Omnipedia: Bridging the Wikipedia Language Gap

Patti Bao[*†], Brent Hecht[†], Samuel Carton[†], Mahmood Quaderi[†], Michael Horn[†§], Darren Gergle[*†]
[*]Communication Studies, [†]Electrical Engineering & Computer Science, [§]Learning Sciences
Northwestern University
{patti,brent,sam.carton,quaderi}@u.northwestern.edu, {michael-horn,dgergle}@northwestern.edu

ABSTRACT
We present Omnipedia, a system that allows Wikipedia readers to gain insight from up to 25 language editions of Wikipedia simultaneously. Omnipedia highlights the similarities and differences that exist among Wikipedia language editions, and makes salient information that is unique to each language as well as that which is shared more widely. We detail solutions to numerous front-end and algorithmic challenges inherent to providing users with a multilingual Wikipedia experience. These include visualizing content in a language-neutral way and aligning data in the face of diverse information organization strategies. We present a study of Omnipedia that characterizes how people interact with information using a multilingual lens. We found that users actively sought information exclusive to unfamiliar language editions and strategically compared how language editions defined concepts. Finally, we briefly discuss how Omnipedia generalizes to other domains facing language barriers.

Author Keywords
Wikipedia; multilingual; hyperlingual; language barrier; user-generated content; text mining

ACM Classification Keywords
H5.m. [Information interfaces and presentation (e.g., HCI)]: Miscellaneous.

General Terms
Human Factors

INTRODUCTION
As the sixth most popular website in the world, Wikipedia has over 270 language editions, dozens of which have more than 100,000 articles [20]. Each language edition of the online encyclopedia has been shown to have a surprising amount of information not contained in any other language edition [11, 15]. This is true of both large Wikipedias like English and smaller ones such as Catalan and Finnish.

The language-induced splintering of information in Wikipedia poses both an opportunity and a challenge. On the one hand, Wikipedia embodies an unprecedented repository of *world knowledge diversity* in which each language edition contains its own cultural viewpoints on a large number of topics [7, 14, 15, 27]. On the other hand, the language barrier serves to silo knowledge [2, 4, 33], slowing the transfer of less culturally imbued information between language editions and preventing Wikipedia's 422 million monthly visitors [12] from accessing most of the information on the site.

In this paper, we present Omnipedia, a system that attempts to remedy this situation at a large scale. It reduces the silo effect by providing users with structured access in their native language to over 7.5 million concepts from up to 25 language editions of Wikipedia. At the same time, it highlights similarities and differences between each of the language editions, allowing users to see the diversity of the represented knowledge. To achieve this goal, Omnipedia extracts the topics discussed in each language edition's coverage of a given concept, then loads them into an interactive visualization that shows which language editions mention which topics and how those topics are discussed.

Consider, for example, the English Wikipedia article "Conspiracy theory". This article discusses many topics, from "Moon landing" to "Kennedy assassination". However, many other language editions also contain articles on this concept, such as "*Verschwörungstheorie*" in the German Wikipedia and "*Teoría conspirativa*" in the Spanish Wikipedia. Omnipedia consolidates these articles into a single "multilingual article" on conspiracy theories, as seen in Figure 1. The small circles on the left represent

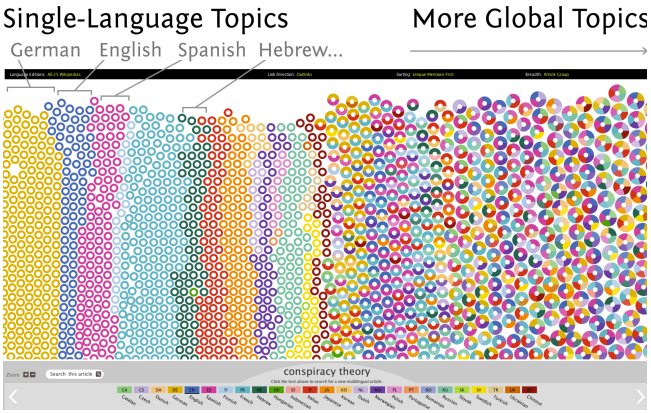

Figure 1. A screenshot of Omnipedia showing the multilingual article "Conspiracy theory" in zoomed-out mode. Small circles on the left indicate topics that are discussed in only a single language edition's coverage of the concept. Bigger circles on the right indicate topics that are discussed in multiple language editions' coverage of "Conspiracy theory".

CHI'12, May 5–10, 2012, Austin, Texas, USA.

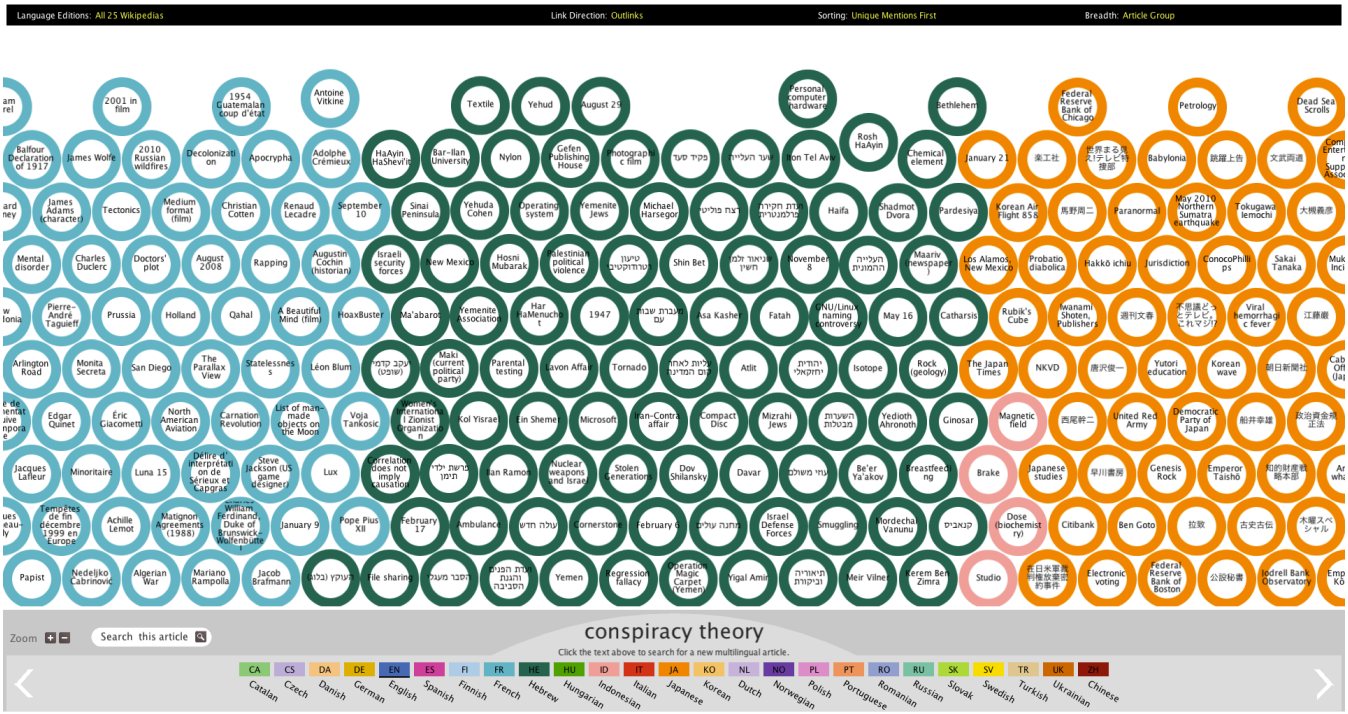

Figure 2. "Conspiracy theory" in zoomed-in mode. The user can see specific topics discussed in each language edition's article. Because the user has panned over to the single language linked topics, s/he can see that the Hebrew Wikipedia (dark green) discusses "Palestinian political violence" while the French Wikipedia (cyan) discusses "Algerian War". Clicking on one of the circles calls up a snippet (Figure 4) from the corresponding Wikipedia article(s) that explains the discussion of each topic in detail.

topics discussed in only one language edition: yellow for German, dark blue for English, and so on. It should be clear that by only reading a single language edition's coverage—even that of English—one misses out on huge amounts of content available in other language editions. Moving toward the right half of Figure 1, one begins to see larger, multi-colored circles that represent topics that are discussed in multiple language editions' coverage of the concept.

Zooming in (Figure 2) allows users to explore content in more detail. For instance, Figure 2 shows that the Hebrew Wikipedia (dark green) has a great deal of exclusive content about Israel-related conspiracy theories. The French Wikipedia (cyan) also has unique content, both pertaining to French history, as indicated by "Algerian War", and of more general interest, such as "Pope Pius XII". Panning right, users begin to find topics that are discussed in more than one language edition. Figure 3 shows the most commonly discussed topics in the "Conspiracy theory" multilingual article, which include "Freemasonry", "United States", and "Central Intelligence Agency". We also see that Judaism is discussed in many language editions' coverage of conspiracy theories, demonstrating that this form of anti-Semitism is unfortunately widespread. To discover precisely how these topics are discussed in various language editions, users can click on a topic circle. This returns a snippet mined from each language edition that covers the clicked topic, with snippets translated into the user-defined language using machine translation (Figure 4).

This simple example offers a brief glimpse of the capabilities of Omnipedia. In addition to the system itself, this paper presents several technical contributions related to building highly multilingual or "hyperlingual" [15] applications. In particular, we introduce natural language processing and information retrieval strategies that are sensitive to the diversity of world knowledge in Wikipedia. These advances have broader implications for human-centered technologies intended to work with a wide variety of users and contexts.

RELATED WORK

There is a growing field of research on multilingual Wikipedia that can be broadly divided into two groups: (1) work that studies multilingual Wikipedia and (2) work that attempts to propagate information from one Wikipedia language edition to another.

The general consensus on multilingual Wikipedia is that while English (the largest language edition) has a content advantage, each language edition has a great deal of unique information not available in other language editions, and most other information exists in only a few language editions [11, 15]. This diversity occurs in terms of *which* concepts are covered as well as *how* concepts are covered [11, 15], the latter of which is what Omnipedia visualizes more directly.

Mounting evidence suggests that some of these differences are due to variation in world knowledge across language-

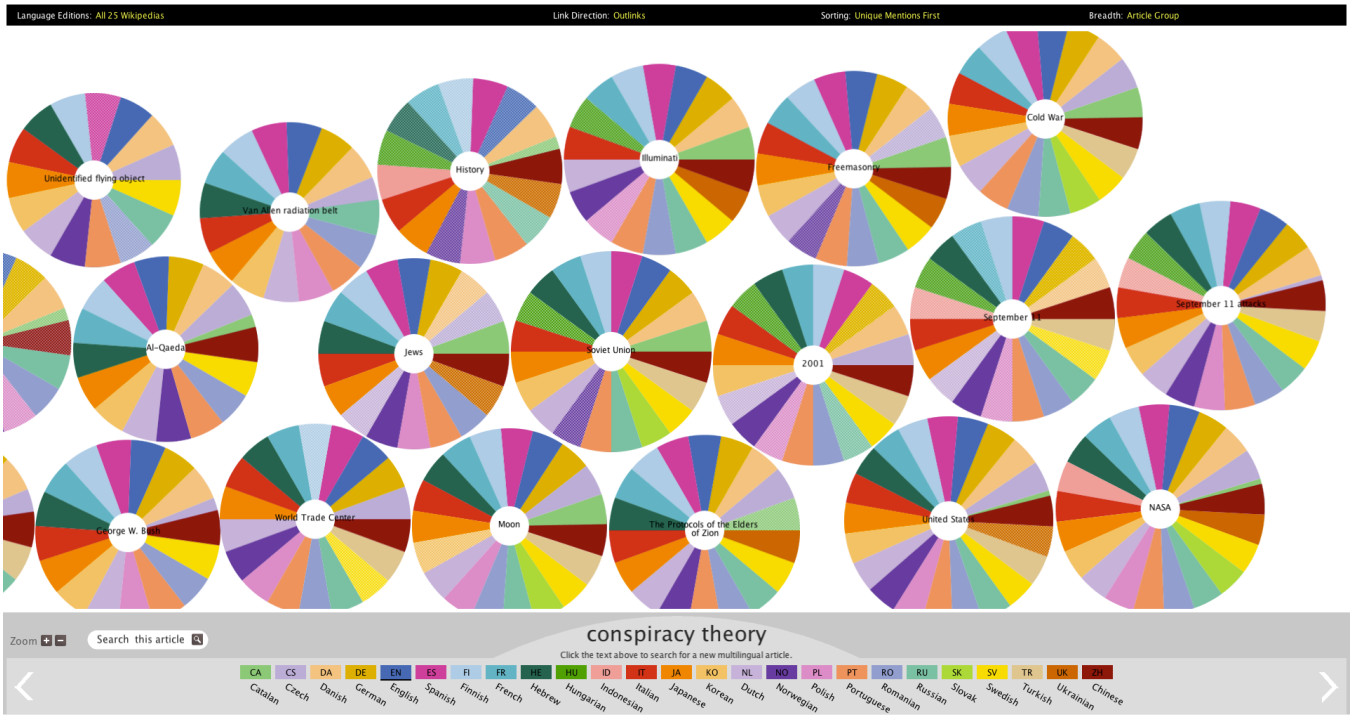

Figure 3. Panning over all the way to the right from Figure 2 reveals the linked topics that are shared among all or nearly all of the language editions' discussion of "Conspiracy theory". These include topics like "Politics", "September 11 attacks", and "NASA".

defined communities, not simply missed opportunities for translation. At the detailed content level, Pfeil et al. [27] analyzed a single concept in four language editions along dimensions of cultural influence. Their findings suggest that certain properties of edits to the corresponding articles could be predicted by the national culture most associated with the corresponding language. Callahan and Herring [7] examined articles about famous persons in the Polish and English Wikipedias and concluded that, while English generally had the most information, language-defined communities played a significant role in differences between the language editions. At a larger scale, Hecht and Gergle [14] used Wikipedia articles about geographic entities to show that there is a great deal of self-focus bias in each language edition of Wikipedia. In other words, the language editions place a disproportionate emphasis on areas within their "culture hearths" (home regions).

The second main area of research on multilingual Wikipedia focuses on propagating information from one Wikipedia language edition to another. These approaches include automated techniques as well as those that rely upon having a "human-in-the-loop". Ziggurat [2] automates the populating of Wikipedia infoboxes in one language edition using data from another language edition. Systems like WikiBhasha [19, 33] and Google Translator Toolkit [30] aim to achieve partial automation by using machine translation to aid humans in propagating information from one Wikipedia (typically English) to another (typically a language edition with a small number of articles). Finally, there have been recent efforts to use the power of human computation to rapidly translate articles between language editions [3, 10].

While these propagation approaches are driven by the commendable goal of making information widely available to speakers of any language, they share one or more of three important limitations—none of which are displayed by Omnipedia. First and foremost, their efforts to reduce the fractured nature of Wikipedia often come at the expense of preserving knowledge diversity across language editions. These approaches typically require decisions to be made about the "correct" content and discard distinctions between languages and viewpoints. Second, many of these systems focus on small numbers of language editions, often two at a time, thus incorporating only a small percentage of the information available in Wikipedia. While in theory these systems can be used to propagate information pairwise

Figure 4. The snippet explaining how Microsoft is discussed in the Hebrew Wikipedia article on conspiracy theory. This is the only part of Omnipedia that relies on live machine translation.

across language editions *ad infinitum*, they typically leave to future work the issue of what to do when conflicts arise between language editions and other major challenges that result from viewing the encyclopedia more globally. Finally, each of these approaches require large changes to Wikipedia itself for readers to benefit—changes that generally have not been realized at a global scale for various reasons including the amount of human labor involved. In contrast to these approaches, Omnipedia requires no changes to Wikipedia itself and can be used immediately.

Basic machine translation can be used to compare two or more articles about the same topic in different language editions, and there are sites like Manypedia [21] that automate this process. Moreover, individuals with skills in multiple languages can do the same without machine translation. However, in both cases, this approach is limited to a very small number of languages and detailed cross-language comparisons are quite difficult.

Although we believe it is the first of its kind in the multilingual space, Omnipedia is not the first system to use visualization to illuminate aspects of Wikipedia. A considerable body of work has used visualization techniques to surface coordination and conflict patterns among Wikipedia editors. Notable examples include History Flow [31], Chromograms [32], Revert Graph [29], and visualizations of trustworthiness metrics [18]. Each of these systems allows users to gain insight into the editing processes underlying a given Wikipedia language edition.

THE OMNIPEDIA SYSTEM

Omnipedia was designed to facilitate exploration of hyperlingual Wikipedia on a concept-by-concept basis. Users typically begin their interaction with Omnipedia by typing in a concept of interest, for instance "Conspiracy theory". Omnipedia will then look up the corresponding multilingual article and display the types of visualizations seen in Figures 1–3 using circles of different sizes and colors to indicate the topics that are discussed in various language editions' coverage of the entered concept.

Each circle denotes a topic that is mentioned in at least one language edition of Wikipedia. We extract these topics by looking at existing and missing hyperlinks to other articles in the same language edition (i.e., outlinks) within each of the current multilingual article's constituent monolingual articles. To determine the degree to which topics are discussed across language editions, Omnipedia merely looks up the corresponding multilingual article for each monolingual outlink. This allows the system to understand that nearly all articles about the concept "Conspiracy theory" discuss the United States even though each monolingual article links to its own language edition's equivalent article (e.g., "*Estats Units d'Amèrica*" (Catalan) and "*Vereinigte Staaten*" (German)). Each topic is thus itself another multilingual article that can be visualized in Omnipedia. By double-clicking on a circle, the user can browse through related topics, just as they can follow hyperlinks in the normal version of each language edition.

Figure 5. The multilingual article "Johnny Cash" with the interface language set to Chinese.

A central design premise for Omnipedia is that it is language neutral. As such, users are able to switch the interface language to any of the 25 supported languages. If a user switches the interface language to Chinese, for example, she is able to look up multilingual articles by their Chinese titles and sees topic titles in Chinese (Figure 5). Because this process relies on the structures built into Wikipedia, it involves no machine translation–an essential requirement for the system as machine translation on this scale would be too slow and, in the common case where thousands of topic titles are visualized, would excessively tax common machine translation APIs. When a topic that is discussed does not have a corresponding monolingual article in the interface language, we use a back-off strategy that, for instance, displays a single-language linked topic title in its native language.

Omnipedia is also language neutral in that it allows users to include or exclude any of the 25 supported languages, creating custom language sets at will. Omnipedia makes several built-in language sets available to the user, including "Languages of the Top Ten Economies", "NATO Languages", "EU Languages", etc. Once a language set is changed, all visualization and algorithmic systems are updated accordingly. For instance, in Figure 5, the user has selected the top ten economies language set.

When designing Omnipedia, we opted for a visualization strategy over a text-based approach largely because text alignment across just two languages is known to be a difficult problem [1, 4, 26], let alone text alignment across 25 languages. In doing so we lose certain advantages of text, like the grouping of related topics into a single, cohesive discussion. We mitigate this situation by using a well-known semantic relatedness algorithm [24] adapted for operation in Omnipedia. When a user clicks on a topic circle to view an explanation snippet, Omnipedia flags highly related topics with a "See Also" tag. For instance, when browsing the multilingual article "Barbecue", a user might be curious as to why "Kashrut" is discussed in the corresponding German article. When she clicks on the "Kashrut" circle, the topics "Israeli cuisine", "Modern Hebrew" and "Pastrami" are highlighted to guide further

exploration. As it turns out, there is an Orthodox Jewish barbecue festival in Memphis every year.

Omnipedia allows users to adjust a breadth setting that determines what is treated as an "article". This helps the user to deal with structural differences that may occur across articles or language editions. For example, large articles are sometimes broken down into a main article and its sub-articles (e.g., "Conspiracy theory" (English) and "List of conspiracy theories" (English), respectively). Because each language edition uses its own syntactic structures to indicate main article/sub-article relationships, we use a hand-coded list of all structures to discover sub-articles. At the broadest breadth setting, Omnipedia groups main articles and sub-articles into one multilingual article (as is done in Figures 1–3). The medium breadth setting shows only what is discussed in the main article(s). Finally, the narrowest breadth setting considers only the paragraphs that precede the first section break of each article, thus showing diversity in how concepts are summarized.

Finally, before moving on to Omnipedia's algorithmic challenges, it is important to briefly describe the dataset on which Omnipedia operates. We most recently updated Omnipedia's Wikipedia database in August 2011. This database contains 15.2 million articles and 445.3 million links. Using a sample of 1,500 multilingual articles, we confirmed the results of prior work [11, 15] in that only 15.9% (on average) of linked topics discussed in a multilingual article were discussed in all of the languages in which the article existed. Moreover, even the largest language edition, English, was missing approximately 60% of linked topics on average. These statistics were generated after running the algorithms described below, and confirm that Omnipedia users have a great deal of unique information and diverse knowledge to explore, no matter their native language(s).

ALGORITHMIC CHALLENGES

Omnipedia relies on a number of natural language processing (NLP) and information retrieval (IR) algorithms to support core functions. However, in the hyperlingual space, algorithmic approaches and their evaluation strategies have been critiqued for ignoring the diversity that exists across language-defined communities and being biased towards English [15, 17]. In this section, we describe novel algorithms and evaluation strategies we developed to achieve Omnipedia's mission of highlighting and respecting diversity across language editions, while at the same time showing information from many language editions in a structured fashion. Both our algorithms and their evaluation strategies have implications for other Wikipedia systems and research that require concept-by-concept access to multiple language editions (e.g., [2, 4]).

Hyperlingual Article Alignment

A fundamental requirement for Omnipedia is an algorithm that can group Wikipedia articles in different languages into multilingual articles. The algorithm should, for example, be able to match the article "Conspiracy theory" (English) to "*Teoría conspirativa*" (Spanish), etc. The accuracy of this algorithm and its ability to respect diversity across language editions are very important to Omnipedia's success because almost every aspect of Omnipedia operates on the multilingual articles that are its output.

We make use of Wikipedia's *interlanguage link* graph to identify and group articles about the same concept in different language editions. Interlanguage links (ILLs) are connections between articles in distinct language editions entered by humans and propagated by bots in certain cases. They are supposed to indicate near conceptual equivalence between pages in different languages. For instance, the article "Texas" (English) contains an ILL to "*Texas* (*staat*)" (Dutch). While there are omissions [28], ILLs have been shown to have relatively good coverage [15].

The standard approach in the literature is to assume that any two articles connected by a path in the ILL graph belong to the same multilingual article [2, 15]. This approach was insufficient for Omnipedia's hyperlingual context due to the fact that it ignores ambiguities in the ILL graph. Ambiguities occur when multiple articles in the *same* language edition are connected via articles in other language editions, meaning that a corresponding multilingual article would have more than one article per language. While only 1.0% of multilingual articles are initially ambiguous, many of them describe concepts that are of general and global interest because the potential for ambiguity increases as more language editions cover a given concept. This becomes a large problem for Omnipedia, as users are likely to be interested in concepts of general and global interest.

One major source of ambiguities in the ILL graph is conceptual drift across language editions. Conceptual drift stems from the well-known finding in cognitive science that the boundaries of concepts vary across language-defined communities [13]. For instance, the English articles "High school" and "Secondary school" are grouped into a single connected concept. While placing these two articles in the same multilingual article may be reasonable given their overlapping definitions around the world, excessive conceptual drift can result in a semantic equivalent of what happens in the children's game known as "telephone". For instance, chains of conceptual drift expand the aforementioned connected concept to include the English articles "Primary school", "Etiquette", "Manners", and even "Protocol (diplomacy)". Omnipedia users would be confused to see "Kyoto Protocol" as a linked topic when they looked up "High school". A similar situation occurs in the large connected concept that spans the semantic range from "River" to "Canal" to "Trench warfare", and in another which contains "Woman" and "Marriage" (although, interestingly, not "Man").

In order for Omnipedia to enable users to explore this vital 1.0% of concepts, we needed an algorithm to split concepts

that were subject to runaway conceptual drift[1]. However, at the same time, this algorithm should respect the fact that different languages may define a concept more widely or narrowly than other languages.

Approach to Resolving Ambiguities

Our approach draws on the conceptual spaces framework from cognitive science [13], in which a concept is a region in a multi-dimensional semantic space. Generally speaking, the higher the average semantic similarity between pairs of concept instances, the smaller the area of the concept. The goal of our approach is thus to split ambiguous concepts by dividing them into regions with higher average semantic similarity. One method would be to attempt to match the average semantic similarity of the 99% of concepts that are not ambiguous. Alternatively, a hyperlingual application designer may want to allow for slightly more conceptual drift (e.g., to include "High School" and "Secondary School" in the same concept), while at the same time eliminating cases like "Woman" and "Marriage".

In order to enable this approach in practice, we introduce an algorithm that allows hyperlingual application designers to adjust the amount of allowable conceptual drift to suit their application needs. Our algorithm strategically removes ILL edges from ambiguous concepts, splitting connected components of the ILL graph into more coherent groups. We removed edges along two dimensions: (1) limiting the number of edges from a given language that can point to the same article in another language (*MaxEdges*), and (2) using a voting scheme that required a certain percentage of language editions to agree on an edge before it could remain (*MinLangs*).

Finally, to measure the semantic similarity of multilingual articles generated by our algorithm, we developed a hyperlingual approach that leverages the same semantic relatedness measure we use to calculate "See Also" topics in Omnipedia. This measure can be used to calculate the semantic similarity between pairs of articles that make up the newly generated multilingual articles[2], regardless of the languages in which those articles are written.

Exploring Parameters

To better understand the ability of our algorithm to generate cohesive multilingual articles as well as its ability to allow designers some flexibility in that cohesiveness, we randomly selected 2,000 ambiguous multilingual articles from our dataset and performed a grid search on the parameters. To establish a reasonable upper bound, we also randomly selected 2,000 unambiguous multilingual articles with pages in two or more languages. For both groups of articles, we calculated the pairwise in-concept semantic similarity for each possible article pair. As a baseline, we did the same for the default state of the ILL graph. For the default state of the ILL graph and the output of our algorithm, we also report the mean "out-concept" similarity, which is the average similarity of articles not in the same concept. Setting *MaxEdges* to any value other than one significantly reduced the average semantic similarity in all cases, so we only report data where *MaxEdges* = 1. Finally, in order to provide an additional perspective on our algorithm's performance, we evaluated our results against the comparable portions of de Melo and Weikum's bilingual German/English dataset [22].

As shown in Table 1, using our algorithm it is possible to match and even exceed the semantic cohesiveness of non-ambiguous multilingual articles, at least with our 25-language dataset. Moreover, for the parameters that result in these high average similarities, performance on the de Melo and Weikum dataset matches and exceeds that of de Melo and Weikum's algorithm. This is true even though our algorithm is far simpler than their complex linear program solution, although their work is focused on graph theory aspects of the problem. Table 1 also shows that our algorithm gives designers significant leeway in allowing for more conceptual drift, meeting the second goal for the algorithm.

The question then becomes, as application designers of Omnipedia, which parameters should we choose? We initially used *MaxEdges* = 1 and *MinLangs* = 70%, matching the semantic similarity of unambiguous concepts. This effectively normalized ambiguity across our entire dataset. However, after examining hundreds of concepts split by our algorithm set to these parameters, we determined that it was too strict to meet Omnipedia's goals of respecting diversity in concept definitions. For instance, "High school" (English) and "Secondary school" (English) were split into separate concepts, even though in many languages these concepts are one and the same. By reducing

MaxEdge	*MinLang*	In-Concept Similarity	Out-Concept Similarity	de Melo Accuracy
1	0%	0.65	0.29	73.7
1	20%	0.67	0.29	77.0
1	**50%**	**0.73**	**0.30**	**81.2**
1	70%	0.78	0.31	87.5
1	90%	0.81	0.33	91.4
1	100%	0.82	0.41	87.5
ILL Graph		0.41	0.26	51.2
Unambiguous Articles		0.78	n/a	n/a
de Melo Algorithm		n/a	n/a	89.7

Table 1. Ambiguity levels of the concepts output by our article alignment algorithm. Bold indicates the parameters used by Omnipedia.

[1] Mistaken ILLs input by Wikipedia editors are another source of ambiguities but algorithmically speaking, these can be treated simply as extreme cases of conceptual drift.

[2] It is standard research practice to use semantic relatedness measures to calculate similarity [6].

MinLang to 50%, we found that we could still maintain a high in-concept similarity, while also including these two articles in the same multilingual concept. Moreover, the algorithm with these parameters had no trouble splitting runaway conceptual drift cases like "Woman" and "Marriage", "River" and "Trench warfare", etc.

Hyperlingual Link Alignment

Early versions of Omnipedia only used explicit outlinks as a proxy for topics discussed in a multilingual article. This approach was motivated by the increasing popularity of "bags of links" models to represent unstructured texts (e.g., [1, 25]) and the use of these models in the multilingual Wikipedia space [1, 15, 21]. However, in a hyperlingual context this form of representation suffers from a failure to recognize the varying linking practices across language editions and individual articles.

In Omnipedia, this problem manifests when an article in a given language edition discusses a topic linked to by other language editions, but no explicit link exists in the given language edition. A bag of links representation would assume that the topic is not discussed in the article, even though it may play a prominent role. Ignoring this "missing link" issue would unfairly bias results toward those language editions with more links on a given topic. Without the link alignment algorithm described in this section, Omnipedia would inform users that the multilingual article "Sociology" would have discussed "Theory" in only one language edition. Moreover, because different language editions have different (and relatively strict) rules about linking to dates, the "September 11" and "2001" circles in Figure 3 would be much smaller, misleading users into thinking the events of that day were a less global source of conspiracy theories.

We resolve this problem with our hyperlingual link alignment algorithm, which is best understood by example. Returning to Figure 3, we find that the Spanish article about conspiracy theories does not link to an article about unidentified flying objects (UFOs) while many of the other language editions do. However, the Spanish Wikipedia does have an article about UFOs ("*Objeto volador no identificado*" (Spanish)). This is the situation in which our hyperlingual link alignment algorithm is applied. Broadly speaking, the algorithm searches the article "*Teoría conspirativa*" (Spanish) for the text "*Objeto volador no identificado*" and its synonyms (redirects in the Spanish Wikipedia). It happens to find the text "*OVNI*", one such redirect, in a discussion of prominent conspiracy theories, indicating that UFOs are indeed a missing link. Omnipedia therefore includes Spanish in the set of languages that discuss UFOs (as seen in Figure 3). This process is repeated for all language editions in Omnipedia.

Evaluation Experiment

To evaluate whether our approach found an unreasonable number of incorrect links or overlooked an unreasonable number of missing links, we turn to the literature on the similar but monolingual problem of wikification [23, 25]. A robust evaluation strategy in this literature involves measuring accuracy against manually labeled data [25]. However, this method runs into the central problem of hyperlingual algorithm evaluation: finding several human labelers for each of many languages is a considerable challenge. We were able to sidestep this problem by leveraging the large supply of labeled data we already have in the form of the links that *do* exist in each language edition. Following Mihalcea and Csomai [23], we strip each test article of all markup (indications of links) and then assess whether or not we can accurately rediscover the links in the article. If our algorithm can do this successfully, we can reasonably suggest that our algorithm can do the same in text that is missing links in the actual hyperlingual dataset.

We ran this experiment on 15,000 randomly selected multilingual articles that have articles in at least two languages. The minimum number of articles considered for a given language was 599 (Indonesian) and the maximum was 12,391 (English). Our algorithm's average recall across all test articles was 75.0% and its average precision was 80.4%. These results are on par with the accuracy of the popular wikification algorithm by Milne and Witten [25], although the problem space is slightly different. While precision was in the upper 70s through high 80s for all languages (range = 76.2–88.7%), recall exhibited greater variance (range = 53.7–86.8%). Chinese had 80.9% recall despite the text processing adaptions necessary to support East Asian languages. However, highlighting the importance of hyperlingual evaluation, we discovered that Finnish and a few of the Slavic languages had lower recall in the mid-50s and low 60s. While this performance is comparable with another popular wikification algorithm [23], it does put these languages at a slight disadvantage in Omnipedia. Future work involves experimenting with different stemming algorithms and taking other approaches to improve recall in these languages.

On average, our algorithm increased the number of linked topics in an article by 51.7%, although a portion of this gain is due only to the differences in date linking practices. For many of the smaller language editions, the number of found topics was much higher. This is not a surprise, as the link density of the English Wikipedia is known to have grown over time [8]. For instance, the algorithm more than doubled the number of Indonesian linked topics per article (134.0%). Even in the Slavic languages with lower recall, we found large numbers of new links (e.g., 67.6% increase for Slovak, 59.8% for Czech). For the older, larger language editions, the number of new links was smaller (e.g., 37.3% for Japanese, 38.7% for English).

Our algorithm's ability to find many new links with good precision and good recall allows Omnipedia to accurately visualize content beyond that just in the link graph. It also problematizes the use of the popular bag of links approach

in a hyperlingual context. For instance, an algorithm or hyperlingual application that compared English and Indonesian articles using a bag of links approach would heavily bias the English Wikipedia.

STUDY

To better understand how people might interact with the diverse information made accessible by Omnipedia, we conducted a study with 27 participants. This allowed us to observe how people gained insights when viewing concepts of their choice through Omnipedia's hyperlingual lens.

Participants and Method

Twenty-seven people (14 female, 13 male, ranging from 18-62 years old) participated in the study, all of whom had accessed Wikipedia at least once in the past 30 days. Participants consisted of 20 native English speakers, four native Chinese (Mandarin) speakers, one native Russian speaker, and two native speakers of English and one other language not supported by Omnipedia. Sixteen users were fluent or proficient in at least one other language besides their native language. These additional languages were Spanish (10 users), English (5), French (1), Japanese (1) Korean (1), and Telugu (1, not supported by Omnipedia). On average, participants had used the English Wikipedia for 6 years (self-reported, *SD* = 1.84). Ten participants had previously seen a non-English Wikipedia, but only three considered themselves frequent users.

Participants arrived at the laboratory and were taken to a private room with a desktop machine and 23" display. The experimenter then provided a 10-minute demonstration of Omnipedia's main features and afterwards proceeded to a separate observation room. Participants were given 30 minutes to freely explore multilingual articles in Omnipedia using any of the languages they wished. Afterwards, the experimenter returned to the room to lead a structured interview asking participants to reflect on their experience. Participants were prompted with specific instances drawn from their interaction logs. We based our analysis on 58,900 words of transcribed interviews, 41,704 logged events (including mouse hovers, clicks, queries, and changes to view settings), and 30 pages of observation notes. We now use this data to characterize user exploration of hyperlingual Wikipedia through Omnipedia and describe some of the insights they shared with us.

Results

Just over half (14) of our users had never seen a non-English Wikipedia prior to the study. When using Omnipedia, all users took the opportunity to access information from a non-English Wikipedia, the five most popular being French, Italian, Russian, German, and Spanish. Twenty-two users switched to one of Omnipedia's built-in language sets at least once during the study (*M* = 3.1 switches, *SD* = 3.49). Twelve of these users also created a custom language set. Users tailored these sets based on their own language proficiency, relevance to the concept of interest, or curiosity about a never-before seen Wikipedia.

On average, users looked up 15 multilingual articles (*SD* = 6.25). They clicked on 60 discussed topics on average (*SD* = 34.7) to load the type of snippets seen in Figure 4. Of all the topics users clicked on, 26.7% had been highlighted as related topics by the semantic relatedness algorithm.

Exploring Similarities and Differences

After seeing the diversity of linked topics among language editions, many users concentrated on the most common topics (or biggest circles), typically citing reasons involving the perceived importance of these topics (e.g., "if it was in all four languages, it must be important" (P7)). Viewing these topics often required users to pan all the way to the opposite end of the visualization (as seen in Figure 3), which they took the effort to do in order to gain insight into what was "well known" worldwide (P18). Many of these users were satisfied with clicking on these topics and reading just one of its multiple snippets.

However, other users read all of the snippets from more globally discussed topics in detail to see if there were "cultural nuances" (P23). Users who engaged in this type of behavior realized that just because multiple language editions shared a link to a topic did not necessarily mean that they agreed on *how* the topic was related. For example, P4 recalled looking up the multilingual article "Boeing 767" and discovering that the English snippet on a plane crash in Egypt included "different perspectives on what happened" while other language editions "just summarized one sentence". Similarly, P12 expected that German coverage of "Siemens" would gloss over the company's support of the Nazi movement during World War II. He was surprised to find that the German Wikipedia's snippet was "the most descriptive about that fact".

Other users spent little time investigating the most common topics, regarding them as "pretty obvious" (P25) or "basic things" (P10) that would not yield the most interesting insights. Instead, they searched for differences in topic coverage by looking at single-language topics (or smallest circles) across language editions. One approach was to examine relative proportions of single-language topics in a given multilingual article. For instance, P17 inferred that American basketball player Dwight Howard was "definitely more famous in the English version than in any other language" based on the considerable number of topics discussed only in the English Wikipedia. Likewise, P24 was not surprised to find that a minor tennis player only had coverage in English while "Rafael Nadal had more single-language links from Spanish and other languages because he's a worldwide figure". She, like others, interpreted these differences as a measure of the concept's (e.g., Rafael Nadal's) global "reach" or "impact".

Users who took this approach also discovered distributions of single-language topics that belied their expectations. P10 was surprised to find that "even Italian and Spanish had something to say" about Jainism, an Indian religion. P6 compared several music genres and was not surprised to

find that hip-hop had "more in English" but was surprised to discover that reggae had "a lot in Japanese".

Another approach for finding differences among language editions involved targeting concepts that might be more likely to reveal differences in perspective. A subset of users actively sought out what they considered to be "globally polarized" (P20) or "heavily charged" (P23) concepts like "Climate skepticism" and "War on Terror". They intentionally included language editions that they thought would reveal "different sides" (e.g., P13, who looked up "The Holocaust" in German, Hebrew, and Polish). In most cases, however, users did not find the extreme differences they anticipated, leading them to reconsider their own expectations regarding hyperlingual Wikipedia.

Discovering New Knowledge

Users actively sought out knowledge not available in their own language editions. For example, the 11 "monolingual" users who were fluent in English but had no more than rudimentary knowledge of another language clicked on topics that, on average, were mentioned in 2.79 language editions (*SD* = 2.06). Thirty-six percent of all topics clicked by these users were not discussed at all in English.

Users often reported clicking on topics discussed in one language because they might have "interesting facts that I hadn't heard of" (P26). For certain multilingual articles, users paid attention to unique topics in a single language edition where they expected a close tie to the language's culture hearth. For example, P6 "focused on the Italian side [of the visualization] just because Sardinia's in Italy". He also looked at Chinese-only topics discussed in "Google" because he thought they might reference Google's search restrictions in China. Similarly, P16 thought "the Japanese-only information will be more authentic since Ayumi Hamasaki is from Japan". Conversely, one user decided to exclude a language (Chinese) from the interface because it "wasn't giving me much" in terms of unique information.

In other cases, users investigated single-language topics from many language editions. For instance, P10 wanted to see "if maybe one culture viewed a certain aspect of 'Beauty' that [she] didn't know". After discovering a number of Japanese-only topics that seemed to emphasize "character", she went on to examine English-only topics and observed that they discussed "beauty in the eye of the beholder" as well as "physical" attributes.

Finally, it is worth noting that the sheer amount of single-language topics was a revelation to the majority of users. Reflecting on their use of Omnipedia, a few users who initially focused on the more global topics wished they had more time to explore the single-language topics, as those may have yielded different insights. P1 even told us in hindsight, "If I had bothered to take my time and go through all the single [language] ones, I think I would have learned more about what the differences were."

In sum, four key insights emerged from users' interactions with Omnipedia. First, users took advantage of the fact that using a hyperlingual lens, they could identify the most commonly and globally discussed aspects of a concept's Wikipedia definition. Second, they were able to discover both similarities and differences in how these topics were discussed among language editions. Third, access to single-language topics allowed users to not only filter interesting topics based on inferences of self-focus bias [14], but also get a big picture view of how much topics were being discussed in different language editions. Finally, users began to comprehend the magnitude of information that was not available to them in the English Wikipedia.

FUTURE WORK

We have focused this paper on Omnipedia's capacity to help users explore hyperlingual Wikipedia. However, we are working to extend Omnipedia to other forms of hyperlingual user-generated content as well. Sites such as Twitter and Flickr suffer from the same language barriers as Wikipedia and have also been shown to display important differences across languages [9, 16]. Future work might treat a Twitter hash tag as an "article" and mine tweets posted in many languages that contain the hash tag for discussed topics. Similarly, a group of related photos (e.g., of the same event) could be used as the "article" and the photos' tags could be considered topics.

As Wikipedia is an extremely popular source of world knowledge for many artificial intelligence, IR and NLP systems, we suspect that the algorithms introduced here will apply outside of the Omnipedia context. For instance, we are working to build an improved hyperlingual semantic relatedness measure based on our link alignment algorithm and our concept alignment strategy. This measure could be used in everything from Omnipedia to cross-language coordination tools [5] to more traditional cross-language information retrieval.

Finally, as we move toward a wider deployment of Omnipedia, we have started collecting user feedback about which information should be directly propagated across language editions and which information is more culturally specific. Users can already flag content in explanations as information that should be in other Wikipedias as well (Figure 4). We believe that some of the non-global information that Omnipedia makes salient should definitely be propagated to other language editions. However, as this occurs, we anticipate that users will be able to focus more clearly on Omnipedia's ability to surface cultural diversity.

CONCLUSION

In this paper, we have made several contributions. First, we introduced Omnipedia, a system that for the first time allows simultaneous access to large numbers of Wikipedia language editions. In doing so, Omnipedia makes salient the diversity of knowledge represented in all considered language editions. Powering Omnipedia are several new algorithms that preserve diversity while solving large-scale

data processing issues. Finally, we demonstrated the kinds of insight Omnipedia affords with a 27-participant study, which led one user to remark "It's ridiculous how many different things are mentioned in different languages that aren't mentioned in others."

ACKNOWLEDGMENTS

We thank our reviewers, our participants, Alan Clark, Lauren Scissors, Candace Brown, Jermaine Dictado, and the CollabLab for their valuable insight. Funding was provided by NSF grant #0953943 and an NSF GRFP award.

REFERENCES

1. Adafre, S.F. and de Rijke, M. 2006. Finding Similar Sentences Across Multiple Languages in Wikipedia. *EACL 2006 Workshop on New Text, Wikis and Blogs and Other Dynamic Text Sources*.
2. Adar, E., Skinner, M. and Weld, D.S. 2009. Information Arbitrage Across Multi-lingual Wikipedia. *WSDM '09*.
3. von Ahn, L. 2011. Three human computation projects. (2011). *SIGCSE '11*.
4. Au Yeung, C.-man, Duh, K. and Nagata, M. 2011. Providing Cross-Lingual Editing Assistance to Wikipedia Editors. *CICL '11*.
5. Bergstrom, T. and Karahalios, K. 2009. Conversation clusters: grouping conversation topics through human-computer dialog. *CHI '09*.
6. Budanitsky, A. and Hirst, G. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*. 32, 1 (2006), 13-47.
7. Callahan, E.S. and Herring, S.C. Cultural bias in Wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology*. 62: 1899–1915.
8. Capocci, A., Servedio, V.D.P., Colaiori, F., Buriol, L.S., Donato, D., Leonardi, S. and Caldarelli, G. 2006. Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia. *Physical Review E*. 74, 3 (2006), 036116.
9. Dong, W. and Fu, W.-T. 2010. Cultural difference in image tagging. *CHI '10*.
10. Duolingo: *http://duolingo.com/*. Accessed: 2011-09-13.
11. Filatova, E. 2009. Multilingual Wikipedia, Summarization, and Information Trustworthiness. S*IGIR Workshop on Information Access in a Multilingual World*.
12. Frequently asked questions - Wikimedia Foundation: *http://wikimediafoundation.org/wiki/Frequently_asked_questions*. Accessed: 2011-09-21.
13. Gärdenfors, P. 2000. *Conceptual Spaces: The Geometry of Thought*. The MIT Press.
14. Hecht, B. and Gergle, D. 2009. Measuring Self-Focus Bias in Community-Maintained Knowledge Repositories. *Communities and Technologies 2009*.
15. Hecht, B. and Gergle, D. 2010. The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context. *CHI '10*.
16. Hong, L., Convertino, G. and Chi, E.H. 2011. Language Matters in Twitter: A Large Scale Study. *ICWSM '11*.
17. Jarmasz, M. and Szpakowicz, S. 2003. Roget's thesaurus and semantic similarity. *RANLP '03*.
18. Kittur, A., Suh, B. and Chi, E.H. 2008. Can you ever trust a wiki?: impacting perceived trustworthiness in wikipedia. *CSCW '08*.
19. Kumaran, A., Datha, N., Ashok, B., Saravanan, K., Ande, A., Sharma, A., Vedantham, S., Natampally, V., Dendi, V. and Maurice, S. 2010. WikiBABEL: A System for Multilingual Wikipedia Content. *American Machine Translation Association (AMTA) Workshop*.
20. List of Wikipedias: http://meta.wikimedia.org/*wiki/List_of_Wikipedias*. Accessed: 2011-09-20.
21. Manypedia: 2011. *http://www.manypedia.com/*.
22. de Melo, G. and Weikum, G. 2010. Untangling the Cross-Lingual Link Structure of Wikipedia. *ACL '10*.
23. Mihalcea, R. and Csomai, A. 2007. Wikify!: linking documents to encyclopedic knowledge. *CIKM '07*.
24. Milne, D. and Witten, I.H. 2008. An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. *WIKIAI '08*.
25. Milne, D. and Witten, I.H. 2008. Learning to link with wikipedia. *CIKM '08*.
26. Oh, J.-H., Kawahara, D., Uchimoto, K., Kazama, J. and Torisawa, K. 2008. Enriching Multilingual Language Resources by Discovering Missing Cross-Language Links in Wikipedia. *WIIAT '08*.
27. Pfeil, U., Zaphiris, P. and Ang, C.S. 2006. Cultural Differences in Collaborative Authoring of Wikipedia. *Journal of Computer-Mediated Communication*. 12, 1, 88-113.
28. Sorg, P. and Cimiano, P. 2008. Enriching the Crosslingual Link Structure of Wikipedia - A Classification-based Approach. *WIKI-AI '08*.
29. Suh, B., Chi, E.H, Pendleton, B.A. and Kittur, A. 2007. Us vs. Them: Understanding Social Dynamics in Wikipedia with Revert Graph Visualizations. *VAST '07*.
30. Translating the world's information with Google Translator Toolkit: 2009. *http://googleblog.blogspot.com/2009/06/translating-worlds-information-with.html*. Accessed: 2011-09-16.
31. Viégas, F.B., Wattenberg, M. and Dave, K. 2004. Studying cooperation and conflict between authors with history flow visualizations. *CHI '04*.
32. Wattenberg, M., Viégas, F.B. and Hollenbach, K. 2007. Visualizing activity on wikipedia with chromograms. *INTERACT '07*.
33. WikiBhasha beta – A multi-lingual content creator for Wikipedia: *http://www.wikibhasha.org/*.