cultural differences in multilingual Wikipedia. We highlight two studies that were not mentioned elsewhere in this work. Hickman et al. (2021) analyze how a boundary dispute over Kashmir between India and Pakistan is represented in English, Hindi, and Urdu Wikipedia, analyzing how the Neutral Point of View principle is upheld. They find there is a sizeable number of cross-language editors between Urdu and English, as well as Hindi and English, but not Urdu and Hindi, attributing this to the popularity of English Wikipedia. Kharazian et al. (2024) studied how the Croatian language version of Wikipedia was usurped by a small group of editors who aimed to promote far-right bias and disinformation about various Croatian political figures, groups, and events. This bias was apparent when comparing the Croatian articles to Serbian and English ones.

5 Conclusion

We presented InfoGap, a reliable method for efficient comparative analysis between two narratives on the same topic written in different languages. We deployed the method to discover differences in LGBT people's portrayals, locating shared facts, as well as information gaps and inconsistencies across 2.7K English, Russian, and French Wikipedia biography pages. InfoGap can be directly applied beyond analyzing differences in multilingual Wikipedia biographies. Analyzing variation in topic coverage is at the heart of much research in the social sciences, from understanding media manipulation strategies (Field et al., 2018), to analyzing differences in argumentation from different stances in a contentious debate (Luo et al., 2020), to analyzing quotation patterns in partisan media (Niculae et al., 2015). Overall, our research lays the foundation for enabling targeted, nuanced textual comparative analyses at scale.

6 Limitations

Applicability to specialized domains. Our method relies on the language understanding abilities of the underlying language model (GPT-4 in our case). While we were able to achieve high accuracy on the LGBTBioCorpus, it is not guaranteed that similarly high-accuracy can be achieved if we were to apply InfoGap to more specialized domains, where domain expertise may be required to assess the equivalence of two facts in different languages, such as comparing Wikipedia articles concerning scientific topics.

Connotation is subjective. In Section 3 we investigated the effect of connotation on the inclusion of facts in different language versions. We acknowledge that connotation is fairly subjective, and may depend on a reader's stance towards the topic and their cultural background. To ensure a high degree of replicability of our results, we have released our all of the finetuned models we applied in §3, including the connotation models.

Ablations of InfoGap components. We did not perform ablations of the components of the X-FactAlign step in the InfoGap pipeline (§2). Our aim was to demonstrate that high-quality automatic cross-lingual comparative analysis is not only possible (§2.3) but provides considerable benefits in downstream analyses (§3). We will perform thorough ablations with a larger number of annotated samples in future work.

7 Ethical Considerations

Data. The dataset used in this study, LGBTBio-Corpus, is publicly available.

Models. We used language models to make classification predictions, limiting their ability to generate offensive content. We used a closed-source model, GPT-4, which entails high costs, and may not be suitable for applying our method to different datasets, especially those containing private information. The distilled version of InfoGap, which uses open-source models, addresses both concerns.

8 Acknowledgements

We thank Shreya Prakash for advice on our regression analyses and hypothesis testing. We thank Miikka Silfverberg and Jai Aggarwal for helpful feedback on the manuscript. FS is supported by an NSERC PGS-D scholarship. VS is supported by the Vector Institute for AI, the CIFAR AI Chair program, and NSERC. We gratefully acknowledge support from the National Science Foundation under CAREER Grant No. IIS2142739, and NSF grants No. IIS2125201 and IIS2203097.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. 2012. Omnipedia: bridging the wikipedia language gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1075–1084.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Paul-Christian Bürkner. 2017. brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80:1–28.

Ewa S Callahan and Susan C Herring. 2011. Cultural bias in wikipedia content on famous persons. *Journal of the American society for information science and technology*, 62(10):1899–1915.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Mitchell Clark and Zoë Schiffer. 2022. Read Tim Cook's email to employees on Ukraine. *The Verge*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Kevin Duh, Ching-Man Au Yeung, Tomoharu Iwata, and Masaaki Nagata. 2013. Managing information disparity in multilingual document collections. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(1):1–28.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. Chapman and Hall/CRC.

Young-Ho Eom, Pablo Aragón, David Laniado, Andreas Kaltenbrunner, Sebastiano Vigna, and Dima L Shepelyansky. 2015. Interactions of cultures and top people of wikipedia from ranking of 24 language editions. *PloS one*, 10(3):1–27.

Fangxiaoyu Feng, Yinfei Yang, Daniel Matthew Cer, N. Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. In *Annual Meeting of the Association for Computational Linguistics*.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.

Anjalie Field, Chan Young Park, Kevin Z Lin, and Yulia Tsvetkov. 2022. Controlled analyses of social biases in wikipedia bios. In *Proceedings of the ACM Web Conference 2022*, pages 2624–2635.

Nick Harvey. 2022. A first course in randomized algorithms.

Brent Hecht and Darren Gergle. 2010. 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*, pages 291–300.

Irene Heim and Angelika Kratzer. 1998. *Semantics in Generative Grammar*. Blackwell.

Seymour M. Hersh. 2004. Torture at abu ghraib. *New Yorker*.

Molly G Hickman, Viral Pasad, Harsh Kamalesh Sanghavi, Jacob Thebault-Spieker, and Sang Won Lee. 2021. Understanding wikipedia practices through hindi, urdu, and english takes on an evolving regional conflict. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–31.

Matthew D Hoffman, Andrew Gelman, et al. 2014. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore. Association for Computational Linguistics.

Zarine Kharazian, Kate Starbird, and Benjamin Mako Hill. 2024. Governance capture in a self-governing community: A qualitative comparison of the croatian, serbian, bosnian, and serbo-croatian wikipedias. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–26.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Suin Kim, Sungjoon Park, Scott A Hale, Sooyoung Kim, Jeongmin Byun, and Alice H Oh. 2016. Understanding editing behaviors in multilingual wikipedia. *PloS one*, 11(5):e0155305.

Paul Laufer, Claudia Wagner, Fabian Flöck, and Markus Strohmaier. 2015. Mining cross-cultural relations from wikipedia: a study of 31 european food cultures. In *Proceedings of the ACM web science conference*, pages 1–10.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proc. ACL*.

Omer Levy, Torsten Zesch, Ido Dagan, and Iryna Gurevych. 2013. Recognizing partial textual entailment. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 451–455, Sofia, Bulgaria. Association for Computational Linguistics.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3296–3315, Online. Association for Computational Linguistics.

David JC MacKay. 2003. *Information theory, inference and learning algorithms*. Cambridge university press.

Paolo Massa and Federico Scrinzi. 2012. Manypedia: Comparing language points of view of wikipedia communities. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, pages 1–9.

Sorin Adam Matei and Caius Dobrescu. 2011. Wikipedia's "neutral point of view": Settling conflict through ambiguity. *The Information Society*, 27(1):40–51.

Richard McElreath. 2018. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Michael Mitzenmacher and Eli Upfal. 2017. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa Bentivogli, and Danilo Giampiccolo. 2012. Semeval-2012 task 8: Cross-lingual textual entailment for content synchronization. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 399–407, Montréal, Canada. Association for Computational Linguistics.

Vlad Niculae, Caroline Suen, Justine Zhang, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Quotus: The structure of political media coverage as revealed by quoting patterns. In *Proceedings of the 24th International Conference on World Wide Web*, pages 798–808.

Aileen Oeberst and Till Ridderbecks. 2024. How article category in wikipedia determines the heterogeneity of its editors. *Scientific Reports*, 14.

Chan Young Park, Xinru Yan, Anjalie Field, and Yulia Tsvetkov. 2021. Multilingual contextual affective analysis of lgbt people portrayals in wikipedia. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 479–490.

Pew Research Center. 2010. Most say wikileaks release harms public interest. Technical report.

Pew Research Center. 2024. Views of ukraine and u.s. involvement with the russia-ukraine war. Technical report.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2015. Connotation frames: A data-driven investigation. *arXiv preprint arXiv:1506.02739*.

Juan Diego Rodriguez, Katrin Erk, and Greg Durrett. 2023. X-parade: Cross-lingual textual entailment and information divergence across paragraphs. *ArXiv*, abs/2309.08873.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2329–2334.

Sheikh Shafayat, Eunsu Kim, Juhyun Oh, and Alice Oh. 2024. Multi-fact: Assessing multilingual llms' multi-regional knowledge using factscore. *arXiv preprint arXiv:2402.18045*.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Proceedings of the international AAAI conference on web and social media*, volume 9, pages 454–463.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

En	Please breakdown the following paragraph into a list of independent facts. All of the facts should be placed in a stringified python list.\n {paragraph}
Fr	Veuillez décomposer le paragraphe suivant en une liste de faits indépendants. Tous les faits doivent être placés dans une liste python sous forme de chaîne de caractères.\n {paragraph}
Ru	Пожалуйста, разбейте следующий абзац на список независимых фактов. Все факты должны быть помещены в строковый список Python (e.g., [’Тим вырос в городе Мальорке, штат Алабама.’,’Его отец был работником верфи.’, ’Мать Тима была домохозяйкой.’,’Кук получил степень бакалавра в области промышленного производства в университете Обёрна в 1982 году.’,’Кук получил диплом MBA в школе Фукуа университета Дьюка в 1988 году.’]).\n {paragraph}"

Table 6: Fact decomposition prompts for each of the languages we consider (§2).

A Fact Decomposition Prompt

We provide the fact-decomposition prompts in Table 6. For Ru, we found that an example was required in order for the GPT-4 response to be consistently structured in the form of a python list of strings, while the other languages (En, Fr) were able to successfully follow this instruction without an example.

B Fact Equivalence Prompt

We provide the prompts for X-FactMatch in Table 7. Concretely, the first row contains the prompt for En → Ru and En → Fr; the second for Fr → En; and the third for Ru → En. In

En	Consider these English facts about {person_name}:\n {src_facts}. Is the last fact in the list inferrerable from the following {tgt_lang} facts?\n {tgt_facts}. Return either yes or no.
Fr	Considérez ces faits français sur {person_name}:\n {src_facts}. Est le dernier fait de la liste inférable de l'une des listes de faits suivantes?\n {tgt_facts}. Retournez oui or non.
Ru	Рассмотрим эти факты на русском языке о {person_name}:\n {src_facts}. Можно ли вывести последний факт из одного из следующих списков фактов?\n {tgt_facts} Возвращает список, содержащий ['да' или 'нет'] — один ответ для каждого списка фактов {tgt_lang}. Все ответы «да/нет» должны быть помещены в список строк Python. (например, ['да', 'нет', 'да'])

Table 7: Prompts for X-FactMatch (§2).

each prompt, the src_facts variable is equivalent to e_{i-2}, e_{i-1}, e_i from §2.2, while tgt_facts contains f_{j-2}, f_{j-1}, f_j, for $j \in [k]$. That is, we use these prompts to determine whether e_i is contained in the other language (e.g., Fr for the En → Fr direction).

C Seed biographies

In Table 8 and Table 9, we list the seed set of biographies, that were used for obtaining InfoGap labels. We performed our human annotation experiment §2.3 for InfoGap on these labels. We then used these labels to distill flan-t5-large and mt5-large for our analyses in §3.

We also used these seed biographies for obtaining connotation labels from GPT-4 for our analysis in §3, which we then also used to distill into flan-t5-large and mt5-large for predicting connotation labels at a larger scale.

D InfoGap Distillation Hyper-Parameters

We report fine-tuning hyperparameters for the HuggingFace Trainer in Table 10. Unspecified values use the default setting of the Trainer (python version: 4.34.1). For all tasks, we used a train/test

En → Fr; Fr → En
Gabriel Attal
Ellen DeGeneres
Tim Cook
Kim Petras
Alan Turing
Caroline Mécary
Abdellah Taïa
Sophie Labelle
Frédéric Mitterrand
Philippe Besson

Table 8: Initial seed set of people for obtaining InfoGap labels with GPT-4 for the En → Fr and Fr → En directions; see §2.3. We used the InfoGap labels on this seed set to finetune a flan-t5-large model; see §3.1.

En → Ru; Ru → En
Pyotr Ilyich Tchaikovsky
Tim Cook
Dmitry Kuzmin
Masha Gessen
Nikolay Alexeyev
James Baldwin
Ali Feruz
Elena Kostyuchenko
Mikhail Zygar
Pyotr Verzilov
Sergey Sosedov
Yekaterina Samutsevich

Table 9: Initial seed set of people for obtaining InfoGap labels with GPT-4 for the Ru → En and En → Ru directions; see §2.3. We used the InfoGap labels on this seed set to finetune a mt5-large model; see §3.1.

Model	Task	hyperparameter	value
`flan-t5`	Fact decomp.	`auto_find_batch_size`	True
		Learning rate	5e-5
		Num. epochs	5
	X-FACTMATCH	`auto_find_batch_size`	True
		Learning rate	5e-5
		Num epochs	5
	Conn. prediction	`auto_find_batch_size`	True
		Learning rate	5e-5
		Num. epochs	5
mT5	Fact decomp.	Batch size	2
		Learning rate	9.5e-4
		Weight decay	0.0
		Gradient accumulation steps	4
		Num. epochs	5
	X-FACTMATCH	Batch size	2
		Learning rate	8.5e-5
		Weight decay	0.4
		Gradient accumulation steps	4
		Num. epochs	5
	Conn. prediction	`auto_find_batch_size`	True
		Learning rate	5e-5
		Num. epochs	5

Table 10: Parameters provided to the HuggingingFace trainer for the `flan-t5-large` and `mt5-large` models.

split of $0.9/0.1$. As for evaluation metrics, we used Rouge-1 for fact decomposition (§2.1), and Micro-F1 for X-FACTMATCH (§2.2) and connotation prediction (§3.3). For the `En` → `Fr` and `En` → `Fr` directions, we apply the `flan-t5` models. We obtained strong validation set performance (0.85 Rouge-1; 0.85 and 0.88 F1s for the connotation prediction and X-FACTMATCH tasks, respectively) using the same hyperparameter settings across all three tasks.

We found that `flan-t5-large` did not generalize well to `Ru`, obtaining poor performance in fact decomposition and often predicting nonsensical Russian strings. We thus resorted to `mt5-large` instead (Xue et al., 2020), since Russian is one of the largest languages in terms of its pre-training data sizes. After a hyperparameter sweep over learning rates, gradient accumulation sizes, and weight decay values, we found much better performance with mT5, obtaining validation set performances of 0.89, 0.79, and 0.86 for fact decomposition, connotation prediction, and X-FACTMATCH tasks, respectively.

All finetuning was completed on a single NVIDIA L40 GPU.

E Connotation modeling

Forced alignment procedure. As mentioned in §3.3, we applied forced alignment to assign decomposed facts back into their original full sentences. Forced alignment is a constrained version of Dynamic Time Warping, where the alignment is monotonic. Forced alignment requires a distance function, we used hubness-corrected distance (Section 2.1).

Connotation prompts . We provide the prompts used for obtaining connotation labels in Table 11. The `content` variable contains up to 3 sentences, s_{i-2}, s_{i-1}, s_i. While we're interested in the connotation towards `person_name` conveyed in the last sentence s_i, we provide the prior two sentences for more context. We prompted for both connotation labels and rationales for the labels, after finding that prompting for a rationale prevented the models from vastly overextending the `neutral` label. This aligns with prior research on text classification, where generating rationales improved accuracy (Wiegreffe et al., 2021).

`En`	The pronoun for `{person_name}` is `{pronoun}`. Does the following text about `{person_name}` imply a positive, neutral, or negative sentiment towards `{person_name}`? Explain why in one sentence. Write your response in JSON format with two keys: label and explanation). `\n {content}` (pos/neutral/neg)
`Fr`	Le pronom du `{person_name}` est `{pronoun}`. Est-ce que le texte suivant au sujet de `{person_name}` implique un sentiment positif, neutre ou négatif envers `{person_name}`? Expliquez pourquoi en une phrase. Écrivez votre réponse en format JSON avec deux clés: étiquette et explication). `\n {content}` (pos/neutral/neg)
`Ru`	Местоимение для {person_name} - {pronoun}. Подразумевает ли следующий текст о {person_name} положительное, нейтральное или отрицательное отношение к {person_name}? Объясните почему в одном предложении. Напишите ваш ответ в формате JSON с двумя ключами: метка и объяснение). \n {content} (положительный/нейтральный/отрицательный)

Table 11: Prompts for obtaining connotation predictions for sentences (§3.3).

Language	Macro-averaged F1
`En`	0.77
`Fr`	0.77
`Ru`	0.86

Table 12: Macro-averaged F1 scores for predicting the connotation towards the subject of a biography from a snippet of text in the biography.

E.1 Validation of connotation label predictions

To validate the connotation labels predicted in §3.3, we sampled 10 positive, 10 negative, and 10 neutral connotation label predictions from Appendix C for each of the 3 languages, thus obtaining 90 datapoints in total. One co-author then annotated each datapoint manually, and compared the annotations against the labels predicted by GPT-4 from the connotation prompt in Appendix E.

We provide the results of this classification in Table 12. We find that the connotation predictions are generally reliable, with all errors stemming from confusion between `neutral` and `positive`, or `neutral` and `positive`, rather than the more severe error of confusing `positive` and `negative` labels. This aligns with observations in previous research on computational modeling of connotation (Park et al., 2021; Rashkin et al., 2015; Sap et al., 2017, among others).

Distillation. Having validated the quality of the GPT-4 connotation predictions, we use the predicted labels to finetune more scalable, lightweight models for predicting the connotation labels. We provide hyperparameter details in Appendix D.

F Regression model fitting

We fit the regression model using the `brms` package (Bürkner, 2017), with 2500 steps (500 warmup) of the NUTS sampler (Hoffman et al., 2014). We used a regularizing $\mathcal{N}(0, 10)$ prior on all the coefficients for the factors.

G Identifying biographies with a positive connotation imbalance across language versions

Plan. We consider the `En` $\rightarrow$ `Fr` direction for an arbitrary bio, without loss of generality. We will use the amount of neutral facts shared by both articles to parameterize a `BetaBinomial` distribution. After fitting this distribution, we will simulate draws from it to predict how much *positive* information should be shared by both articles. When the actual amount of shared positive connotation facts is much lower than the amount predicted by the fitted `BetaBinomial` distribution, we can consider this an imbalanced biography for the `En` $\rightarrow$ `Fr` direction.

Implementation. We first set the prior for the neutral fact distribution to uniform (prior to observing the actual neutral overlap ratio): $\text{Beta}(1, 1)$. We leverage the useful fact that the posterior distribution after observing x neutral facts $e_{i(1)}, \ldots, e_{i(x)}$ in both `En` and `Fr` out of n total facts in `En` is $\text{Beta}(1 + x, 1 + n - x)$ (MacKay, 2003). We can then simulate draws from the `BetaBinomial` distribution, first drawing a sample from $\text{Beta}(1+x, 1+n-x)$, followed by predicting amount of `En` facts that *should* also be found in `Fr`. The number of trials is fixed to the total number of positive facts in the `En` article.

Thus, this binomial distribution tells us the number of positive facts we would expect to see in both articles, if positive facts were not omitted at a higher rate than neutral facts. We can then draw $S = 1000$ samples, counting the number of times K the *expected* amount of shared positive connotation facts is higher than the *actual* amount. When K/S is close to 1.0, there is a large amount of positive information being omitted the `Fr` article, compared to the `En` one. We use $1 - K/S$ as a p-value, with an $\alpha = 0.05$.

We emphasize further that this method can be applied in either direction (e.g., `En` $\rightarrow$ `Fr`, or `Fr` $\rightarrow$ `En`), as well as for finding negatively imbalanced biographies, where one language version includes negative content at a rate much higher than expected under the neutral rate.

H Hoeffding's inequality

We provide the full statement of Hoeffding's inequality for easy reference (Mitzenmacher and Upfal, 2017; Harvey, 2022):

Theorem 1 (Hoeffding's inequality)**.** *Let* $X_1, \ldots, X_n$ *be independent random variables such that* X_i *always lies in the interval* $[0, 1]$. *Define* $X = \sum_{i=1}^{n} X_i$. *Then* $Pr[|X - E[X]| \geq t] \leq 2\exp(-t^2/2n)$.