War and Pieces: Comparing Perspectives About World War I and II Across Wikipedia Language Communities

Ana Smith
Dept. of Computer Science
Cornell University
alsmith@cs.cornell.edu

Lillian Lee
Dept. of Computer Science
Cornell University
llee@cs.cornell.edu

Abstract

Wikipedia is widely used to train models for various tasks including semantic association, text generation, and translation. These tasks typically involve aligning and using text from multiple language editions, with the assumption that all versions of the article present the same content. But this assumption may not hold. We introduce a methodology for approximating the extent to which narratives of conflict may diverge in this scenario, focusing on articles about World War I and II battles written by Wikipedia's communities of editors across four language editions. For simplicity, our unit of analysis representing each language communities' perspectives is based on national entities and their subject-object-relation context, identified using named entity recognition and open-domain information extraction. Using a vector representation of these tuples, we evaluate how similarly different language editions portray how and how often these entities are mentioned in articles. Our results indicate that (1) language editions tend to reference associated countries more and (2) how much one language edition's depiction overlaps with all others varies.

1 Introduction

Wikipedia's expansive content and multiple language editions have made it an invaluable resource, particularly for the training of large language models and translation models in natural language processing (NLP). Less work has gone into quantifying the *differences* among language editions though. In particular, military conflicts, with their political implications and charged nature due to casualties, may be described in distinct ways by different language editions. While community guidelines ensure some quality control and consistency across articles, Table 1 shows that in descriptions from German (DE), English (EN), French (FR), and Italian (IT) Wikipedia articles about the World War I battle at Verdun, there is still disagreement about whether the German objective was to "bleed" the French army. Instead of glossing over this difference, we aim to quantitatively measure it.

There are challenges to measuring these differences, though. Language editions may differ because of (1) linguistic differences in expression; (2) lack of information access, especially due to language barriers; and (3) an author's subjective preferences for sources. There is work on identifying subjectivity in Wikipedia (Recasens et al., 2013; Pavalanathan et al., 2018). But these supervised approaches, while successful, are limited by their need for explicit annotations. This work instead uses unsupervised methods to measure reporting tendencies of Wikipedia articles about battles in World Wars I and II from four language versions — German (DE), English (EN), French (FR), and Italian (IT).

We narrow our scope of analysis to national entities and their contexts, posing the following computationally-amenable question about the representation of such entities:

> *RQ1: How do combatant entity distributions vary among articles from different language editions about the same event?*

Although an author's preferred writing language is not equivalent to an author's nationality, language editions are known to reflect geopolitics in images (He et al., 2018), cultural topics (Tian et al., 2021), and community participation (Shi et al., 2019). Therefore, we hypothesize the following:

> *H1: Languages associated with particular combatants will emphasize that combatant more than others.*

While entity distributions alone facilitate comparisons, the context in which those entities appear may also contribute to subtle differences in perspective. We incorporate context by using (subject,

Proceedings of the 6th LaTeCH-CLfL Workshop, pages 94–104
October 16, 2022.

DE	*Summary: Germany did* not *intend to “bleed” France* In contrast to subsequent representations by the Chief of Staff of the German Army, Erich von Falkenhayn , [3] the original intention of the attack was not to "bleed" the French army without spatial targets. With this assertion made in 1920, Falkenhayn tried to give the unsuccessful attack and the negative German myth of the "blood mill" an alleged meaning.
EN	*Summary: Germany* did *intend to inflict mass casualties on France* Falkenhayn wrote in his memoir that he sent an appreciation of the strategic situation to the Kaiser in December 1915, "...French General Staff would be compelled to throw in every man they have. If they do so the forces of France will bleed to death." The German strategy in 1916 was to inflict mass casualties on the French, a goal achieved against the Russians from 1914 to 1915, to weaken the French Army to the point of collapse.
FR	*Summary: Germany did* not *intend to “bleed” France* According to the version that Falkenhayn gives of his plan in his Memoirs after the war 15 , the goal is to engage in a battle at the loss ratio favorable to the German army, and therefore to discourage France to obtain the stop of the fights... Recent historical works, notably those of the German historian Holger Afflerbach, cast doubt on the version of Falkenhayn who claimed to want to "bleed dry" the French army.
IT	*Summary: Germany* did *intend to “bleed” France* ... [I]n Verdun the purpose of the Falkenhayn offensive was to "bleed the French army to death drop by drop." In the plans of the German Chief of General Staff , the moral and propaganda importance of an attack on Verdun would have meant that all the French effort was poured into the defense of a stronghold considered to be of primary importance for France.

Table 1: Segments of different-language articles that provide contrasting accounts of a supposed German strategy to “bleed” France in the Battle of Verdun. (Google Translate was used for German (DE), French (FR), and Italian (IT); English (EN) is the original.)

relation, object) *tuples* filtered for the geopolitical entities used above, asking the second question:

> *RQ2: How are tuples from different language editions grouped or separated when clustered?*

Differences between language editions are expected, but the gap between languages associated with Germany and Italy and the languages associated with the United States, Britain and France might be expected to have more overlap in their accounts of battles, given wartime alliances:

> *H2: The German (*DE*) and Italian (*IT*) language editions of Wikipedia will overlap more in facts than the English (*EN*) and French (*FR*) language editions.*

Contributions. In a quantitative analysis of entity distributions related to language-country association, we find a language edition associated with a particular country does tend to emphasize that country more than other language editions do (H1 validated). An additional contribution is an approach to reveal conflicting or corroborating tuples by using a downstream diagnostic *battle outcome* inference task. The results of this task indicate that several factors discussed in more detail below affect representation quality.

We demonstrate that though there are more instances of standalone tuples, clustering facts based on similarity across language editions and averaging their representation yields a representation that is more linearly correlated with battle outcome. The results of our outcome prediction task suggest that different language editions provide complementary information and models benefit from using all language versions rather than just one.

In this work, we describe multilingual Wikipedia articles. But there are parallels to news articles from different broadcasters and countries that produce documents covering the same events. A possible extension is to identify domain-specific indicators of differences in opinion in scenarios where a pre-built lexicon is not immediately available, but multiple perspectives are. Another possible application of this methodology is as a diagnostic tool to identify potential sources of bias in Wikipedia datasets.

2 Related work

There is prior work extracting relations between and events involving geopolitical entities from text (O’Connor et al., 2013; Chambers et al., 2015; Makarov, 2018; Han et al., 2019; Stoehr et al., 2021); see Hürriyetoğlu et al. (2021) for a recent collection of papers. We focus on managing and comparing descriptions of such relations across different language communities (McCarthy et al., 2021; Scharf et al., 2021). (Of course, multilingual parallel and comparable corpora have been a mainstay of machine translation since its beginnings.)

2.1 Multilingual Wikipedia

Our research is primarily a study of the relationship between a Wikipedia article's content and its relationship to the corresponding article in another language edition. Other work compares Wikipedia language editions from the perspective of the geography associated with an article (Lieberman and Lin, 2009), the imagery of articles (He et al., 2018; Porter et al., 2020), and perspectives of colingual groups on common topics (Tian et al., 2021). Our project is closely aligned in spirit with other analyses of how wars are described across different language communities in Wikipedia (Gieck et al., 2016; Zhou et al., 2015; Bridgewater, 2017; Kubś, 2021)

2.2 Wikipedia and information extraction

Wikipedia has served various purposes outside of its obvious role as an open-edited, free encyclopedia. After years of studies on Wikipedia's information quality (Stvilia et al., 2007; Arazy et al., 2011; Kumar et al., 2016), more recent work focuses more on leveraging it to answer questions (Chen et al., 2017), populate knowledge bases (Hoffmann et al., 2011; Wu and Weld, 2008), and generate summary tables (Liu et al., 2019). The former line of work more directly questions the quality of Wikipedia content. We do not assess the quality of information directly, but rather assess the prevalence of certain pieces of information. Our work is similar to the latter line of work in that we attempt to simplify Wikipedia content to a few phrases for analysis. Our work differs from prior work in that it does not extract snippets from a larger body of text to fill in answers. Rather, it compares snippets from multiple language editions.

	WWI			WWII		
Rank	Lang	No.	≉ En	Lang	No.	≉ En
1	EN	606	—	EN	2958	—
2	FR	373	23%	FR	1358	10%
3	IT	327	7%	IT	888	10%
4	DE	225	16%	DE	788	5%

Table 2: Number of retrieved distinct identifiers for Wikipedia articles listed under the WWI or WWII battle categories. (Recall that we restricted attention to Latin-script languages for countries with the most casualties.) "≉ En" columns: % of articles in that language without an English-language equivalent.

3 Data Collection

Our corpus of battle descriptions is collected from multiple language editions of Wikipedia. To identify potential candidate articles for download, we take the names of articles listed under the English language categories "Battles of World War I" and "Battles of World War II"[1] and corresponding categories in other language editions (e.g., Battaglie_della_prima_guerra_mondiale) identified by interlanguage Wikilinks for German, French, and Italian. These languages were selected because they are the primary languages employing Latin script used by combatant countries with the largest recorded casualties.[2]

Different language editions do encompass different sets of articles, with some articles available in only a subset of data. So even if the communities are comprised of the same individuals with the same aims in every language edition, the output is non-equivalent for all languages. In total, our dataset has 765 distinct WWI battles and 3430 distinct WWII battles. See Table 2 for the distribution across language editions.

After the names of battle articles in different languages are collected, they are disambiguated by linking them to a Wikidata item identifier known as a QID, obtained by querying the WikiData API. QIDs link articles across different language editions, and we use the reduced set of QIDs to identify all language editions of each article. Though there is still a bias for articles grouped under the "Battles of World War I" and "Battles of World War II" categories, this additional step reduces the likelihood that we are collecting data only visible from English Wikipedia. For example, the DE version of Wikipedia tends to have fewer articles, possibly because they conceptualize warfare differently (e.g., campaigns instead of actions).

Full-text content is then downloaded from Wikipedia using the PetScan interface[3]. The next section discusses how this data is further cleaned

[1] https://en.wikipedia.org/wiki/Category:Battles_of_World_War_I, https://en.wikipedia.org/wiki/Category:Battles_of_World_War_II

[2] We repeat that the restriction to Latin script is an attempt to minimize processing differences between languages. Moving to a larger set of more diverse languages is a direction for future work.

[3] https://en.wikipedia.org/wiki/Wikipedia:PetScan

WWI				WWII			
DE	EN	FR	IT	DE	EN	FR	IT
german	german	german	german	german	german	german	german
british	british	british	british	japanese	japanese	japanese	japanese
french	french	french	french	british	british	british	british
russian	germans	germans	germans	soviet	italian	french	italian
army	russian	france	russian	american	soviet	germans	soviet
germans	ottoman	russian	italian	us	french	soviet	germans
division...	france	ottoman	germany	allied	germans	american	french
...(german empire				germans	allied	us	american
italian	russians	germany	france	italian	us	germany	us
german empire	belgian	italian	russians	french	american	france	allied
austria	germany	armenian	russia	army	germany	allied	germany
france	allied	austro	austrian	americans	france	italian	italy
hungary	armenian	austria	austro	france	japan	japan	france
reserve division	italian	hungary	ottoman	infantry...	axis	americans	americans
army corps	russia	turkish	turkish	...division			
russians	austria	somme	army	category	united states	united states	japan
category	belgium	allied	belgian	polish	dutch	soviets	soviets
austrian	uk	russians	allied	dutch	chinese	polish	polish
reserve corps	hungary	ottomans	italy	germany	the united states	dutch	axis
weblinks	romanian	italy	meuse	japan	italy	italy	army
germany	turkish	serbian	belgium	the red army	italians	category	chinese

Table 3: Top 20 most frequent non-pronoun, non-individual-human terms per language (after →Spanish→English translation) automatically tagged as geopolitical named entities in our World War I (left) and World War II (right) corpora.

and partitioned.

4 Associated Languages and Entities

Initially, all battle articles listed under the battle categories in each of the four languages are collected. But because this work compares language editions, only the intersection of the four language editions is used. This results in 131 articles for World War I and 414 articles for World War II. This subset is then processed as described below.

4.1 Processing articles

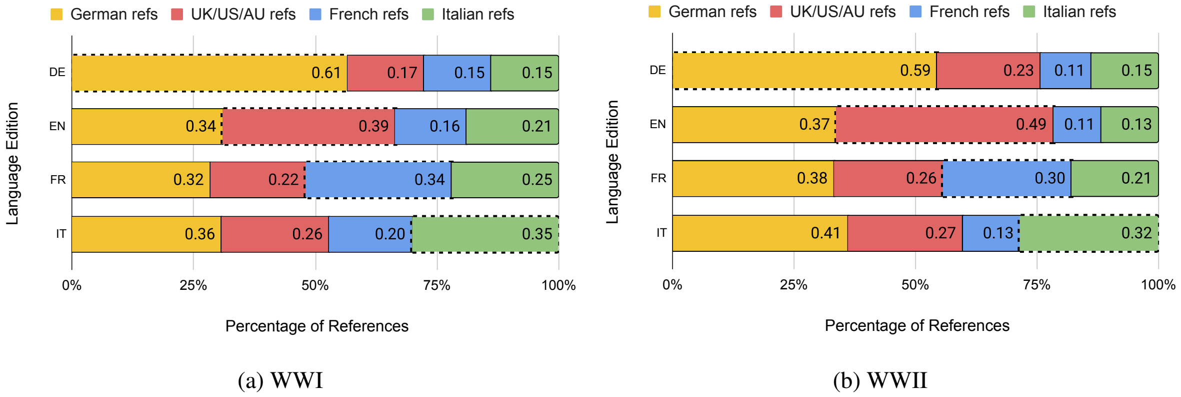
Our approach requires the use of open domain information extraction, which has until recently been largely restricted to English, so all articles must be translated to English for our method. To compensate for translation noise in our non-English articles, all articles (including English articles) are translated to "new", fifth language, Spanish, and then to English using Google Translate.[4] Importantly, we subject English to potential translation errors to avoid privileging it as the only language under consideration that would not have undergone translation otherwise.

Translation. Using different language revisions enables us to probe differences across groups of editors employing the same language. On the other hand, although the original language of the articles is expected to give the most accurate distinctions, we choose to work with translated versions of the articles so that we can apply a standardized set of NLP tools developed for English. To avoid privileging the originally-English articles, all language versions are first translated to a *new* language (Spanish, given that there are many high-quality machine translation models between Spanish and other languages) before being then retranslated into English via Google Translate.

Text cleaning The collected articles are in xml format, complete with internal links, templates, and other artifacts. The article text is sentence- and word-tokenized; then, internal links are simplified to the alt-text only, and we remove templates including infoboxes, inline references, and text starting from the section headers "References" and "See also".

Named entity tagging. Though there are two major sides in these wars, there are numerous combatants. We use the named entity tagger to identify geopolitical entities and persons. Manual inspection of the entities in the context of the article is used to identify ties to a single political entity. Al-

[4] We employed only European languages to stay within a family of relatively related languages; future work can be more ambitious about language choices.



(a) WWI

(b) WWII

Figure 1: Comparison of per-article proportions of self-references vs. other-references, as discussed in §4.3.

liances of entities (e.g., Allied Powers) are considered separately.

Associating entities with nations/alliances. Recall our first research question is related to the combatant distributions across language editions. One difficulty is associating a particular entity with a combatant nation, due to issues with granularity and type of reference: American entities may be referred to as *the United States* (nation name), *Eisenhower* (leadership), *333rd Field Artillery Battalion* (military unit), or *they* (pronoun). To address this issue, a list of nations, leaders grouped by nation, and military units by nation are collected for each article from English Wikipedia categories and pages. (We exclude pronouns and entities not clearly identified by nationality as existing coreference tools did not prove reliable enough on our data.) Though this does not encompass all entities mentioned in our corpus, it does capture prominent entities.

4.2 Entity-count statistics

In total, there are 88,317 entities in our WWI corpus and 274,713 entities in our WWII corpus. Table 3 lists the most common non-pronoun non-person grammatical subjects in our World War I and World War II data. The most prominent national entity across all language editions by far is Germany. This is to be expected given that in both wars Germany was engaged with combatants on both the Eastern Front and the Western Front, whereas most other combatants only appear on one Front. In the World War I corpus, the British are the second most common national entity subject. In the World War II corpus, the Japanese are the second most prominent national entity. Not shown here is a list of PERSON entities. The most frequent persons listed in those tables are, surprisingly, *battle* in WWI and, unsurprisingly, *Hitler* in WWII. Our tags do contain noise. The word *battle* should *not* be tagged as a person, but it was tagged so across all language editions. The spaCy (Honnibal and Montani, 2017) en_core_web_sm model was used to obtain named entity and part-of-speech information.

4.3 Associated-language test (RQ1)

In the introduction, we hypothesized that languages associated with a combatant country would reference that combatant as a subject more than any other combatant. To evaluate our hypothesis, we compare the relative proportion of counts per article of *self-references* (i.e., references to a nation by its associated language) to *other-references* (i.e., references to a nation by other languages). Each other-reference is normalized by the number of other languages (i.e., 3) for a more balanced comparison to other self-references. Though doing so reduces statistical power, instances are grouped by war for better analysis.

Figure 1 is a stacked barplot of the self-reference and other-reference proportions in our dataset. To test significance between populations, we use the Mann-Whitney U test implementation in scipy (Virtanen et al., 2020), as our population sizes differ between the "self" country reference group and the "other" countries reference group and are non-normally distributed. When using a Bonferroni correction of 2 on a p-value threshold of 0.01 since a test was run for each war, our p-values for both WWI (5.46e-6) and WWII (8.61e-4) are significant at <0.005. Though the data are not normally distributed, the self-reference distribution suggests that our hypothesis H1 is supported (i.e., languages associated with particular nations are more likely to mention those nations than ones that are not).

A breakdown of references by language edition and country reveals more nuance, with self-references highlighted by the dashed borders. The significance of the above test may be attributed in part to DE's many self-references and other language editions' many other-references to DE. This is likely because Germany's engagement on both Eastern and Western fronts made it a more common reference overall. That said, for every language version, the proportion of self-references is greater than references to that country in other language editions. This indicates there is indeed a tendency to emphasize the countries commonly associated with these languages. We consider H1 validated.

5 Tuple Clusters

While the entities alone indicate a preference for language editions to reference their associated countries more, the context in which they occur may aid our understanding of why these differences in distribution occur. We hypothesized that overlap among languages may be more likely between English and French accounts and German and Italian accounts than any combination of the two. But overlap alone says little about why accounts may differ.

We simplify article text to (subject, object, relation) tuples. Solely as a means to validate the quality of representation, a domain-specific outcome inference task is used. The intuition is that a better representation should enable a linear classifier to learn a correlation between outcome and text, among other properties.

5.1 Extracting tuples and clustering

Tuple extraction. Once all articles are translated, (subject, relation, object) tuples are extracted with the Stanford NLP Toolkit's OpenIE implementation (Angeli et al., 2015). This system was chosen instead of a neural approach to limit the possibility that information is hallucinated or generated that was not in the original text (such problems are known to occur in neural models such as Imojie (Kolluru et al., 2020)).

One problem is that essentially redundant tuples may be considered distinct. Consider the following tuples:

1. EN: ('sides', 'suffered casualties with', 'numbers of soldiers succumbing to freezing')
2. EN: ('sides', 'suffered casualties with', '**large** numbers of soldiers succumbing to freezing')

The only difference between (1) and (2) is the adjective "large" in the object. To address this problem, we group tuples by subject and relation per article section (e.g., == *Aftermath* ==) and take only the tuple within each group with the longest object (in tokens). No subject should be a substring of another subject, and no relation should be a substring of another relation. Hence, tuple (2) would be retained and (1) discarded.

Tuple representation. Following Kristof et al. (2021), averaged word embeddings are used to represent text content. As a baseline, we compare this against a 1- to 3-gram bag-of-words.

We begin with a basic representation of tuple t that doesn't distinguish between subject, object, and verb (relation) status:

$$v_{sro} = \frac{1}{|t|} \sum_{w \in t} \text{emb}(w) \qquad (1)$$

where emb() is a mapping of w to a pretrained vector. This reflects our naive hypothesis that treating an entity (e.g., France) as an object is not distinct from treating it as the subject. We also compare a pretrained embedding (GLoVe (Pennington et al., 2014)) and an embedding trained on our corpus (using fasttext) only to assess the extent to which the context of World War conflict influences a model. Though GLoVe is trained on more data, the nature of conflict may contravene typical associative assumptions and domain-specific words (especially entities) may be dropped. Both vectors are of dimension 100. This dimension was chosen because previous studies suggest that dimensions on the order of 100 are relatively similar in performance but better than those with dimensions on the order of 10 (Rodriguez and Spirling, 2021). In the case of GLoVe, a random vector was assigned to out-of-vocabulary words. The fasttext embeddings were trained using a character n-gram of maximum size 3 and a learning rate of 0.05. These embeddings are trained over the combined corpus (both WWI and WWII). Words appearing in fewer than 0.1% of tuples are excluded to manage the number of features and prevent overfitting.

The first representation neglects the structure denoted by the tuple. But this may be harmful in cases where distinguishing the subject and the object tuple matters (e.g., (France, defeated, Germany) is

1st lang	Tuples contributed to cluster
DE	('German armed forces', 'lost will', 'resist')
DE	('German positions', 'against Army is', 'United Kingdom')
DE	('British troops', 'Only announced', 'their victory at Battle of Havrincourt')
DE	('German forces', 'lost will', 'resist')
EN	('Germans', 'could consolidate', 'their positions')
EN	('American forces', 'face', 'difficult task')
EN	('Germans', 'encouraged', 'Allies')
EN	('Germans', 'were', 'weakening')
FR	('German divisions', '6 at', 'least')
FR	('German army', 'withdraw until', 'November 11 1918')
IT	('advance', 'would', 'would also backed by 300 machine guns')

Table 4: An example multilingual (s, r, o) cluster obtained from articles on the 1918 Battle of Havrincourt. The component tuples, while from four distinct languages, generally correspond to the "tuple" that the Germans were unable to hold their position against British troops.

distinct from (Germany, defeated, France)). To address this, a 300 dimensional representation is concatenated to v_{sro}. The mean vector for each word in the subject (s), relation (r), and object (o) is calculated as above and concatenated as follows:

$$v^{(t)} = [v_s; v_r; v_o] \tag{2}$$

Though the structure of $v^{(t)}$ ensures that the word *France* as an object is distinct from *France* as a subject, similar tuples may be written in the passive voice in one language and not another. To combat the issue of word order, $v^{(t)}$ is concatenated to v_{sro} to form the second feature vector used:

$$v^{(t)}_{final} = [v_s; v_r; v_o; v_{sro}] \tag{3}$$

Clustering tuples into tuples. The ultimate goal is to group similar tuples from different language versions in such a way that we minimize the size of the clusters — so that the included tuples should be more similar — while maximizing heterogeneity of within-cluster source languages, that is, the number of source languages represented in the cluster. To address both limits, we implement a hierarchical K-means clustering algorithm with thresholds for cluster sizes. Euclidean distance is used to measure (dis)similarity among instances. Clusters are recursively split until they contain fewer than 16 instances. Table 4 shows an example cluster.

Because word embeddings may associate words by type (e.g., tuples with *Germany* and *France* as subjects appear in the same cluster), an additional one-hot vector is prepended to $v^{(t)}_{final}$ to split tuple clusters along country lines when clustering.

$$v^{(t)}_{cluster} = [a_{\text{de}}; a_{\text{en}}; a_{\text{fr}}; a_{\text{it}}; v^{(t)}_{final}] \tag{4}$$

Here, $a_{\text{<language>}}$ is 1 if the associated language occurs in the subject of the tuple, otherwise 0. A single cluster can be represented by the mean of all $v^{(t)}_{cluster}$ tuple representations in the cluster. It is this mean vector that is used in the following experiments.

5.2 Validating representation quality

To assess the quality of the proposed representations, we use the outcome of the battle as a target to evaluate the extent these representations implicitly attribute advantages to (or minimize disadvantages of) combatants. For this task, the input is a tuple representation and the output is the *outcome* (e.g., 0 if Germans won, otherwise 1). Not every tuple is expected to directly correspond to the outcome, but any tuple that does should benefit from a better representation as indicated by an increase in model precision. In our experiments, we employ 3-fold cross-validation; for each fold, we fit a logistic regression model using the scikit-learn implementation (Pedregosa et al., 2011). The regularization parameter C is tuned over the range [0.01, 0.1, 0.5, 1.0, 3.0]. The results of evaluating the model on a held-out test set are shown in Table 5.

Results. The bag-of-words (*bow*) representation presents a competitive baseline, particularly for WWI, as do the smaller v_sro representations. The WWII corpus benefits from the word embedding representation across the board, though. (Bear in mind that it is approximately 4 times larger than the WWI corpus.) Additionally, averaging the tuple representations per cluster yields even better outcome inference results — for example, on WWII using fasttext, F1 goes from .567 for unclustered to .662 for clustered — likely because of the larger context on which it draws in comparison to a single tuple.

Though there are fewer instances, using clusters is more advantageous in outcome inference than using individual tuples suggesting that the context derived from grouping similar tuples is useful for corroborating outcomes. Part of this effect may be due to complementary information from different language editions. Using $v^{(t)}_{cluster}$, we turn

feature	WWI tuples			WWI clusters			WWII tuples			WWII clusters		
	F1	recall	prec	F1	recall	prec	F1	recall	prec	F1	recall	prec
majority	0.372	0.500	0.297	0.286	0.500	0.201	0.378	0.500	0.304	0.317	0.500	0.232
$\#words$	0.372	0.500	0.297	0.375	0.500	0.299	0.378	0.500	0.304	0.349	0.500	0.268
$\#tuples$	0.372	0.500	0.297	0.375	0.500	0.299	0.378	0.500	0.304	0.349	0.500	0.268
bow_{sro}	0.467	0.523	0.560	**0.609**	**0.610**	**0.635**	0.502	0.545	0.617	0.573	0.586	0.608
bow_{final}	**0.475**	0.520	0.545	0.604	0.605	0.622	0.508	0.547	0.611	0.583	0.591	0.607
v_{sro} (G)	0.392	0.506	**0.616**	0.536	0.555	0.585	0.468	0.533	0.636	0.602	0.616	0.650
v_{final} (G)	0.431	0.516	0.581	0.531	0.539	0.548	0.512	0.553	0.638	0.606	0.621	0.660
v_{sro} (F)	0.435	0.521	**0.616**	0.562	0.573	0.604	0.537	0.569	0.658	0.633	0.642	0.675
v_{final} (F)	0.468	**0.533**	0.613	0.602	0.602	0.616	**0.567**	**0.586**	**0.663**	**0.662**	**0.669**	**0.706**

Table 5: Battle outcome inference results using several representations. (F) denotes the use of fasttext vectors, while (G) denotes GLoVe. On the left side of each table are the results obtained when using individual tuples as instances. On the right side are the results obtained when using the mean of a cluster's tuple representations as instances.

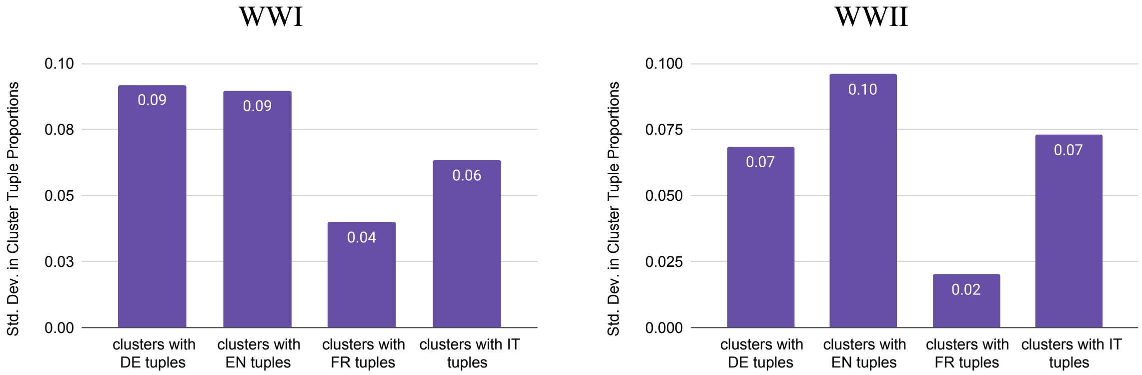


Figure 2: Bar chart showing the standard deviation in the proportion of language tuples in a subset of clusters defined by the presence of at least one tuple of a particular language. The cluster subsets defined by the presence of FR tuples in both WWI and WWII tend to have a balanced mix of tuples from DE, EN, and IT.

to our second research question regarding overlap between language editions with clusters.

5.3 Measuring cluster composition (RQ2)

To measure language heterogeneity in the tuples, each language (l_1) is paired with every other unique language (l_2) counted in the cluster. The count of occurrences of l_2 is then divided by the total number of tuples for that language. Correcting in this way, rather than using simple overlap, is intended to reduce the effects of population size (e.g., there being more EN tuples than FR tuples means the former are more likely to end up in any cluster by chance). This in turn helps us to better assess semantic (dis)agreement among language editions.

Figure 2 shows that the tuples with FR language tend to co-exist in a balanced manner with tuples from other languages in both the WWI and WWII data; this is true even though FR has the fewest tuples of all the language editions. One possible explanation may be that though the French language version contains fewer tuples, each tuple tends to be corroborated by other language versions. See Table 6 for the total number of tuples. In contrast, EN tends to be the most variable in its proportions. Though FR clusters include EN tuples in a similar proportion () to all other tuples, EN includes a much smaller proportion of FR tuples (). These results partially contradict our hypothesis that the overlap would be greatest between FR and EN and between DE and IT. We consider H2 as not validated.

6 Conclusion

In this work, we introduced a methodology for identifying information upon which language editions agree and disagree by applying open-domain information extraction and unsupervised learning to English translations of articles. Our results indicate that (1) language editions tend to mention their associated country more than other language editions mention the same country and (2) the FR language

Lang	WWI		WWII	
	Tuples	Clusters	Tuples	Clusters
DE	181,456	78.5%	526,290	77.5%
EN	184,795	81.0%	504,085	69.8%
FR	107,879	69.6%	376,489	69.8%
IT	133,041	69.5%	532,223	76.3%

Table 6: Counts and cluster coverage of tuples extracted from the World War I and World War II corpora using the Stanford OpenIE system. The "Clusters" columns indicate the proportion of clusters in which the languages appear.

edition align with other language editions' accounts more than the reverse. Result (1) confirms other work on geopolitical tendencies of multilingual Wikipedia. Result (2) implies that FR Wikipedia may have a more limited though balanced account than other language editions. More qualitative analysis is needed though.

Limitations There are limitations to using machine translation for historical analysis. To avoid issues regarding nuance, articles are reduced to a set of simple (subject, relation, object) tuples. The vector representations used were also evaluated on downstream tasks before use in our second experiment.

Future work There are several possible directions for future work. Regarding tasks, it may be of interest to NLP practitioners to understand the impact the information imbalances have on downstream tasks such as translation. For the language communities themselves, it may be useful to be aware of the gaps in the accounts they are writing. To make this more useful for them, an important step would be to expand to other languages; our analysis is limited to four languages. Future work should include articles from languages correlated with combatants on the Western and Pacific front.

Ultimately, more conclusive results will require a better model of the community dynamics and citation practices of editors, especially over time as well as more qualitative analysis of the differences between language editions. We aim to continue this work with the hope it encourages interest and advances in the overlap of computational, historical, and cultural analysis.

7 Acknowledgments

We thank the anonymous reviewers, Austin Benson, Tianze Shi, Arthur Spirling, and the Cornell NLP group for helpful comments. This work was supported in part by a Cornell Provost Diversity Fellowship.

References

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354.

Ofer Arazy, Oded Nov, Raymond Patterson, and Lisa Yeo. 2011. Information quality in wikipedia: The effects of group composition and task conflict. *Journal of Management Information Systems*, 27(4):71–98.

Matt Bridgewater. 2017. History writing and Wikipedia. *Computers and Composition*, 45:36–50.

Nathanael Chambers, Victor Bowen, Ethan Genco, Xisen Tian, Eric Young, Ganesh Harihara, and Eugene Yang. 2015. Identifying political sentiment between nation states with social media. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 65–75, Lisbon, Portugal. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Robin Gieck, Hanna-Mari Kinnunen, Yuanyuan Li, Mohsen Moghaddam, Franziska Pradel, Peter A. Gloor, Maria Paasivaara, and Matthäus P. Zylka. 2016. Cultural differences in the understanding of history on Wikipedia. In *Designing Networks for Innovation and Improvisation*, pages 3–12, Cham. Springer International Publishing.

Xiaochuang Han, Eunsol Choi, and Chenhao Tan. 2019. No permanent Friends or enemies: Tracking relationships between nations from news. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1660–1676, Minneapolis, Minnesota. Association for Computational Linguistics.

Shiqing He, Allen Yilun Lin, Eytan Adar, and Brent J Hecht. 2018. The_tower_of_babel. jpg: Diversity of visual encyclopedic knowledge across wikipedia language editions. In *ICWSM*, pages 102–111.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 541–550. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Ali Hürriyetoğlu, Hristo Tanev, Vanni Zavarella, Jakub Piskorski, Reyyan Yeniterzi, Osman Mutlu, Deniz Yuret, and Aline Villavicencio. 2021. Challenges and applications of automated extraction of socio-political events from text (CASE 2021): Workshop and shared task report. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 1–9, Online. Association for Computational Linguistics.

Keshav Kolluru, Samarth Aggarwal, Vipul Rathore, Soumen Chakrabarti, et al. 2020. Imojie: Iterative memory-based joint open information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5871–5886.

Victor Kristof, Aswin Suresh, Matthias Grossglauser, and Patrick Thiran. 2021. War of words II: Enriched models of law-making processes. In *Proceedings of the Web Conference 2021*, WWW '21, page 2014–2024, New York, NY, USA. Association for Computing Machinery.

Jakub Kubś. 2021. Historical narratives in different language versions of wikipedia. *Academic Journal of Modern Philology*, (12):83–94.

Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, pages 591–602. International World Wide Web Conferences Steering Committee.

Michael Lieberman and Jimmy Lin. 2009. You are where you edit: Locating Wikipedia contributors through edit histories. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Tianyu Liu, Fuli Luo, Pengcheng Yang, Wei Wu, Baobao Chang, and Zhifang Sui. 2019. Towards comprehensive description generation from factual attribute-value tables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5985–5996.

Peter Makarov. 2018. Automated acquisition of patterns for coding political event data: Two case studies. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 103–112, Santa Fe, New Mexico. Association for Computational Linguistics.

Arya D. McCarthy, James Scharf, and Giovanna Maria Dora Dore. 2021. A mixed-methods analysis of western and Hong Kong–based reporting on the 2019–2020 protests. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 178–188, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Brendan O'Connor, Brandon M. Stewart, and Noah A. Smith. 2013. Learning to extract international relations from political context. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1094–1104, Sofia, Bulgaria. Association for Computational Linguistics.

Umashanthi Pavalanathan, Xiaochuang Han, and Jacob Eisenstein. 2018. Mind your POV: Convergence of articles and editors towards Wikipedia's neutrality norm. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–23.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Emily Porter, P. M. Krafft, and Brian Keegan. 2020. Visual narratives and collective memory across peer-produced accounts of contested sociopolitical events. *Trans. Soc. Comput.*, 3(1).

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659.

Pedro Rodriguez and Arthur Spirling. 2021. Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *Journal of Politics*.

James Scharf, Arya D. McCarthy, and Giovanna Maria Dora Dore. 2021. Characterizing news portrayal of civil unrest in Hong Kong, 1998–2020. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 43–52, Online. Association for Computational Linguistics.

Feng Shi, Misha Teplitskiy, Eamon Duede, and James A Evans. 2019. The wisdom of polarized crowds. *Nature human behaviour*, 3(4):329–336.

Niklas Stoehr, Lucas Torroba Hennigen, Samin Ahbab, Robert West, and Ryan Cotterell. 2021. Classifying dyads for militarized conflict analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7775–7784, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Besiki Stvilia, Les Gasser, Michael B Twidale, and Linda C Smith. 2007. A framework for information quality assessment. *Journal of the American society for information science and technology*, 58(12):1720–1733.

Yufei Tian, Tuhin Chakrabarty, Fred Morstatter, and Nanyun Peng. 2021. Identifying distributional perspectives from colingual groups. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 178–190.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. 2020. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272.

Fei Wu and Daniel S Weld. 2008. Automatically refining the wikipedia infobox ontology. In *Proceedings of the 17th international conference on World Wide Web*, pages 635–644. ACM.

Yiwei Zhou, Alexandra Cristea, and Zachary Roberts. 2015. Is Wikipedia really neutral? A sentiment perspective study of war-related Wikipedia articles since 1945. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 160–168, Shanghai, China.